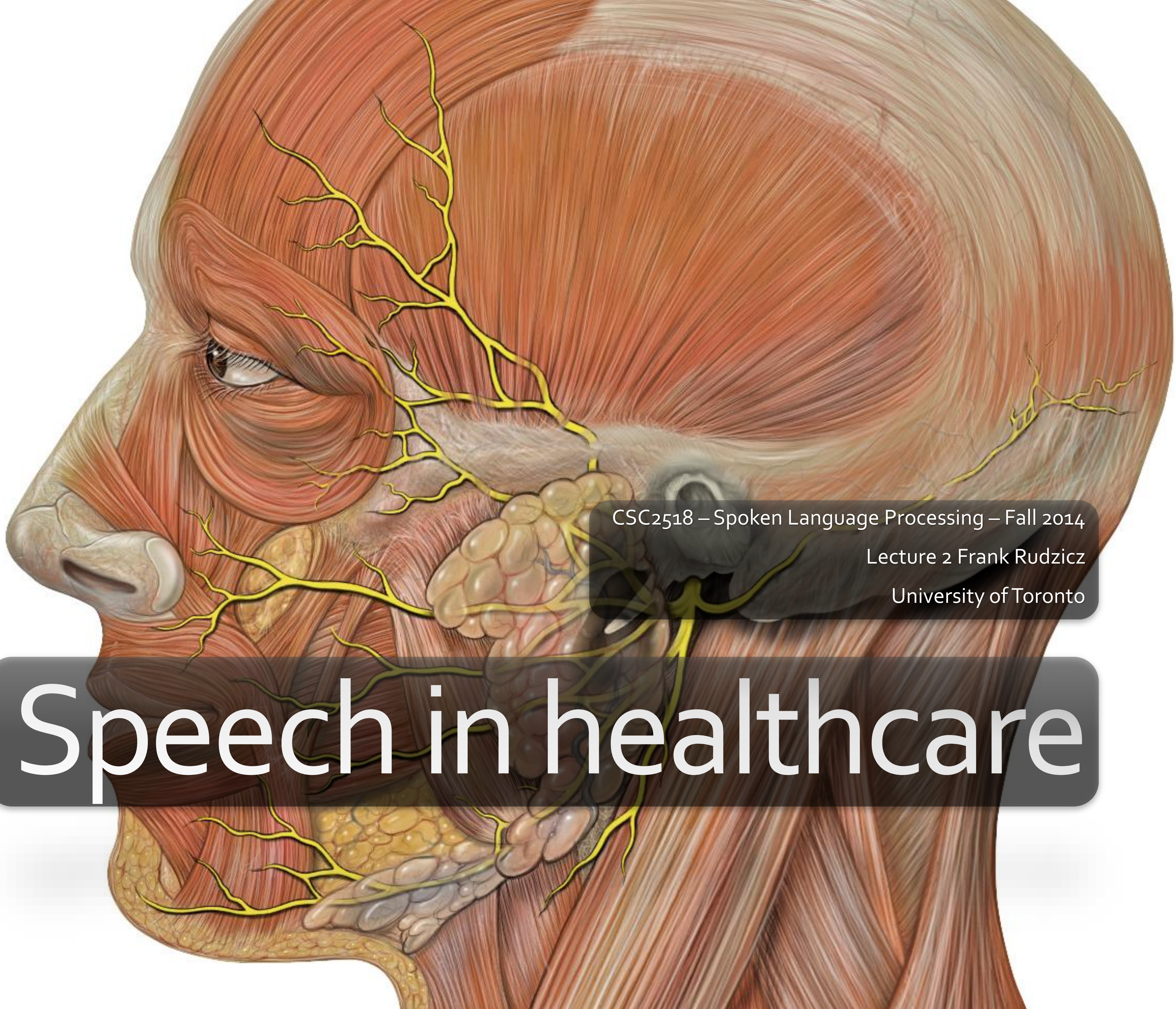


CSC2518 – Spoken Language Processing – Fall 2014

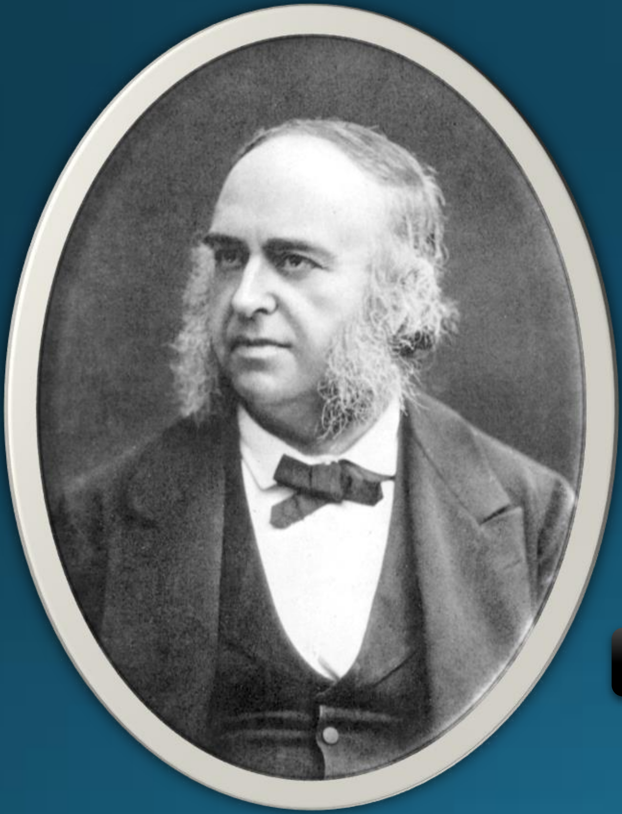
Lecture 2 Frank Rudzicz

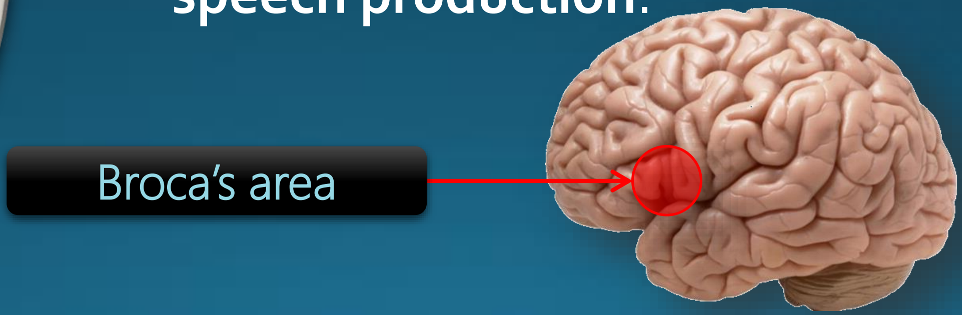
University of Toronto

Speech in healthcare

Studying how systems break down

- Observing how **closed systems** ***fail*** can be a **valuable method** in discovering how those systems **work**.
- **Paul Broca** (left) discovered, in 1861, that a **lesion** in the **left** ventro-posterior **frontal lobe** caused **expressive aphasia.**
- This was the first **direct** evidence that **language function** was **localized**.
  - It hinted at a **mechanistic** view of **speech production**.





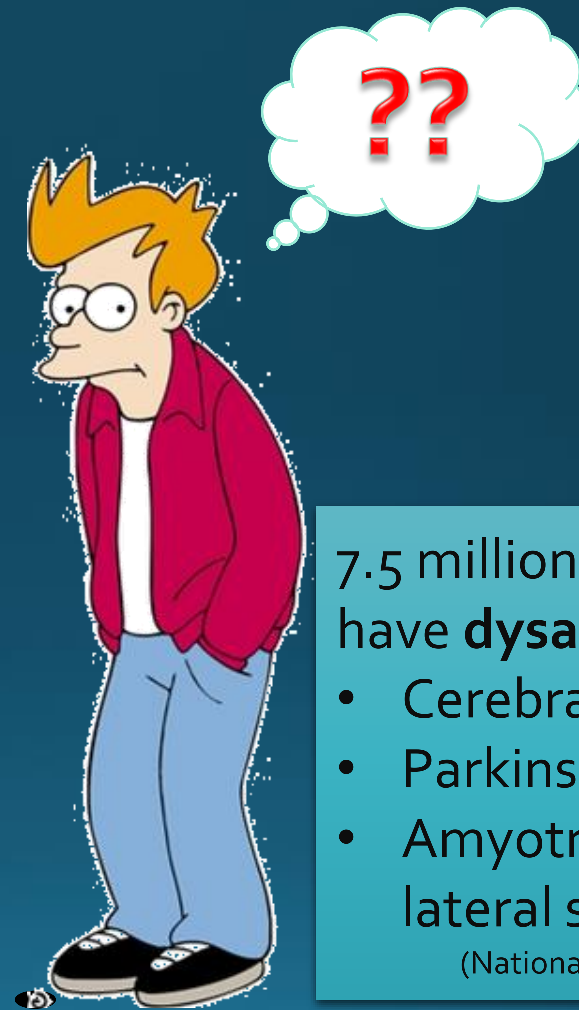
UNIVERSITY OF TORONTO

Today

- Physical production disorders (e.g., cerebral palsy)
  - Capturing data
  - Using those data in speech recognition
  - Speech output devices
- Physical perception disorders (e.g., deafness)
  - Hearing aids
- Cognitive problems (e.g., Alzheimer's disease)
  - Neural origins
  - Assistive technologies

Dysarthria

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



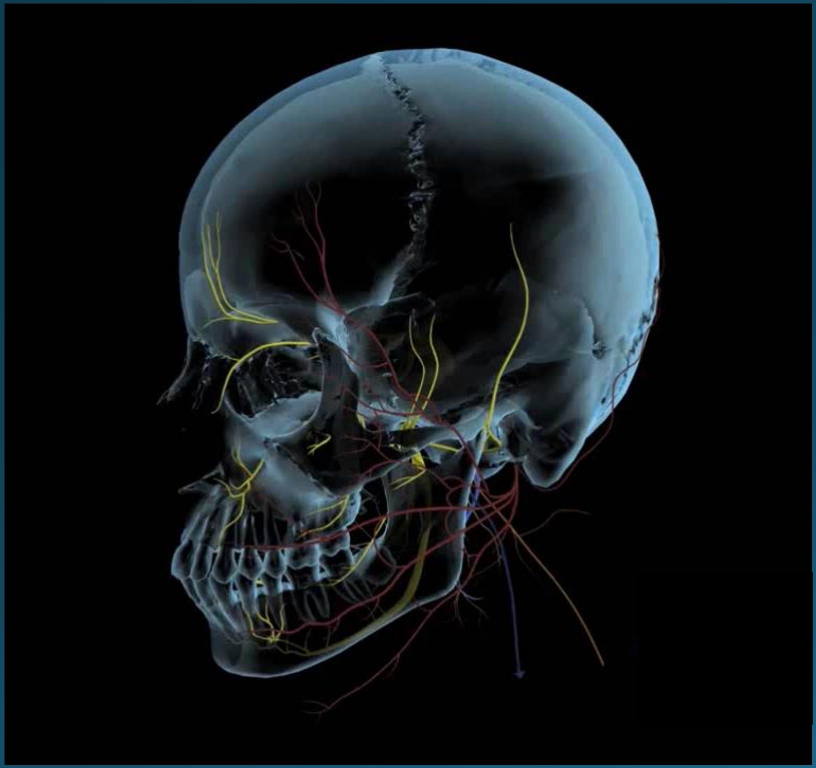
7.5 million Americans have **dysarthria**
- Cerebral palsy,
- Parkinson's,
- Amyotrophic lateral sclerosis)

(National Institute of Health)

University of Toronto

Neural origins

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



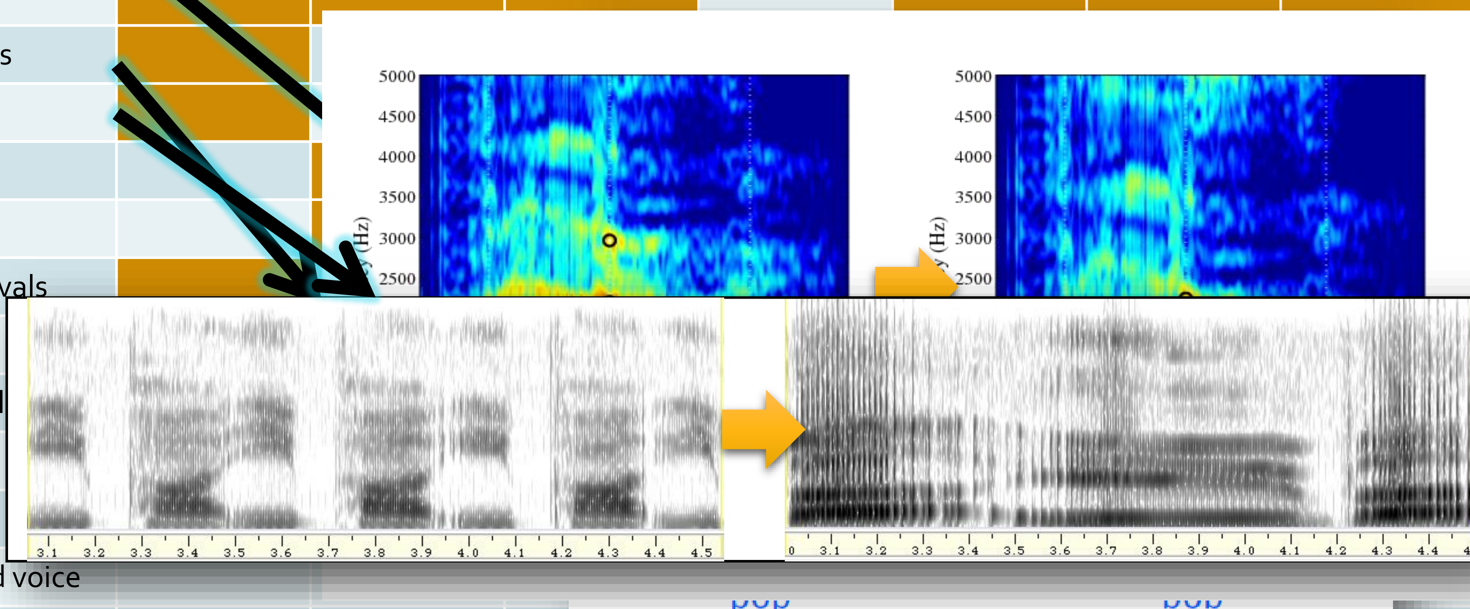
Type	Primary lesion site
Ataxic	Cerebellum or its outflow pathways
Flaccid	Lower motor neuron (≥1 cranial nerves)
Hypo-kinetic	Basal ganglia (esp. substantia nigra)
Hyper-kinetic	Basal ganglia (esp. putamen or caudate)
Spastic	Upper motor neuron
Spastic-flaccid	Both upper and lower motor neurons

(After Darley *et al.*, 1969)

UNIVERSITY OF TORONTO

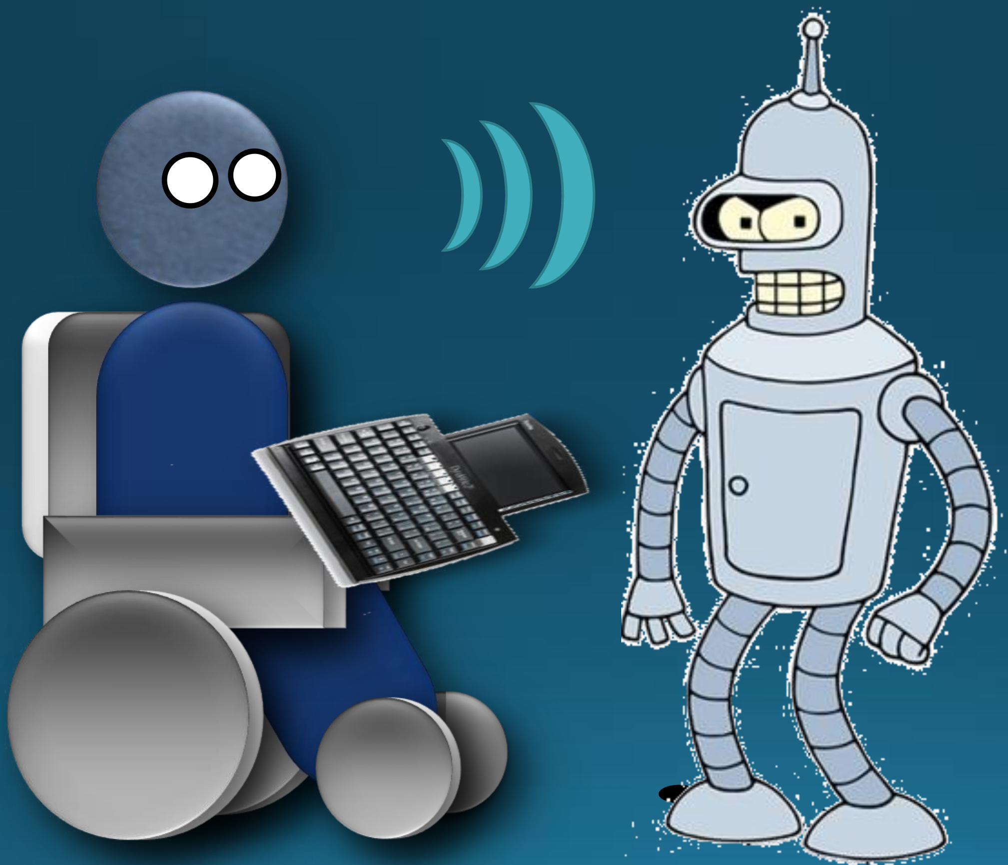
Characteristics of dysarthria

	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 sil							
Variable rate							
Breathy voice							
Strain-strangled voice							
…							



Dysarthria

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



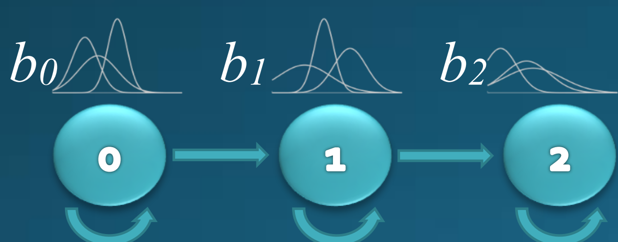
Can we use ASR for dysarthria?

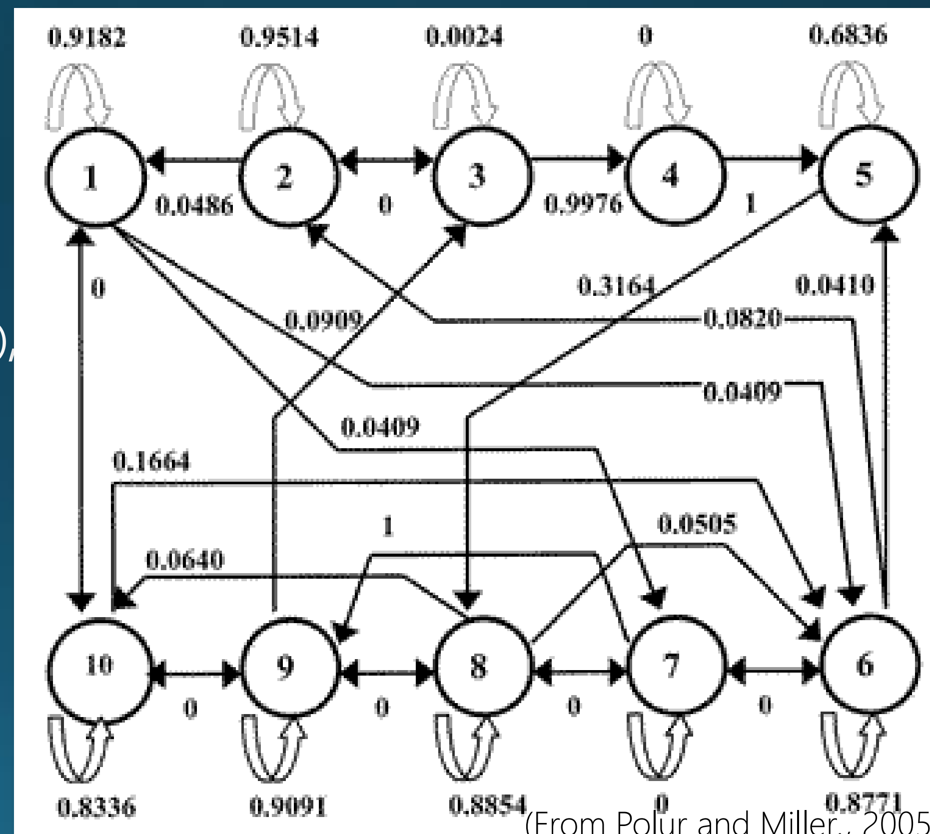
UNIVERSITY OF TORONTO

Accounting for 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**.

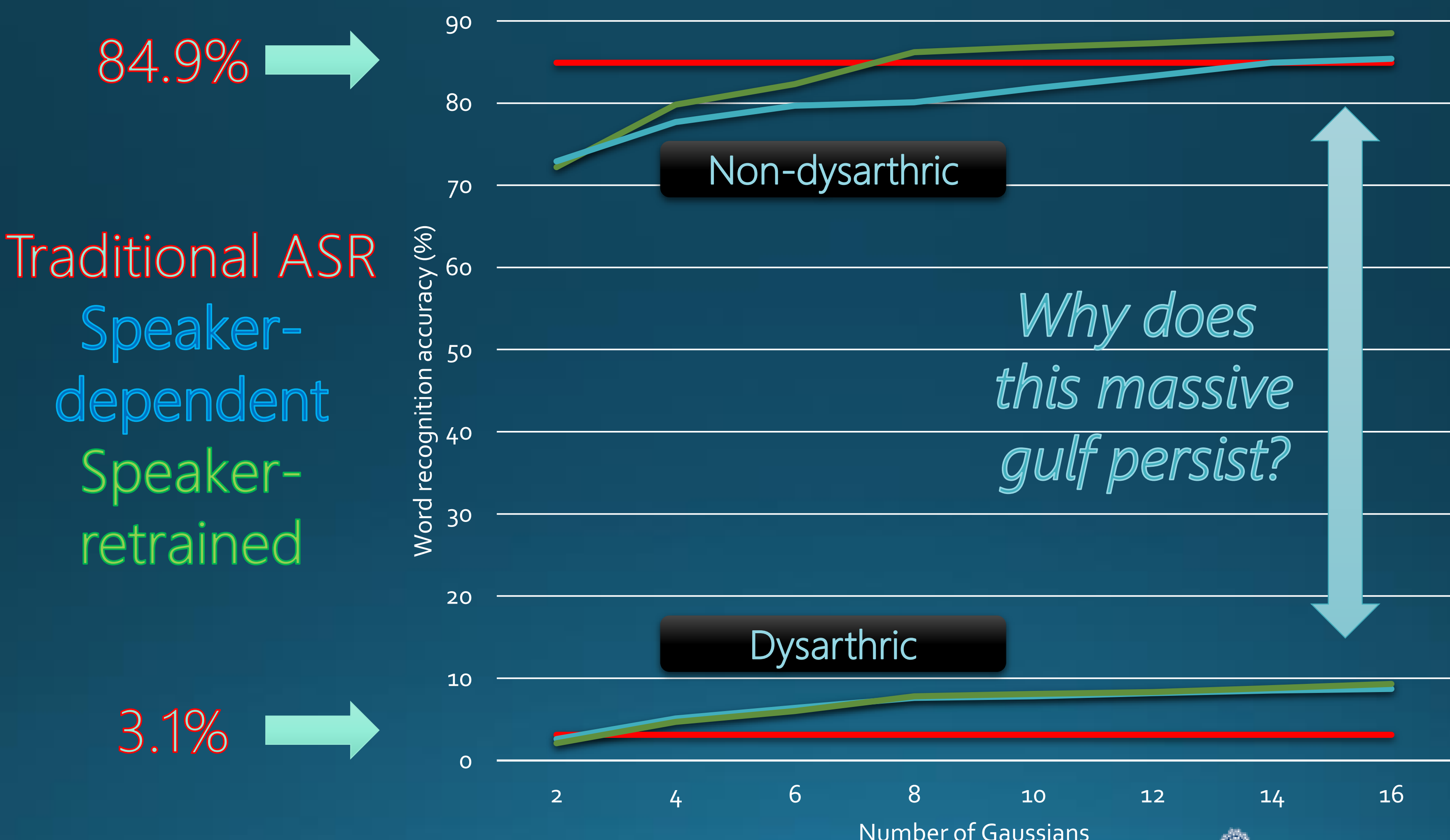




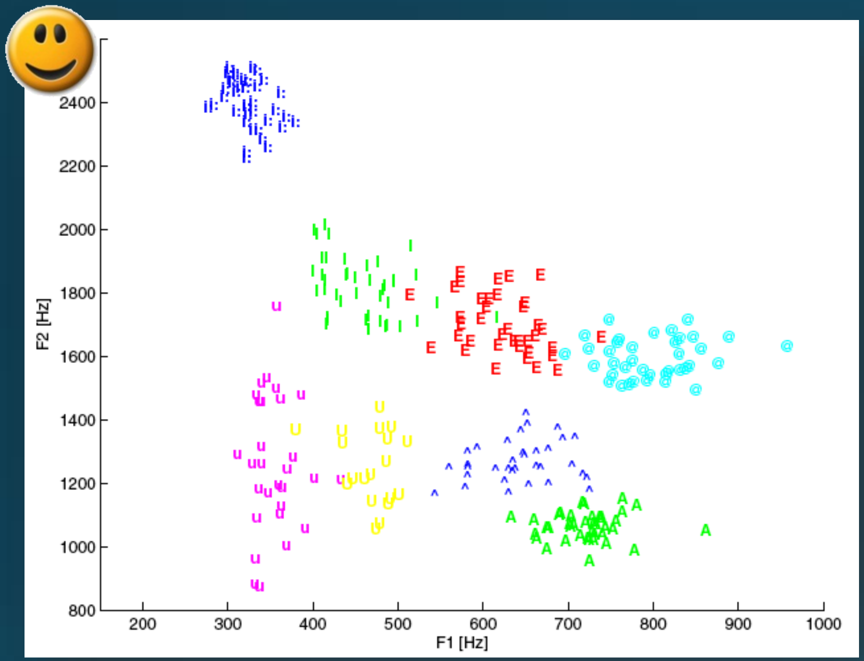
(From Polur and Miller., 2005)

UNIVERSITY OF TORONTO

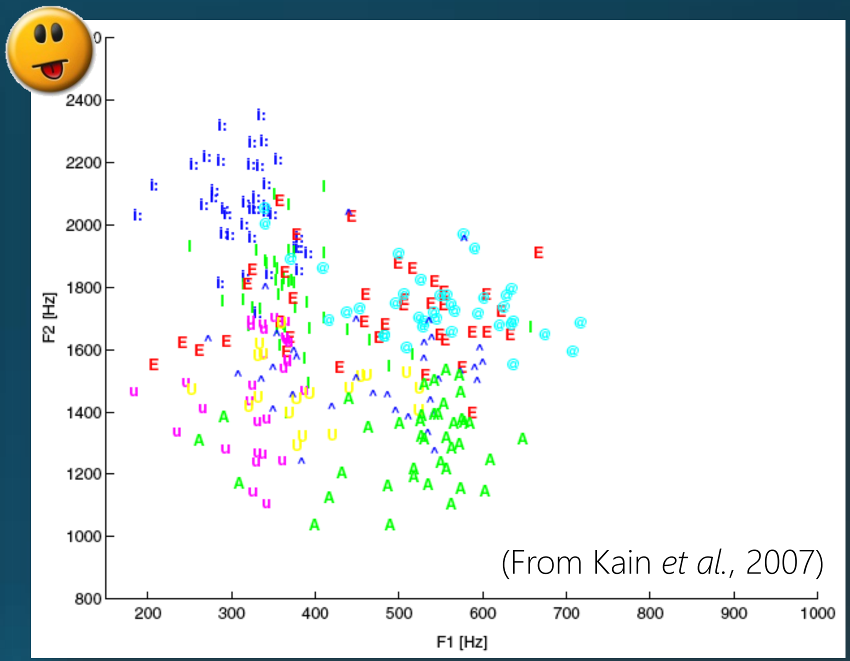
Adjusting to the individual



Acoustic ambiguity



Non-dysarthric

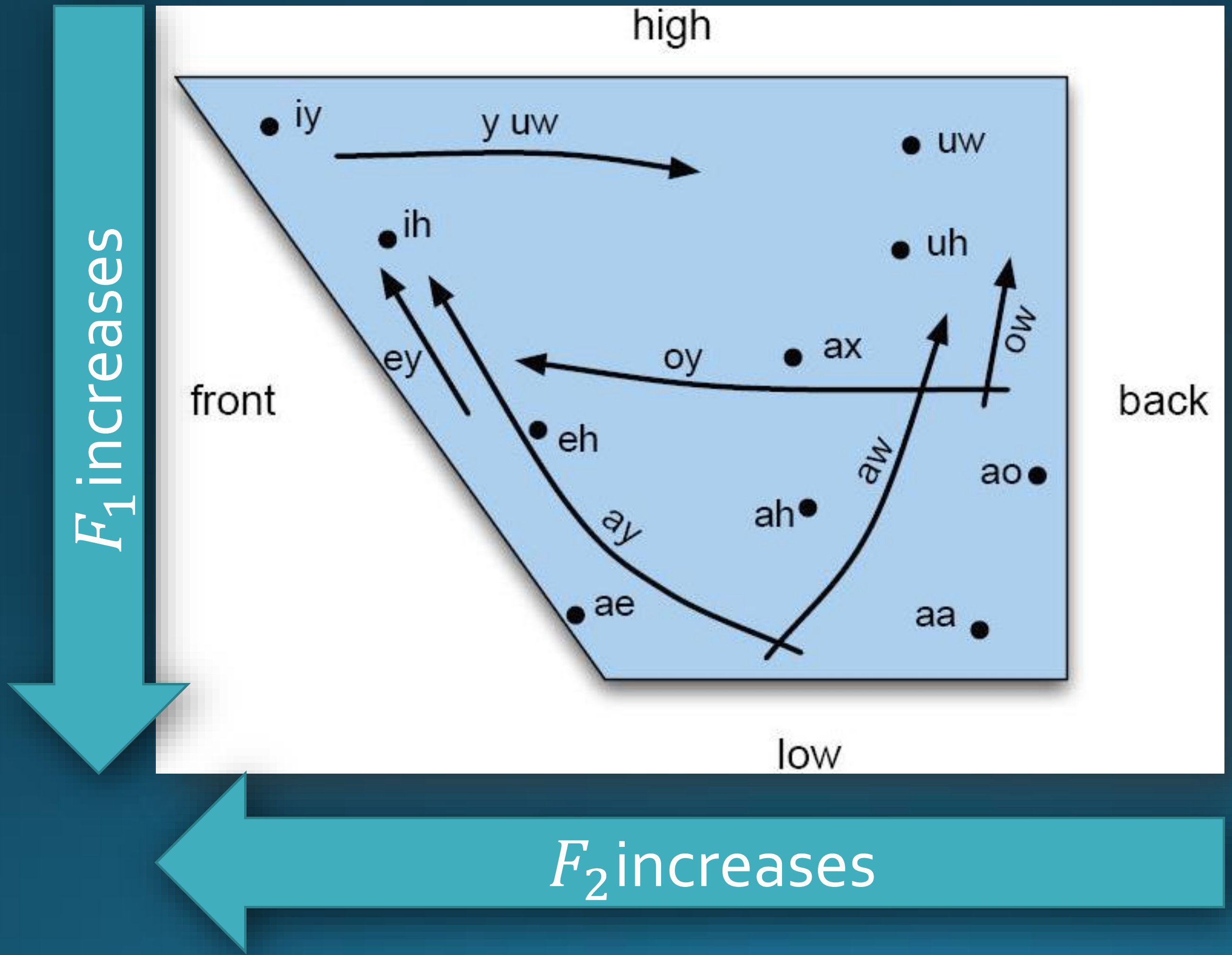


(From Kain *et al.*, 2007)

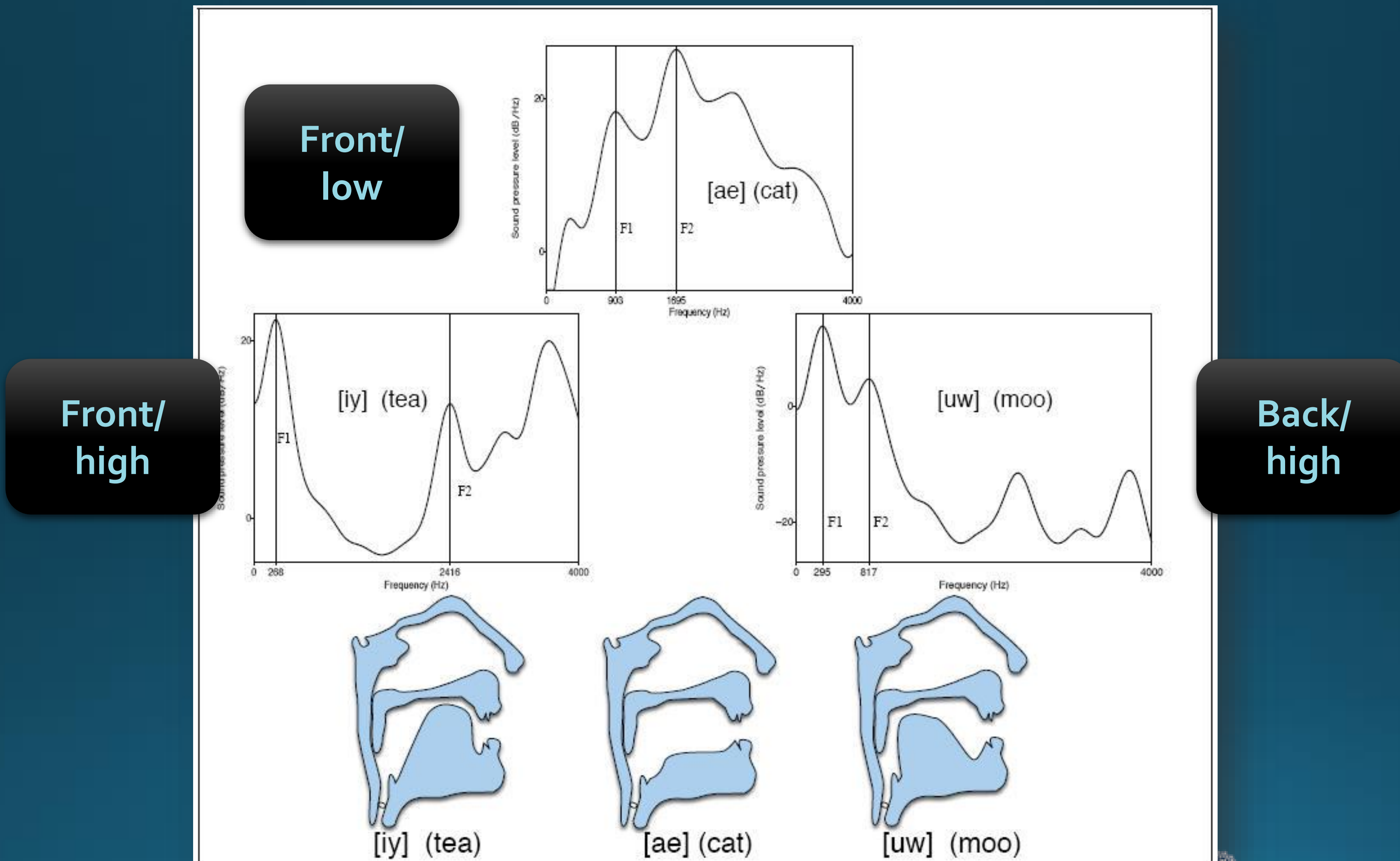
Dysarthric

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

The vowel trapezoid



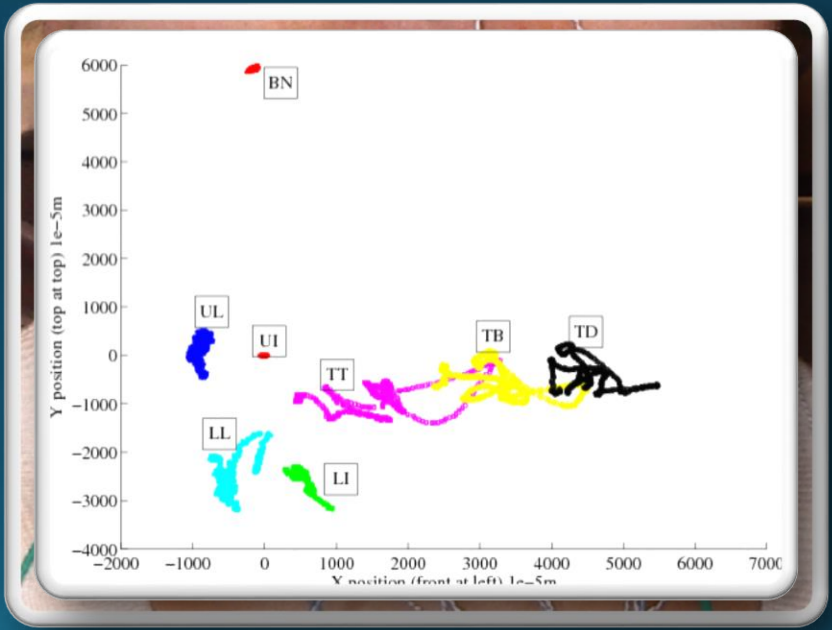
Formants and tongues



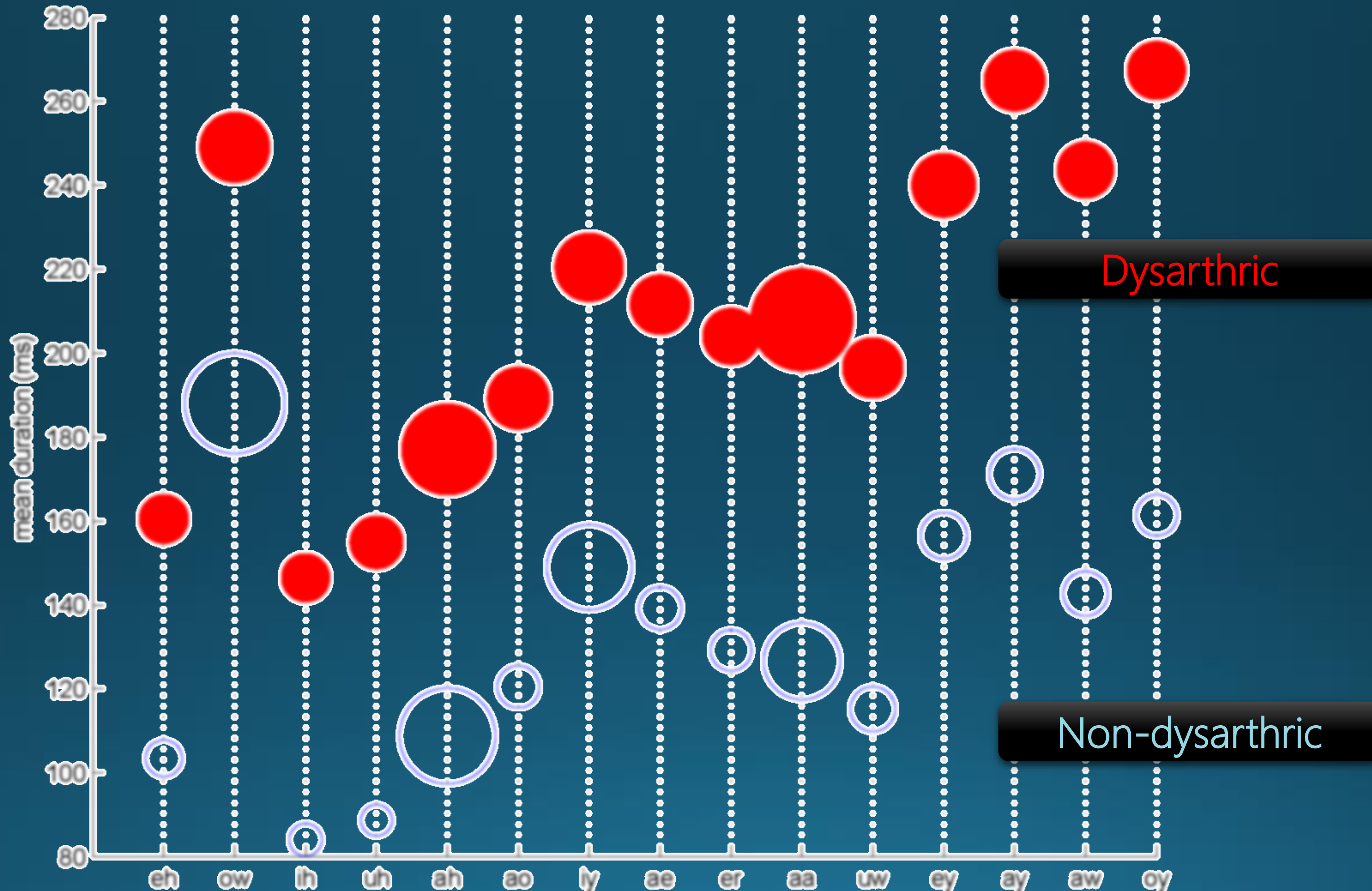
UNIVERSITY OF TORONTO

The TORGO database

- TORGO was built to train augmented ASR systems.
  - **9** subjects **with cerebral palsy**, **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.



Vowel durations in TORGO



UNIVERSITY OF TORONTO

Information in TORGO

	Speaker	$H(Acous)$	$H(Artic)$	$H(Ac \mid Ar)$
Dysarthric	M01	66.37	17.16	50.30
	M04	33.36	11.31	26.25
	F03	42.38	19.33	39.47
	Average	47.34	15.93	38.68
Control	MC01	24.40	21.49	1.14
	MC03	18.63	18.34	3.93
	FC02	16.12	15.97	3.11
	Average	19.72	18.60	2.73

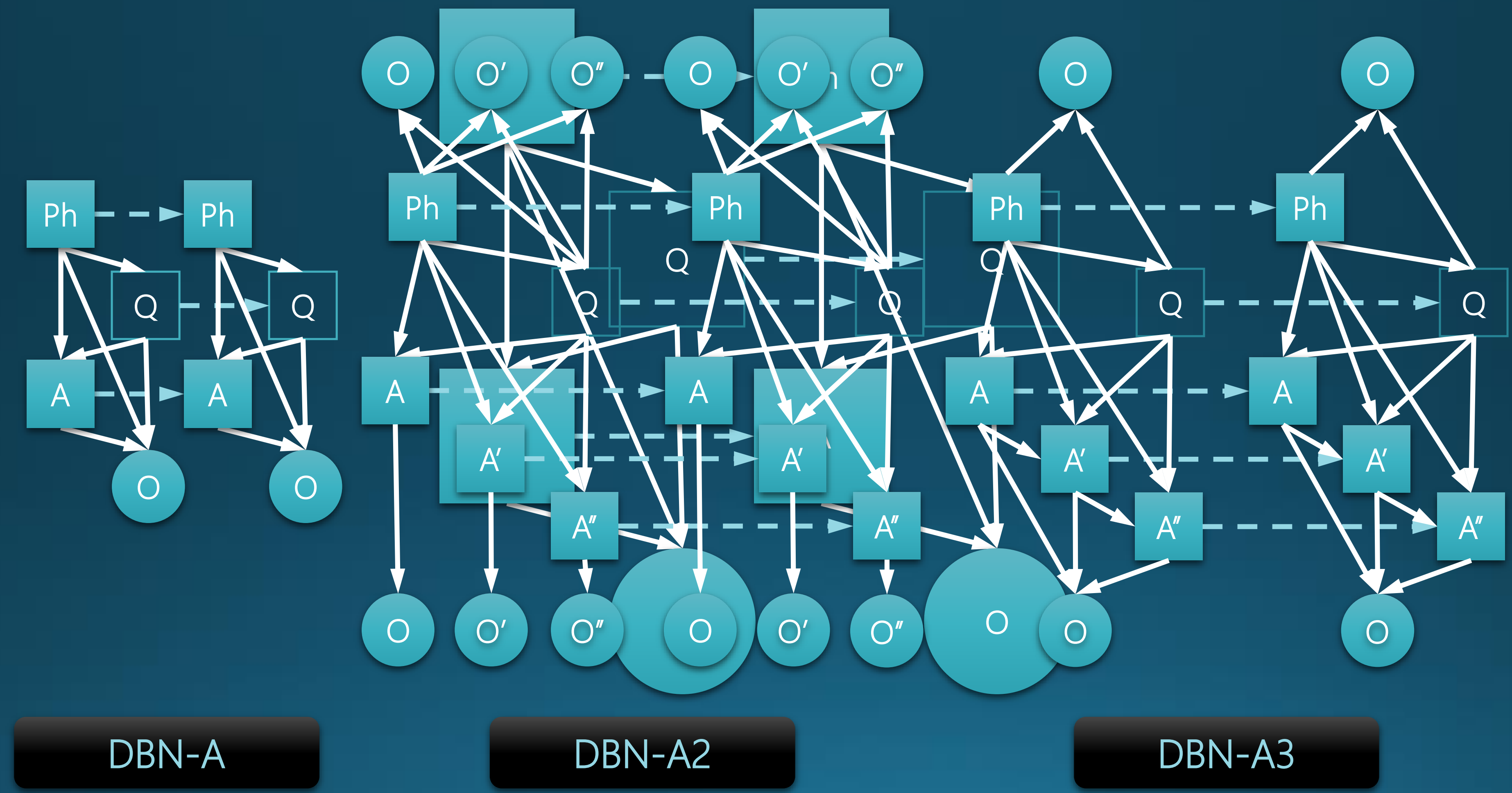
Dysarthric **acoustics** are far more statistically disordered than the control data *but*

Dysarthric **articulation** is *just as* statistically ordered as the control data *yet*

Dysarthric acoustics are far less **predictable** from articulation.

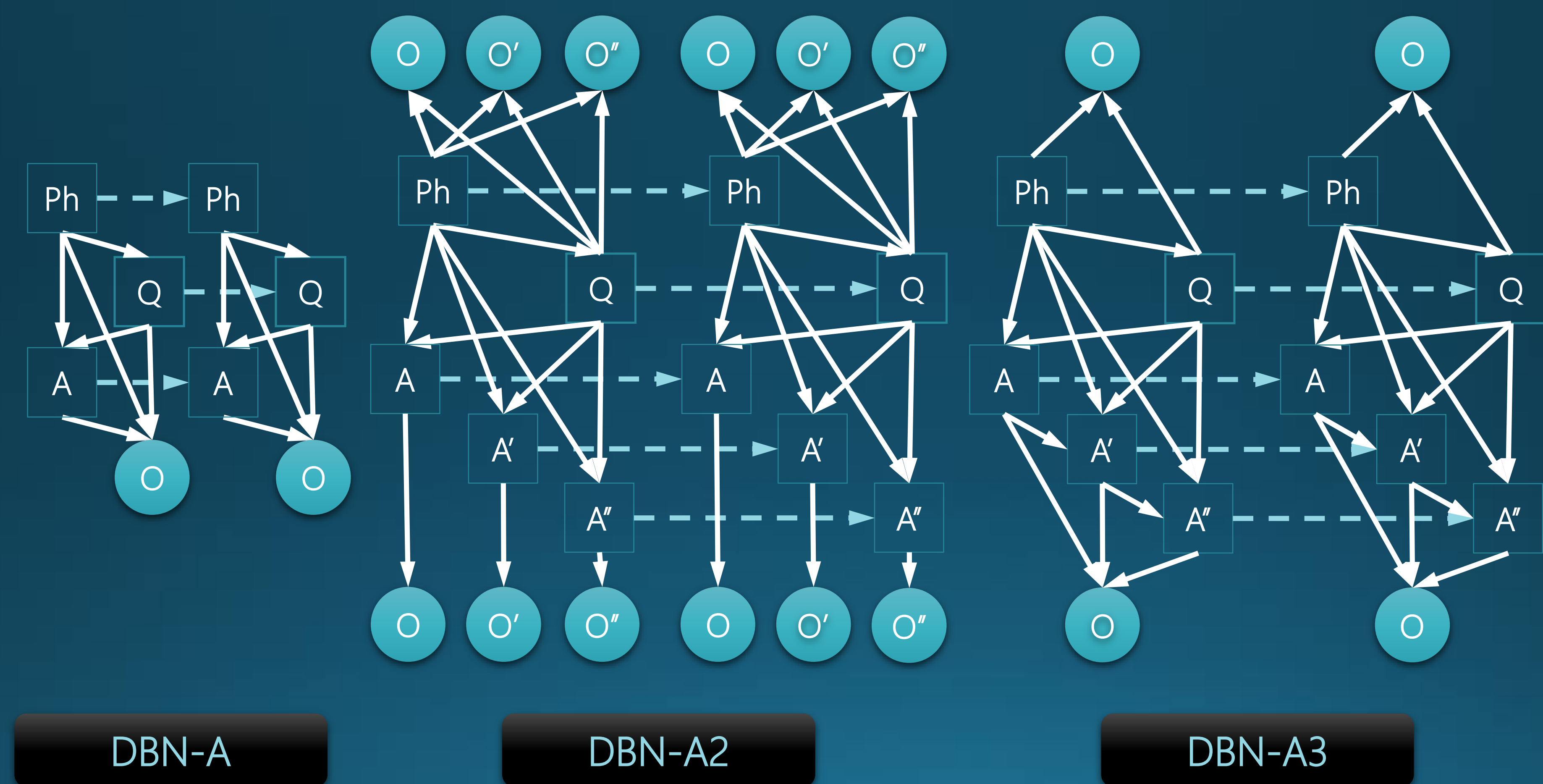
UNIVERSITY OF TORONTO

Dynamic Bayes nets and EMA



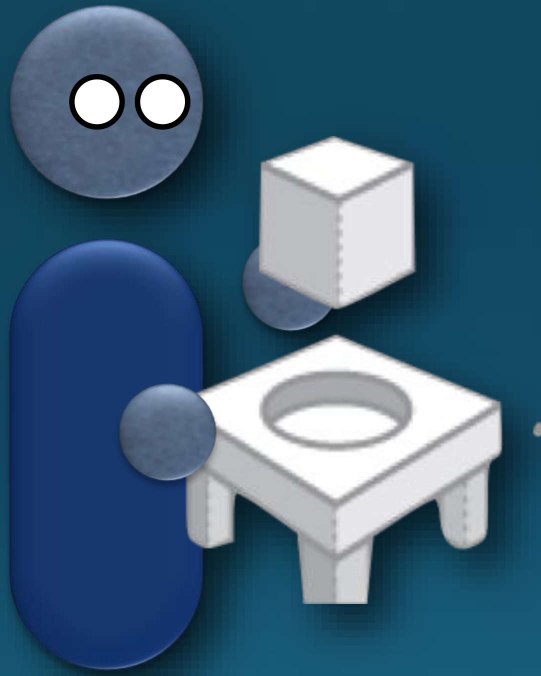
UNIVERSITY OF TORONTO

Dynamic Bayes nets and EMA



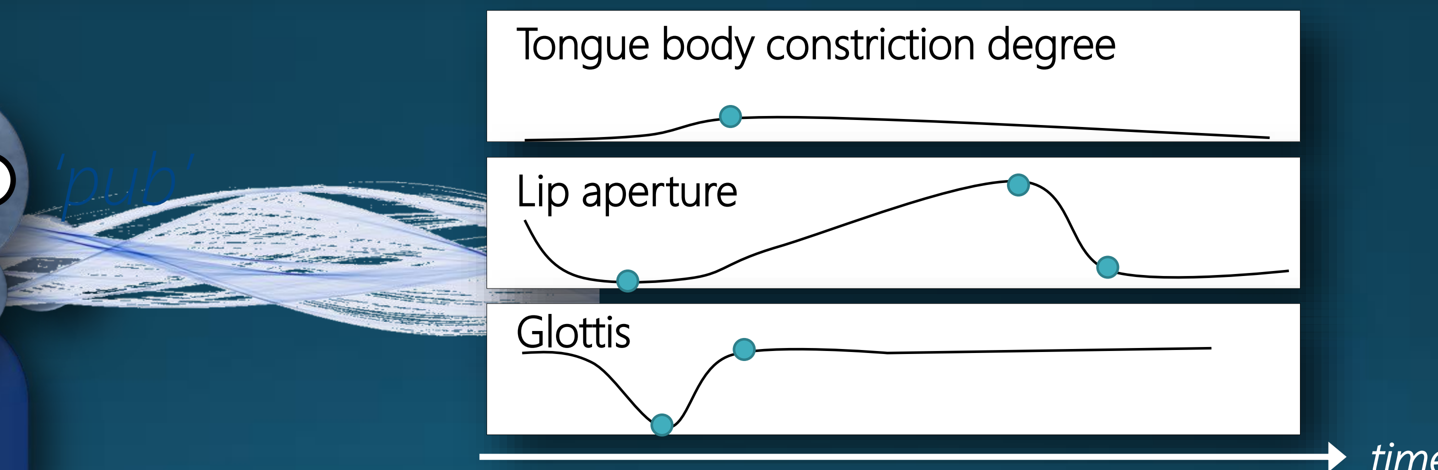
UNIVERSITY OF TORONTO

Beyond discrete articulation



Dynamic speech gestures

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



We require a **theoretical framework** to represent relevant and continuous articulatory motion.

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							
…							

Smaller vowel space might be replicable by modifying spring coefficents.

Task-dynamics:

$$\mathbf{M}z'' + \mathbf{B}z' + \mathbf{K}(z - \mathbf{z^0})$$

Aspects to consider

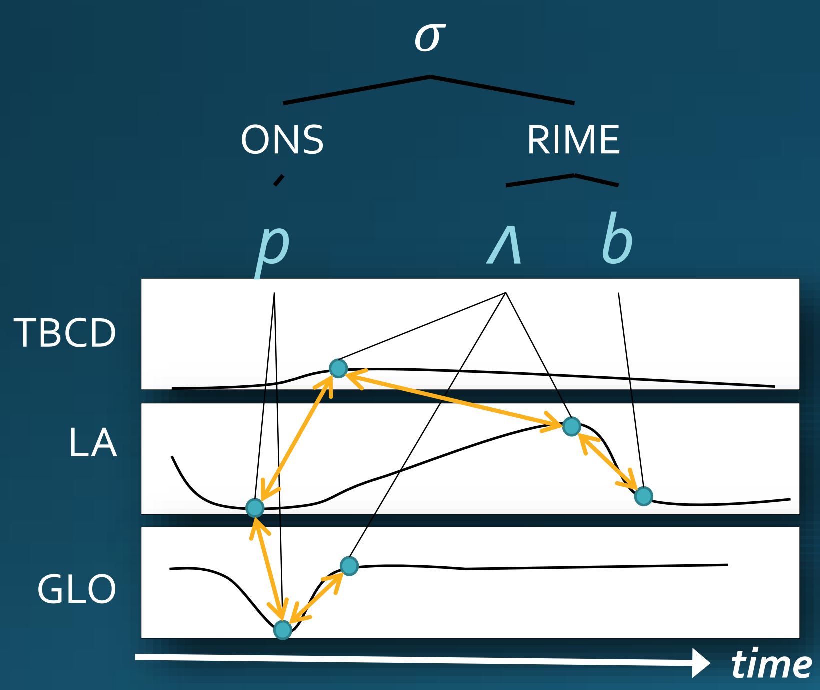
- A model of physical speech production should include:

1. **Timing**.
    a) Inter-articulator co-ordination.
    b) Rhythm.

2. **Feedback**.
    a) Acoustic, proprioceptive, and tactile.

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).

UNIVERSITY OF TORONTO

1. Timing

- Cerebellar **ataxia** often **prohibits** control over more than one articulator at a time.
  - **Apraxia** generates incorrect motor **plans**, wholly **distorting** gestural **goals**, hence timing.

- **Dysarthric** speech **nearly equally** consists of **steady-states** (49.95%) and **transitions** (50.05%) (Vollmer, 1997).
  - **Typical** speech consists of ~**82.14**% steady-states.

Ataxia *n.* lack of voluntary coordination of muscle movements, often associated with cerebellar damage.

UNIVERSITY OF TORONTO

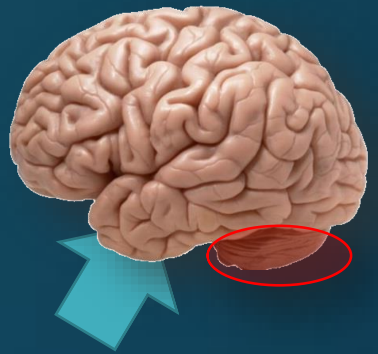
1. Timing/rhythm

- **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)

2. Feedback

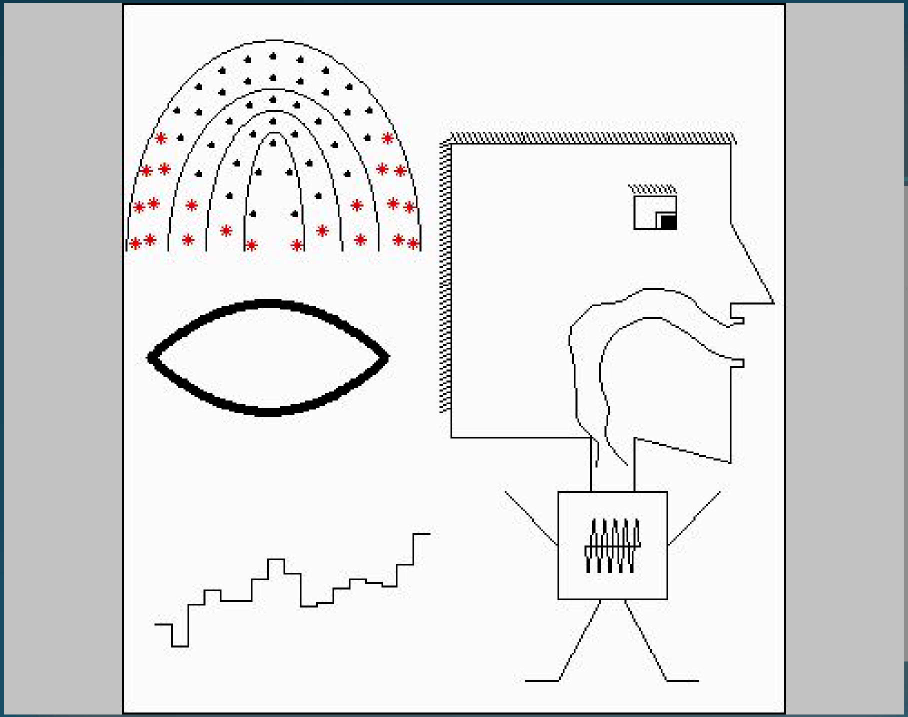


- Dysarthria can affect **sensory** cranial nerves.
- **Parkinson's disease** reduces **temporal** discrimination in **tactile**, **auditory**, and **visual** stimuli.
  - Likely explanation is that **damage** to the **basal ganglia** **prohibits** the formation of **sensory targets** (Kent *et al.*, 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**.

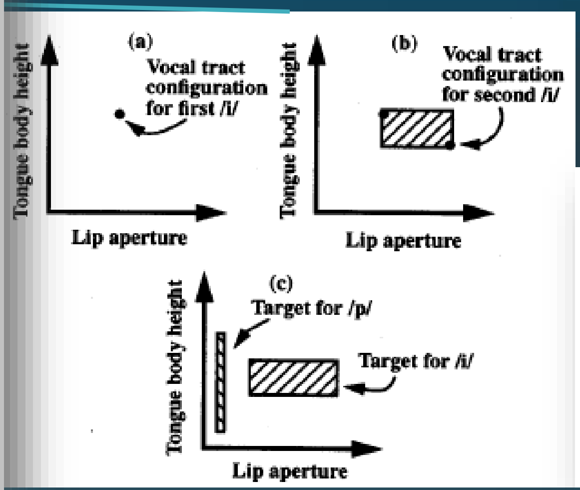
UNIVERSITY OF TORONTO

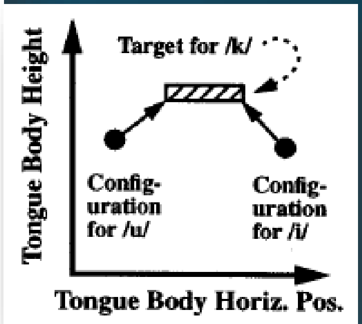
2. Feedback and DIVA

- The DIVA model is **supposed** to model feedback, but is largely **speculative** on **neurological** aspects.
- Here, **sound targets** and **somatosensory targets** are **learned** during ‘babbling’ and **modify** articulatory **goals**.

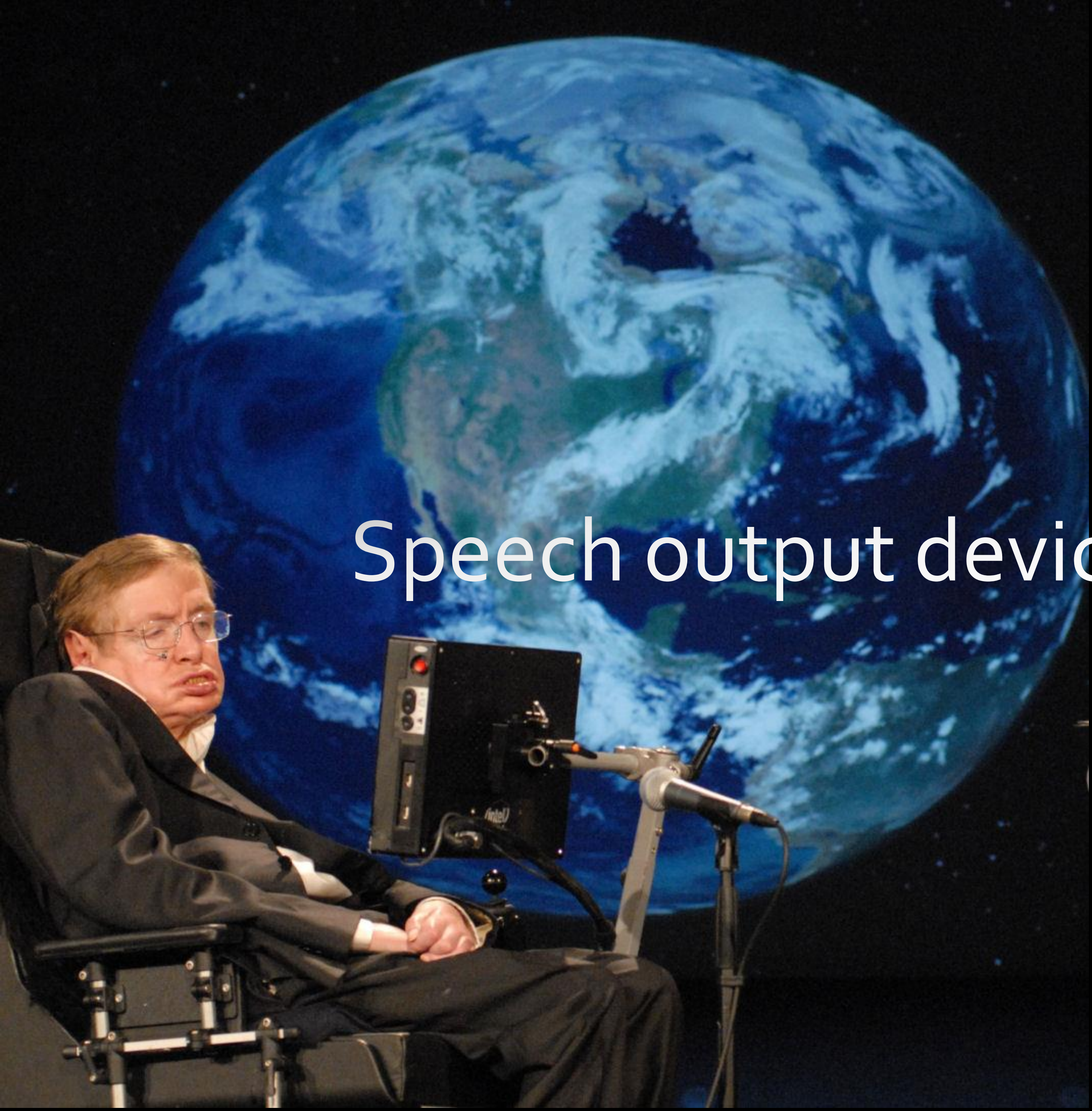


- This is meant to imitate the cerebellum (or basal ganglia).





Speech output devices

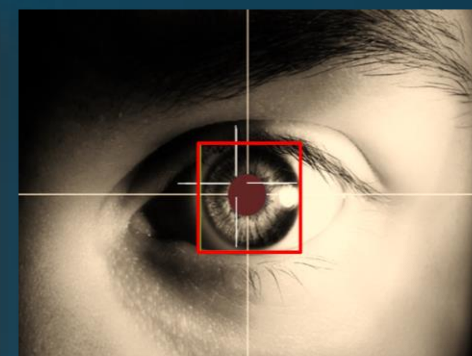


Augmentative/Alternative Communication (AAC)

- There are several 'physical' means to enter text.


Switches


Touch

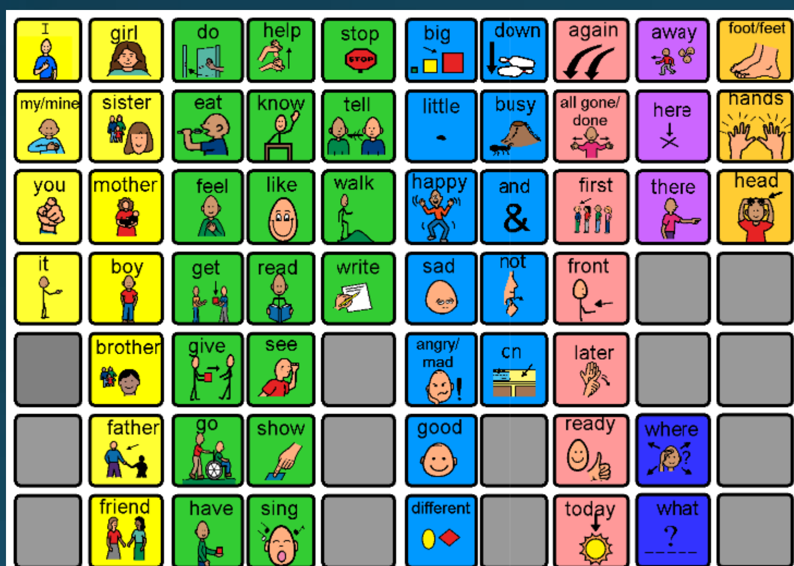

Eye

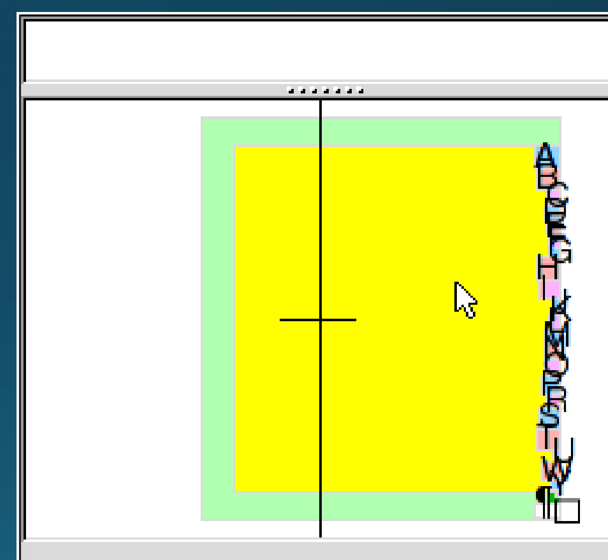
- Each can depend on the physical limits of the user.

UNIVERSITY OF TORONTO

Speech output devices

- There are several ‘soft’ means to enter text.
  - **Scanning** involves a **cursor** moving at a constant rate through an **array of symbols** until one is selected.

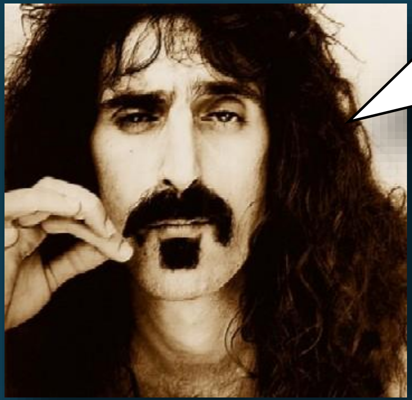




- **Word prediction** (with *N*-grams) can be invaluable.

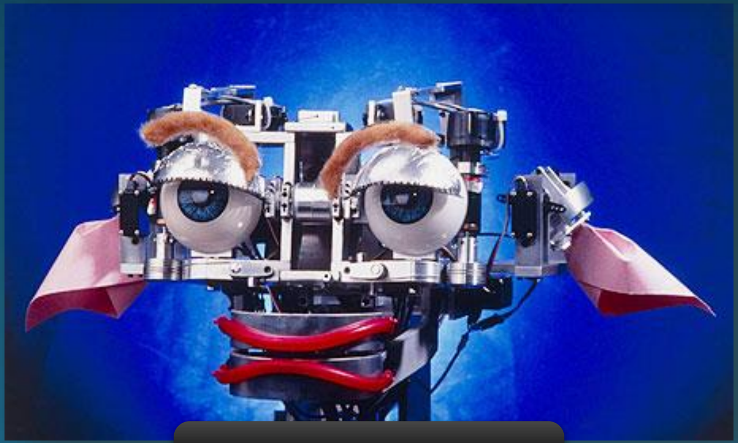
UNIVERSITY OF TORONTO

Speech output devices need to devise speech output



The computer can't tell you the **emotional** story. It can give you the exact mathematical design, but what's missing is the **eyebrows**.

Frank Zappa



Kismet

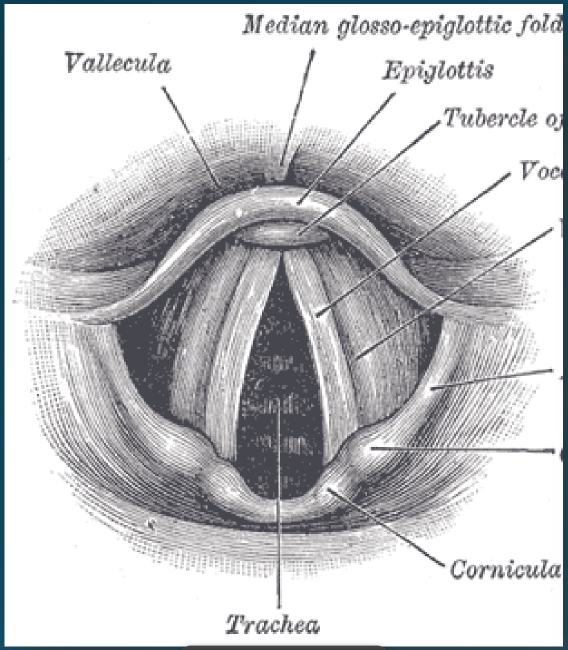
Emphasis can modify meaning

- ***I*** never said she stole my money. (Someone else said it)
- I ***never*** said she stole my money. (It never happened)
- I never ***said*** she stole my money. (I just hinted at it)
- I never said ***she*** stole my money. (Someone else stole it)
- I never said she ***stole*** my money. (She just borrowed it)
- I never said she stole ***my*** money. (She stole someone else's)
- I never said she stole my ***money***. (She stole my heart).

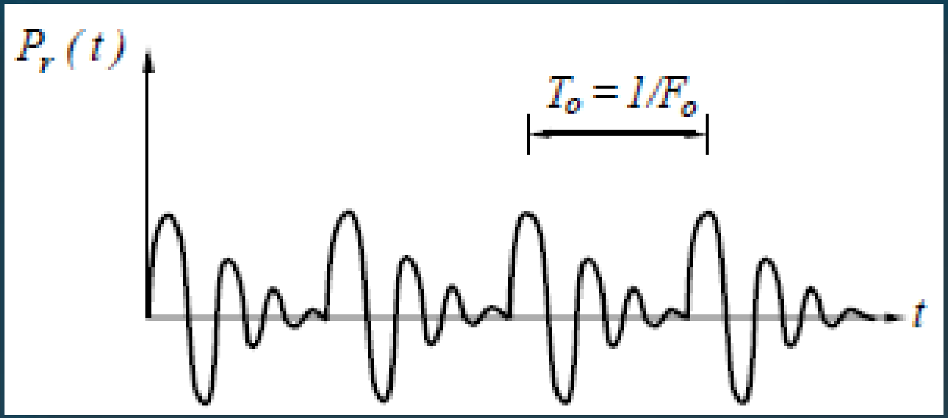
What ***is*** emphasis?

Reminder: F_0

- F_0: *n.* (**fundamental frequency**), the rate of vibration of the **glottis** – often very **indicative** of the speaker.



Glottis



	Avg F_0 (Hz)	Min F_0 (Hz)	Max F_0 (Hz)
Men	125	80	200
Women	225	150	350
Children	300	200	500

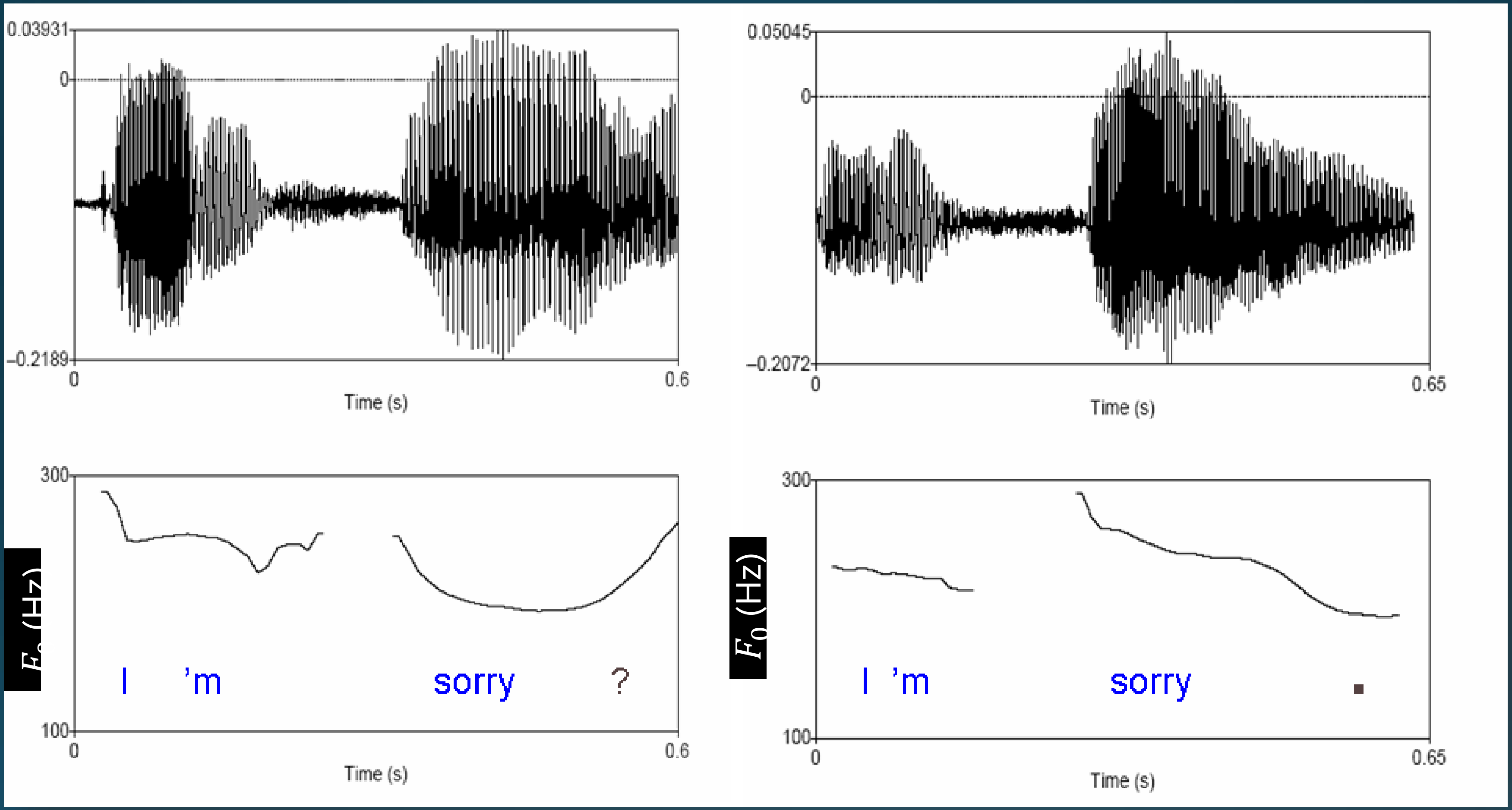
UNIVERSITY OF TORONTO

Prosody

- **Sonorant**: *n.* Any **sustained** phoneme in which the **glottis** is vibrating (i.e., the phoneme is '**voiced**').
  - Includes some consonants (e.g., /*w*/, /*m*/, /r/).

- **Prosody**: *n.* the **modification** of speech acoustics to convey some **extra-lexical** meaning:
  - **Pitch**: Changing of F_0 over time.
  - **Duration**: The length in time of sonorants.
  - **Loudness**: The amount of **energy** produced by the **lungs**.

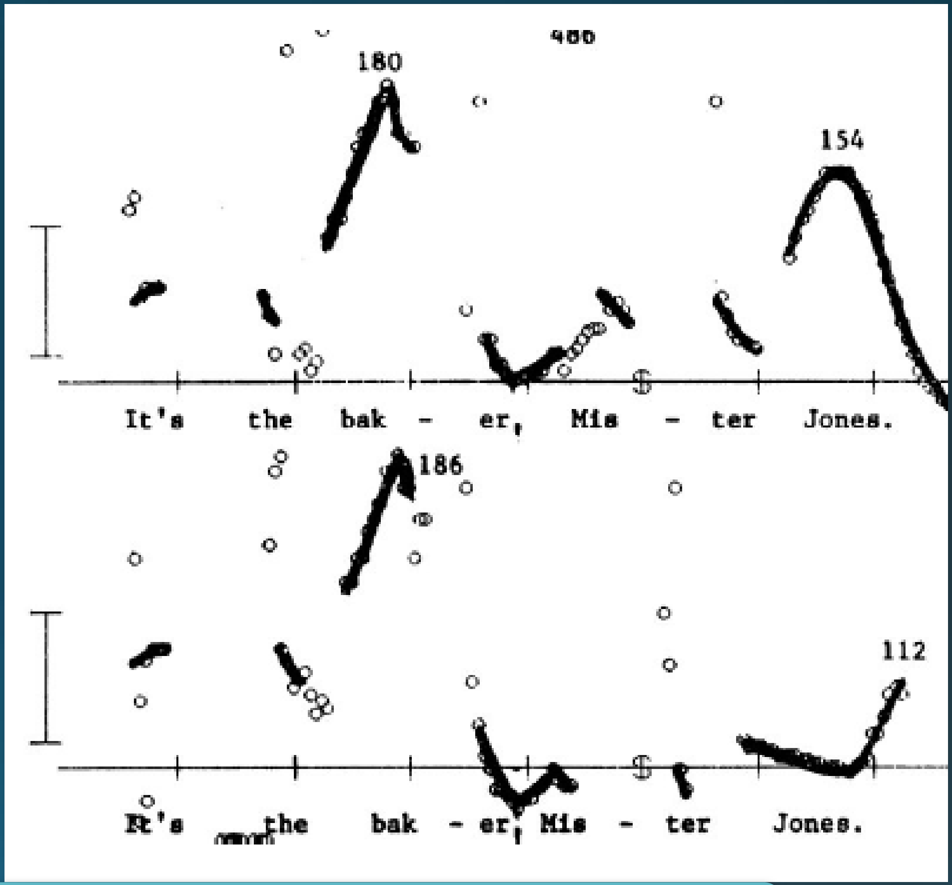
Pitch prosody



UNIVERSITY OF TORONTO

Pitch can modify meaning

- e.g., Mr. X asks you the name of the baker, whose name is 'Jones'.
- e.g., Mr. Jones asks you the profession of Mr. X.



Pitch tends to rise when uttering **novel** or important information.

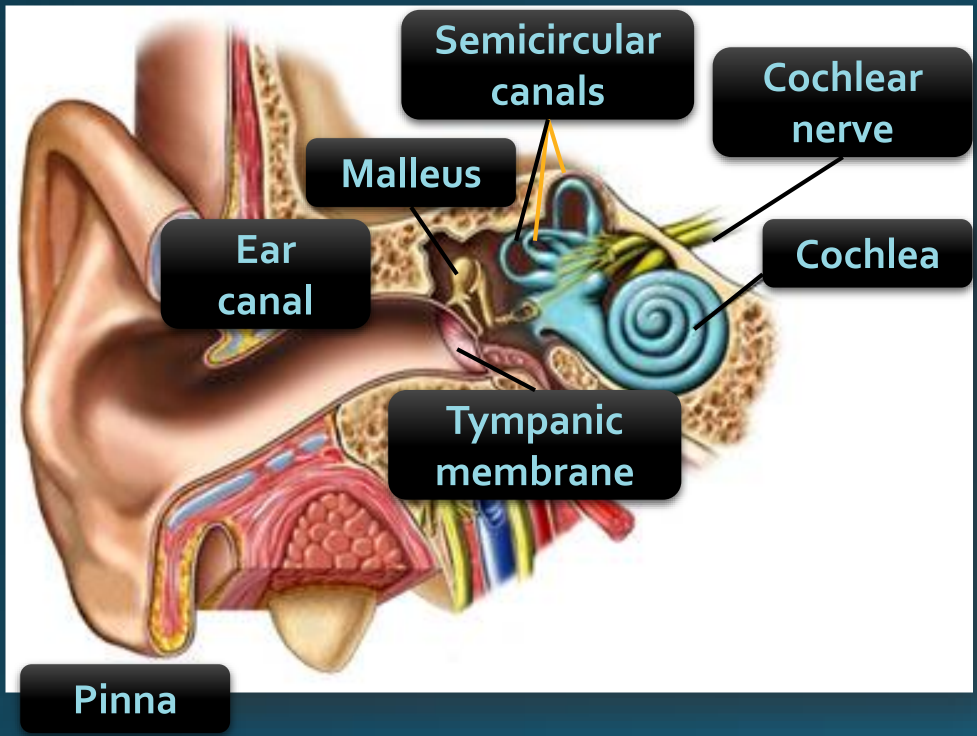
Speech output devices

- **Rate enhancement** remains a challenge.
  - In addition to **word prediction**, **semantic compaction** and **lemmatization** can increase output to ~12 words/minute.
- AAC can **improve independent speech** in children with autism or developmental delays in 89% cases (Millar *et al.*, 2006).
- Use of AAC devices **significantly improves** quality of life, including social interaction and employment.
  - >90% unemployment rate for severely disabled individuals.

Physical perception

UNIVERSITY OF TORONTO

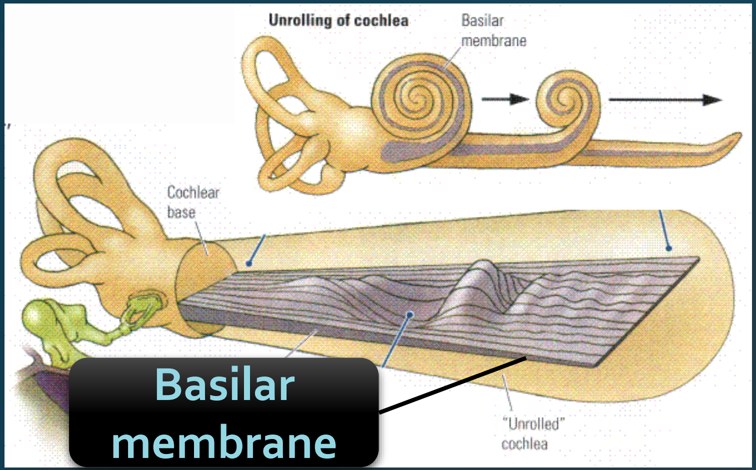
The inner ear

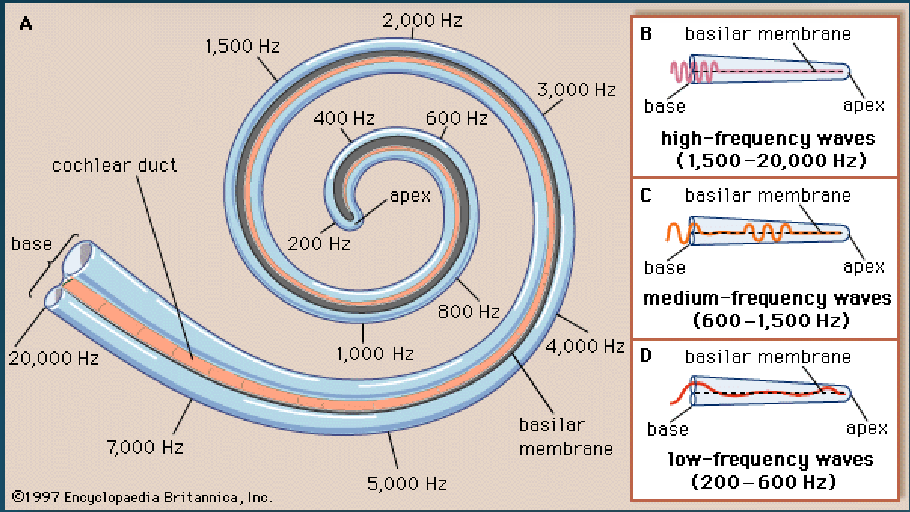


- Time-variant waves enter the ear, vibrating the **tympanic membrane**.
- This membrane causes tiny bones (incl. **malleus**) to vibrate.
- These bones in turn vibrate a structure within a shell-shaped bony structure called the **cochlea**.

UNIVERSITY OF TORONTO

The cochlea and basilar membrane





- The **basilar membrane** is covered with tiny hair-like nerves – some near the **base**, some near the **apex**.
- **High** frequencies are picked up near the base, **low** frequencies near the apex.
- These nerves fire when activated, and communicate to the brain.

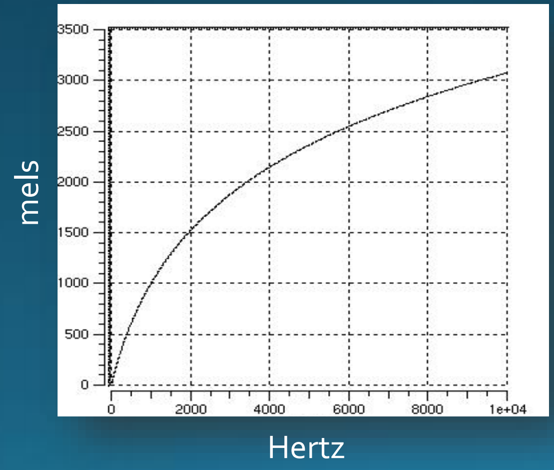
UNIVERSITY OF TORONTO

The Mel scale

- Human hearing is **not** equally sensitive to **all** frequencies.
  - We are **less** sensitive to frequencies > 1 kHz.

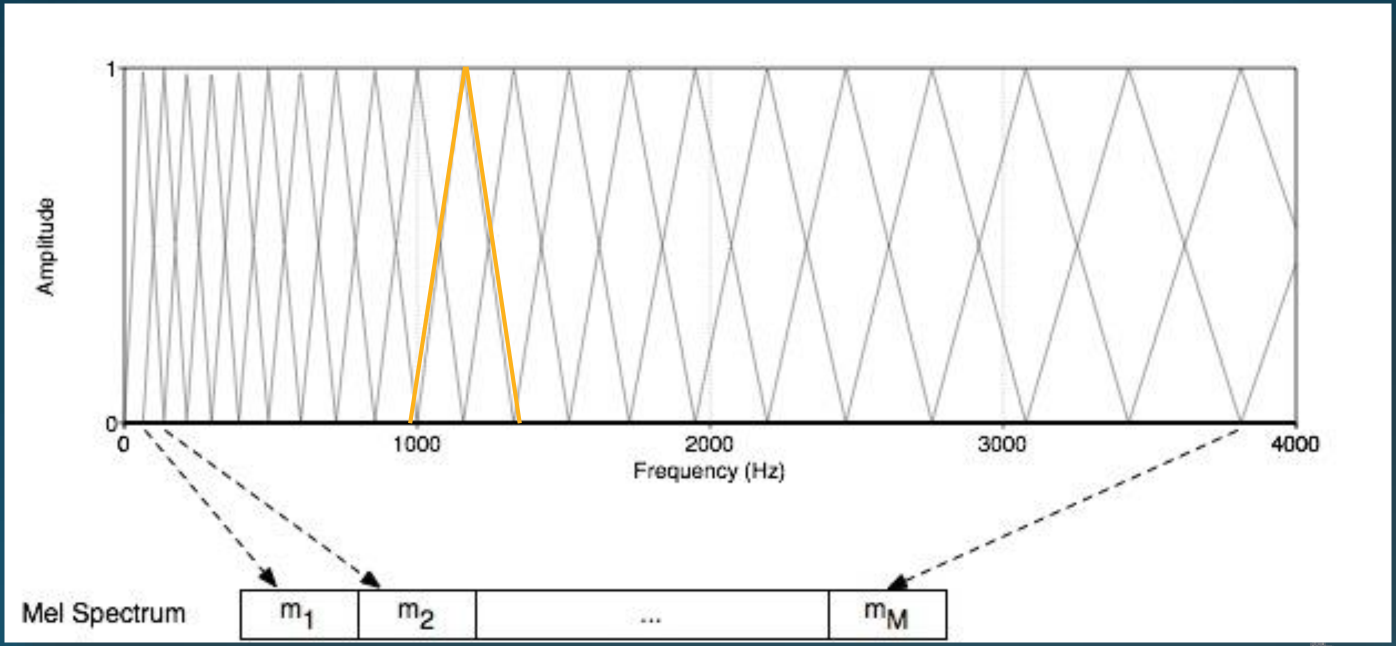
- A **mel** is a unit of pitch. Pairs of sounds which are **perceptually** equidistant in pitch are separated by an equal number of **mels**.

$$Mel(f) = 2595 \log_{10}\left(1 + \frac{f}{700}\right)$$



The Mel scale filter bank

- To mimic the response of the **human ear** (and because it often **improves** speech recognition), we often discretize the spectrum using M triangular **filters**.
  - **Uniform** spacing before 1 kHz, **logarithmic** after 1 kHz



Problems of physical perception

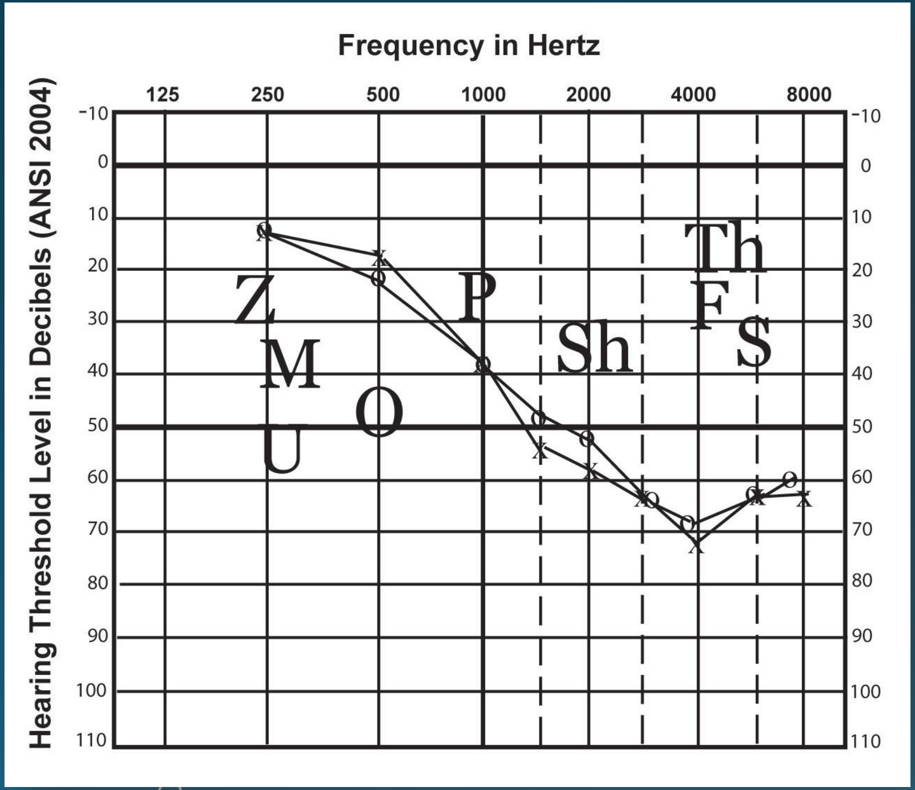
- 0.1% of children are born with **pathological hearing loss**, including auditory nerve damage.
- ~33% of adults over 60 have **acquired hearing loss**.

- **Conductive** deafness interferes with sound to the inner ear.
- **Sensorineural** deafness involves the auditory nerve itself.

- **Tinnitus** involves noise (e.g., pulsing, hissing, ringing) that can be acute and debilitating.

Assessing physical perception

- **Otologists** and **audiologists** administer audiograms, which measures hearing loss across tones (and words) at various frequencies and amplitudes.



UNIVERSITY OF TORONTO

Overcoming physical perception

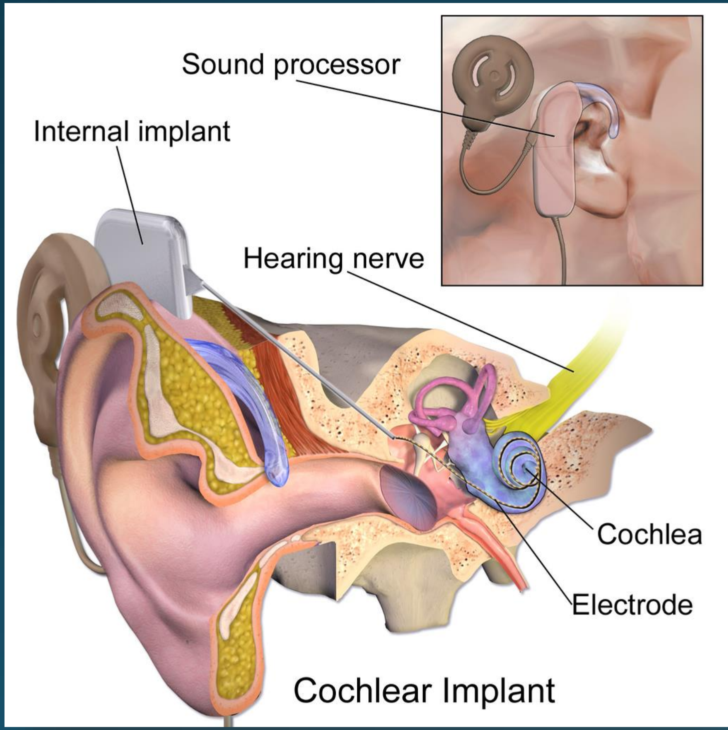
- **Hearing aids** usually **amplify** sound in certain frequencies.

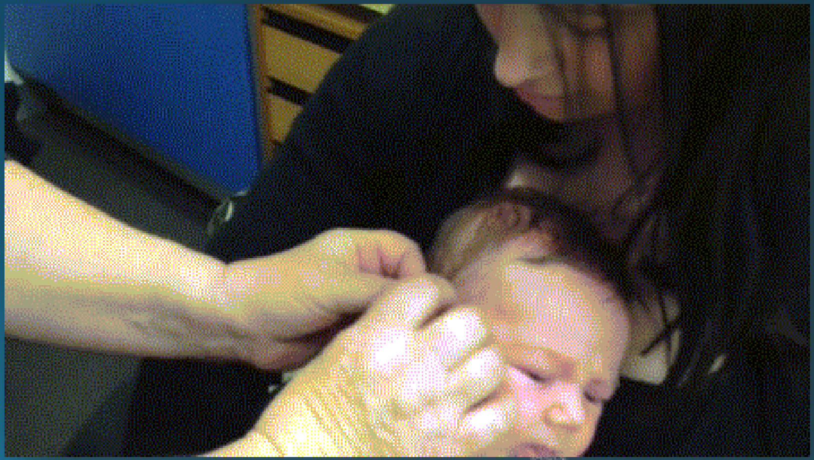
- Issues include:
  - **Occlusion effect** where person perceives "hollow" or "booming" echo-like sounds of their own voice caused by reverberations that normally pass *out* of the open air canal.
  - **Lombard effect** where people modify their own voice to compensate.
  - **Compression effect** where louder sounds need to be 'capped' to avoid further hearing damage.

UNIVERSITY OF TORONTO

Overcoming physical perception

- **Cochlear implants** replace the basilar membrane and stimulate the auditory nerve directly.

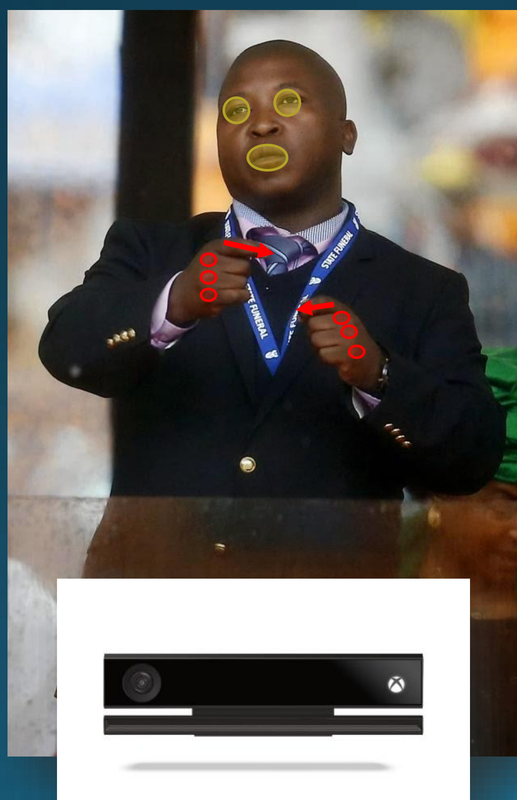




UNIVERSITY OF TORONTO

Overcoming physical perception

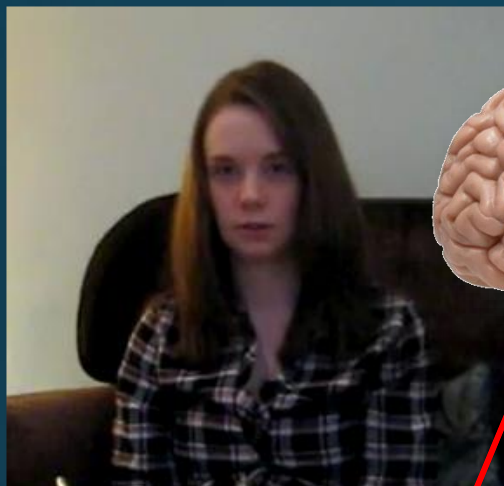
- **Sign language** interpreted by **vision-processing software**.
  - Inexpensive devices like the **Kinect** can do advanced finger and face tracking.

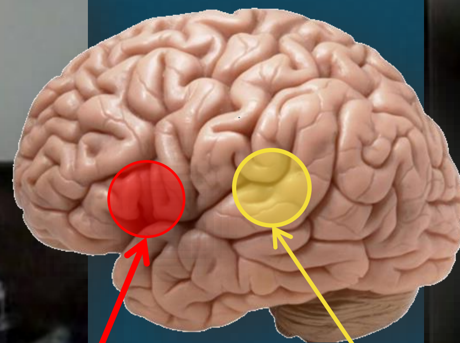


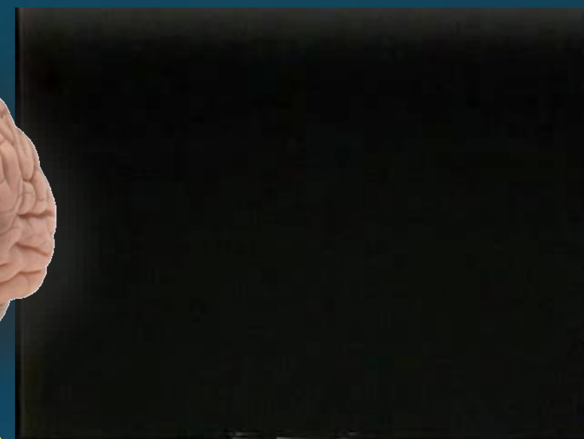
- **Subtitles** automated with **ASR**.
  - An **automated transcriber** must **reduce lexical content** while **preserving semantic content** to fit the timeframe of movie dialogue.


Cognitive issues

Deeper into the brain – Aphasia







Broca's aphasia

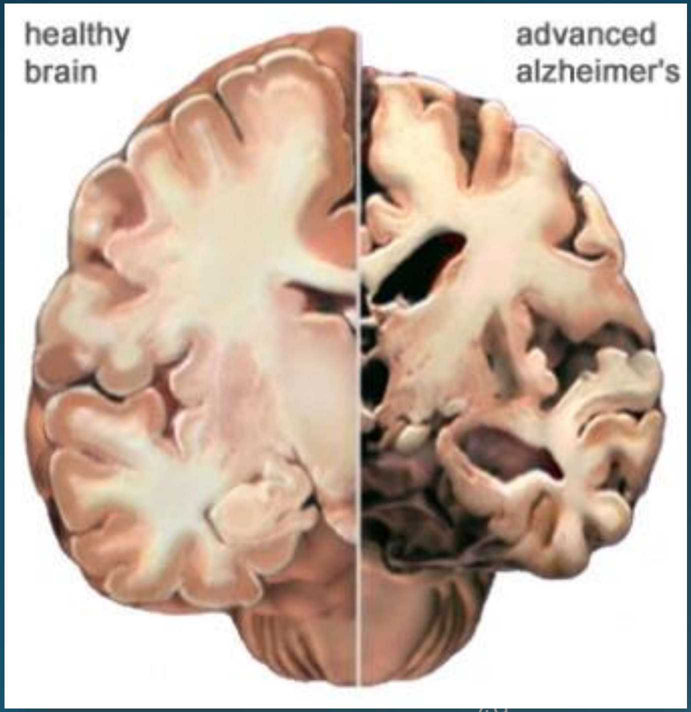
Wernicke's aphasia

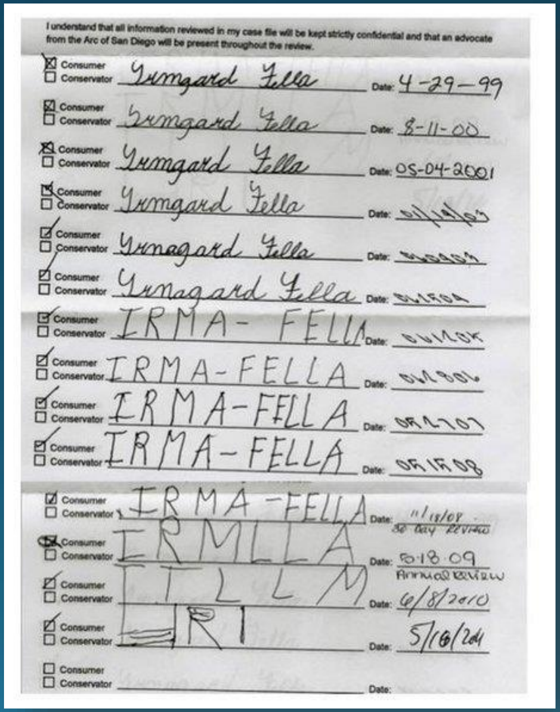
- **Reduced** hierarchical **syntax**.
- **Anomia**.
- **Reduced** "mirroring" between **observation** and **execution** of **gestures** (Rizzolatti & Arbib, 1998).

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

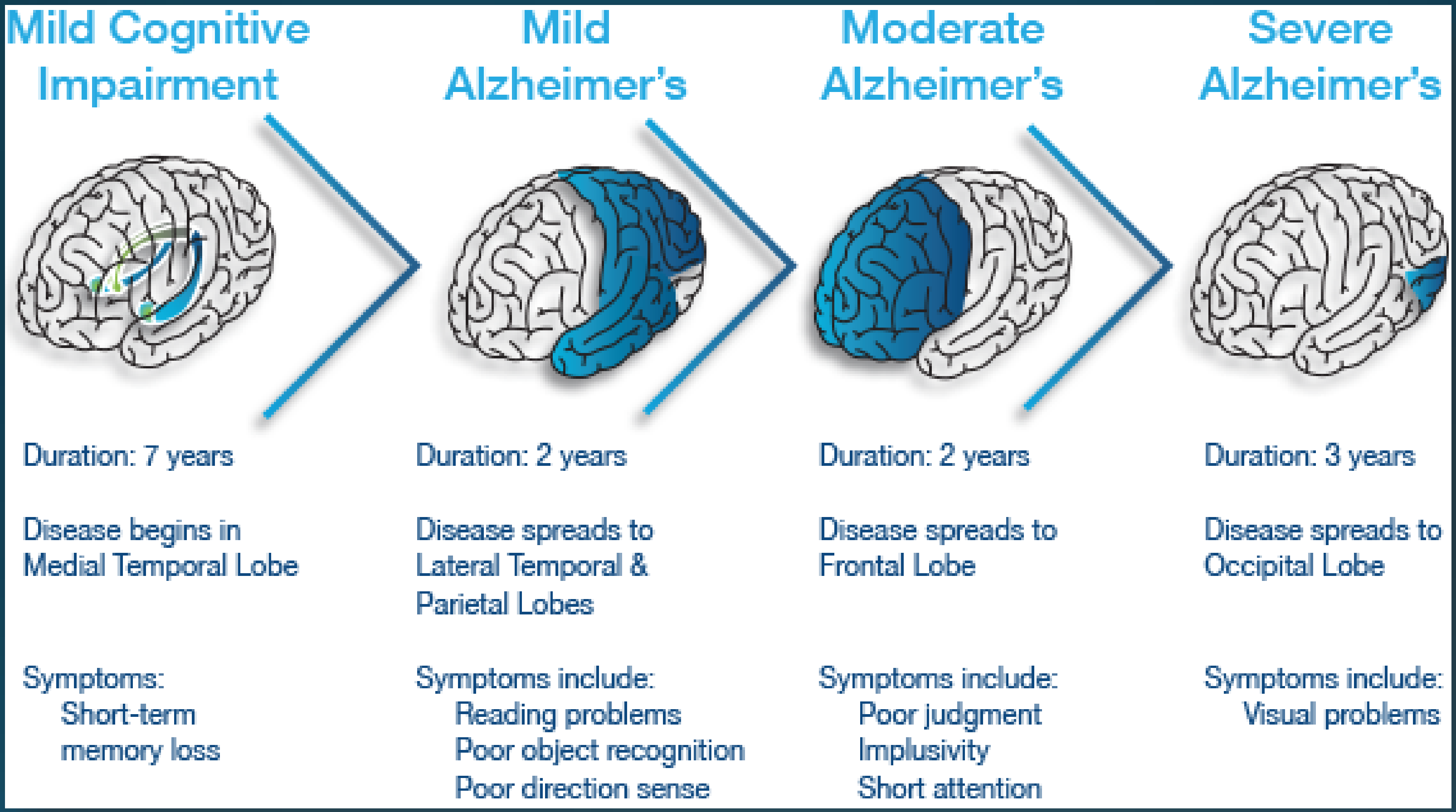
Alzheimer’s disease

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





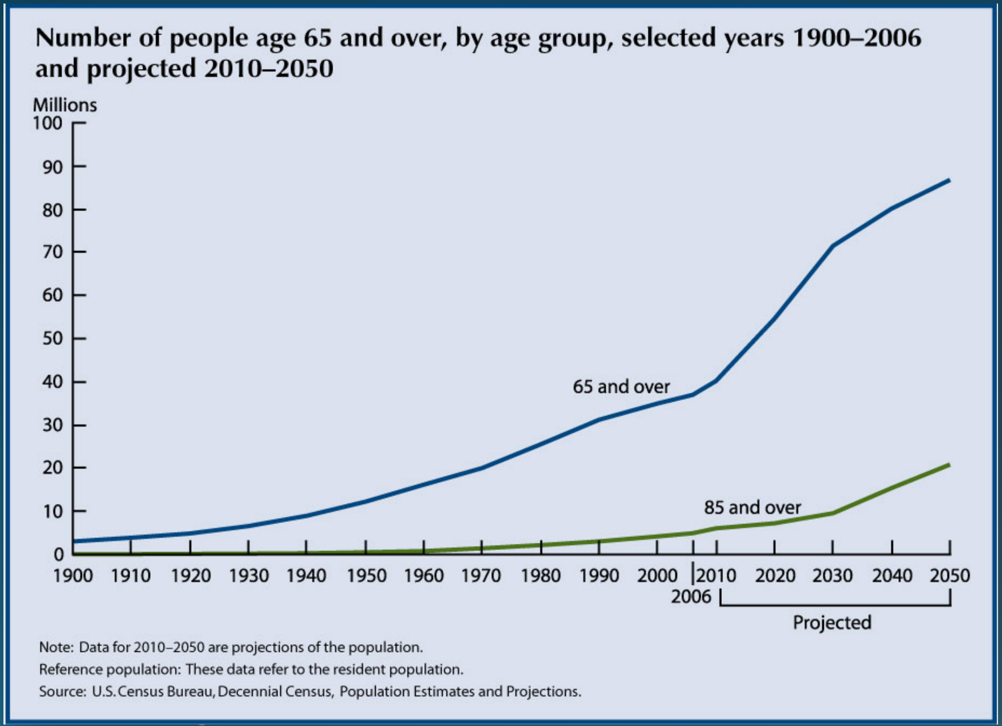
Alzheimer's disease progression



UNIVERSITY OF TORONTO

Demographic crisis

- **Caregivers** often assist individuals with AD, either at **home** or in **long-term care facilities**.
  - **>$100B** are spent annually in the U.S. on caregiving AD.
  - As the population ages, the incidence of AD may **double** or **triple** in the next decade (Bharucha *et al.*, 2009).



UNIVERSITY OF TORONTO

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.

Language in AD and dementia

- Common features in dialogue in AD: *Repetition*, *incomplete words*, and *paraphrasing* (Guinn and Habash, 2012).
  - *Pauses*, *filler words*, *formulaic speech*, and *restarts* were **not**.
    - Surprisingly, this seems to contradict Davis and Maclagan (2009), and Snover *et al.* (2004).

- Effects of AD on *syntax* remains controversial.
  - **Agrammatism** could be due to **memory deficits** (Reilly *et al.*, 2011).

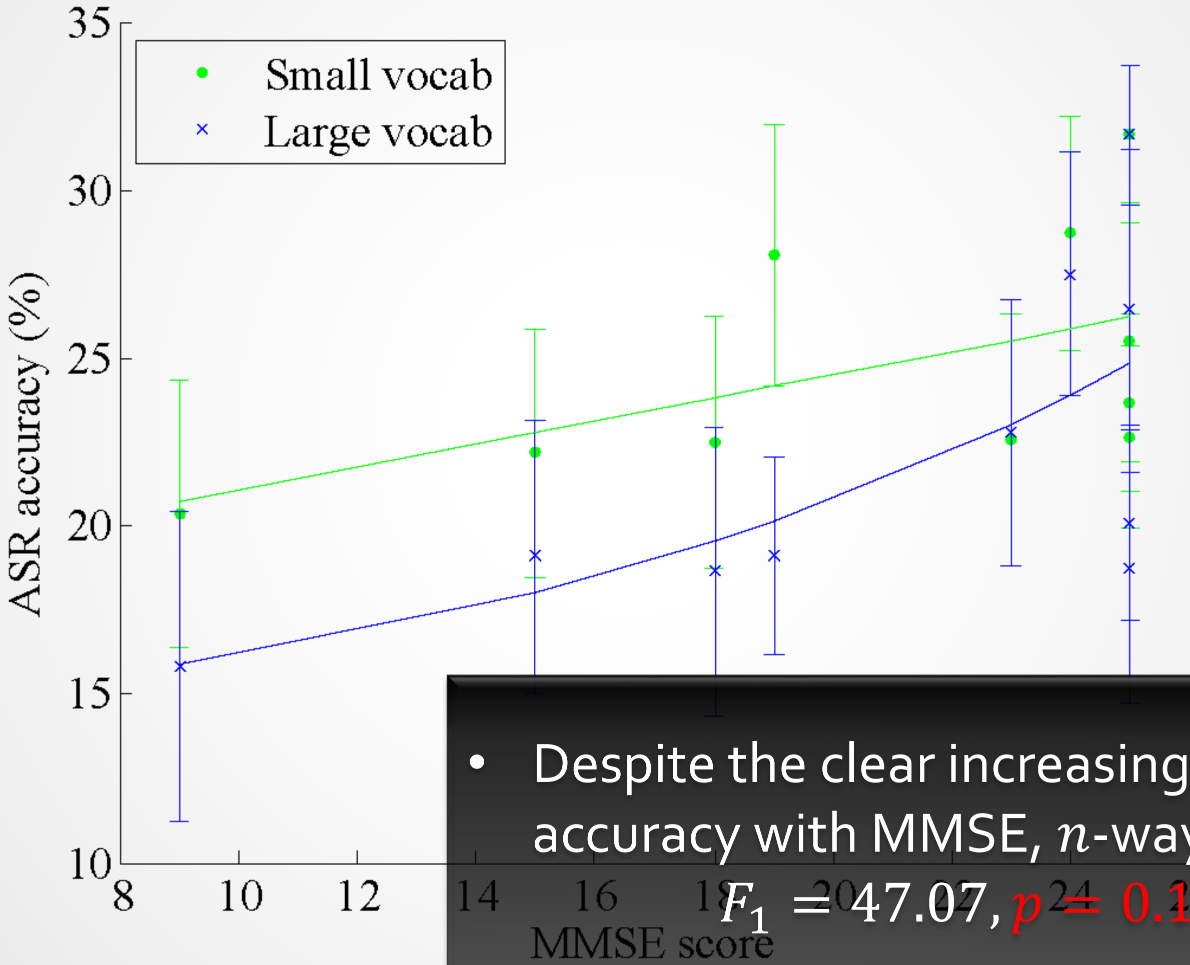
- Pakhomov *et al.* (2010) found *pause-to-word* and *pronoun-to-noun ratios* were discriminative of frontotemporal lobar degeneration.

- Roark *et al.* (2011) found *pause frequency* and *duration* were indicative of mild cognitive impairment.

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: $F_1 = 47.07$, $p = 0.164$

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

Communication strategies

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

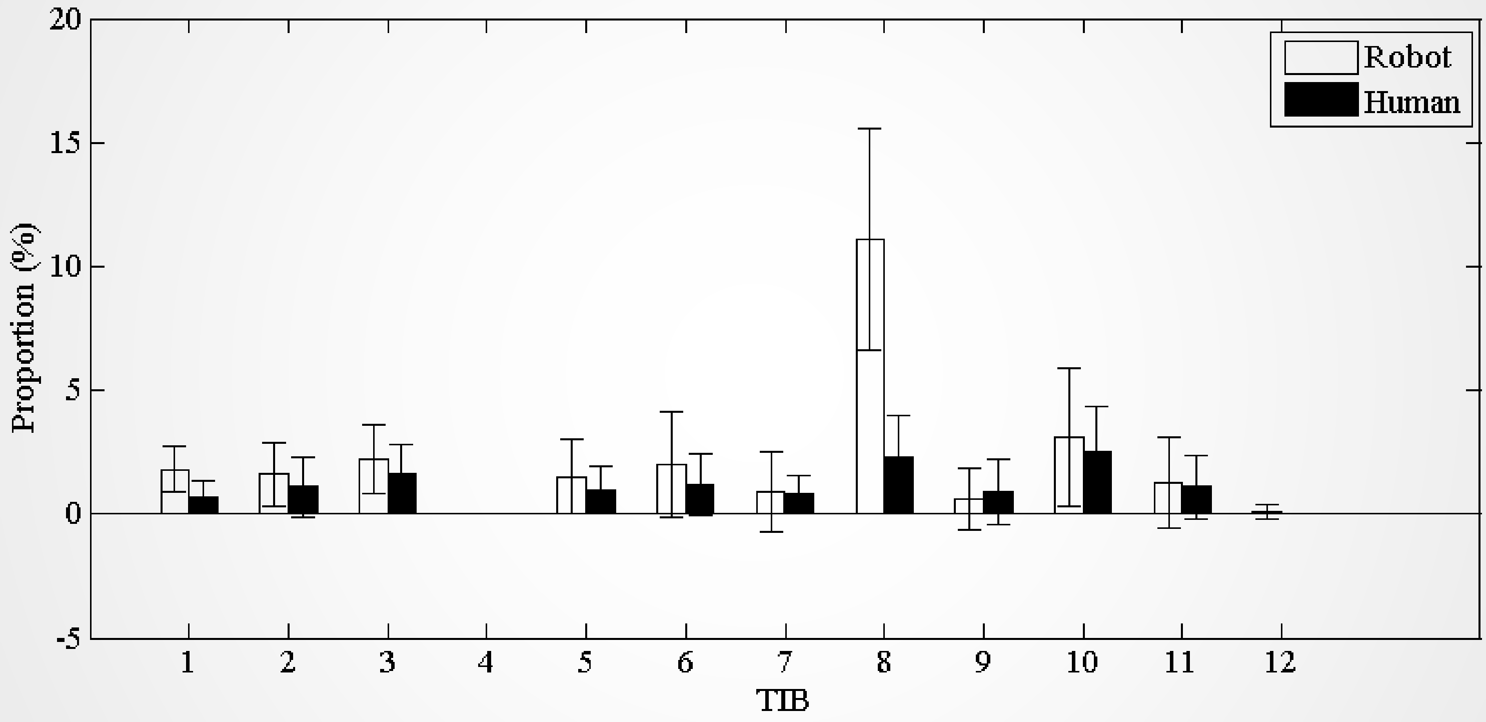
- Caregivers are commonly trained to use **communication strategies** (Small et al., 2003) , such as:
  - Using a **relatively slow** rate of speech,
  - **Repeating** misunderstood prompts **verbatim,**
  - Posing **closed-ended** questions (e.g., yes/no questions),
  - **Simplifying** the **syntactic complexity** of sentences,
  - Giving one question or **one direction at a time**, and
  - Using pronouns minimally.

How to identify breakdowns?

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

- >33% of moderate AD dyads display '**trouble-source repair**', which is related to TIB (Orange, Lubinsky, Higginbotham, 1996).
  - **Most common trouble**: discourse (e.g., inattention, working memory)
  - **Most common repair**: *wh*-questions and hypotheses (e.g., "*Do you mean …?*").

How to identify breakdowns?

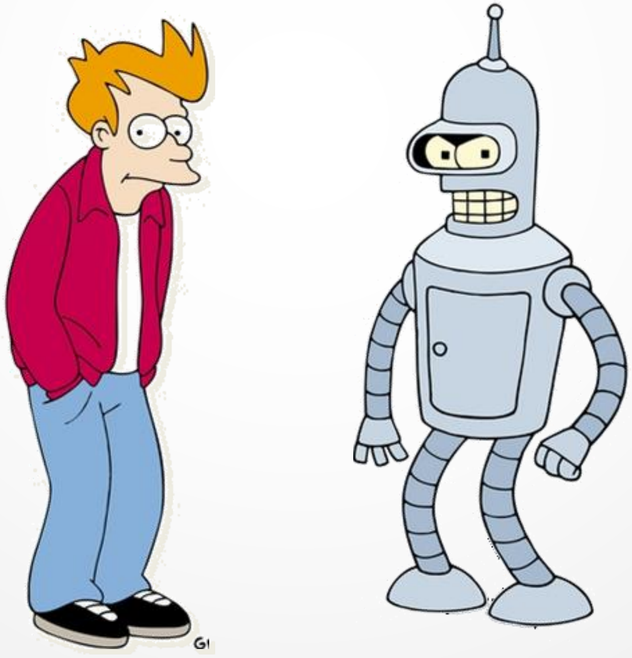


- People with AD were much ($t(18) = -5.8, p < 0.0001$) more likely to exhibit **TIB 8 (lack of uptake)** with the robot …

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

How to identify breakdowns?

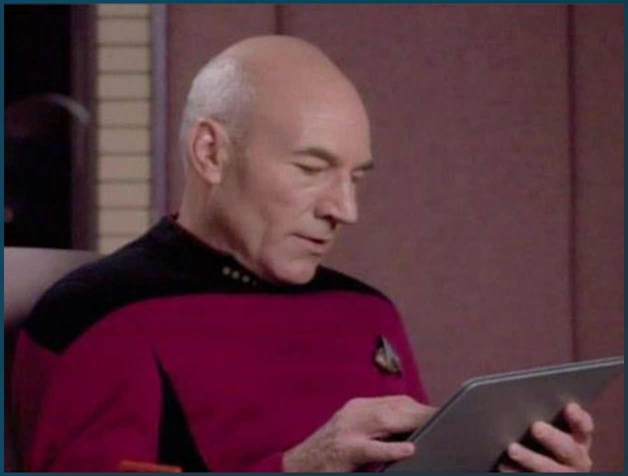
- … people with AD were much more likely ($t(18) = -4.78$, $p < 0.0001$) to have **successful** interactions with a **robot** (18.1%) than with a non-familiar **human** (6.7%).



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

Interfaces for automated dialog

- Are **alternative modes** appropriate?
  - e.g., could a digital assistant be useful on **tablets** or on the **TV**?
  - How do we **measure success**? Engagement? Emotion?



- Can these systems be doing something *else* in the 'background'?

UNIVERSITY OF TORONTO

Diagnosing language disorders

- **Recent work** aims to **diagnose language disorders**. E.g.,
  - primary progressive aphasia and its subtypes, and
  - **Parkinson's disease**.

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

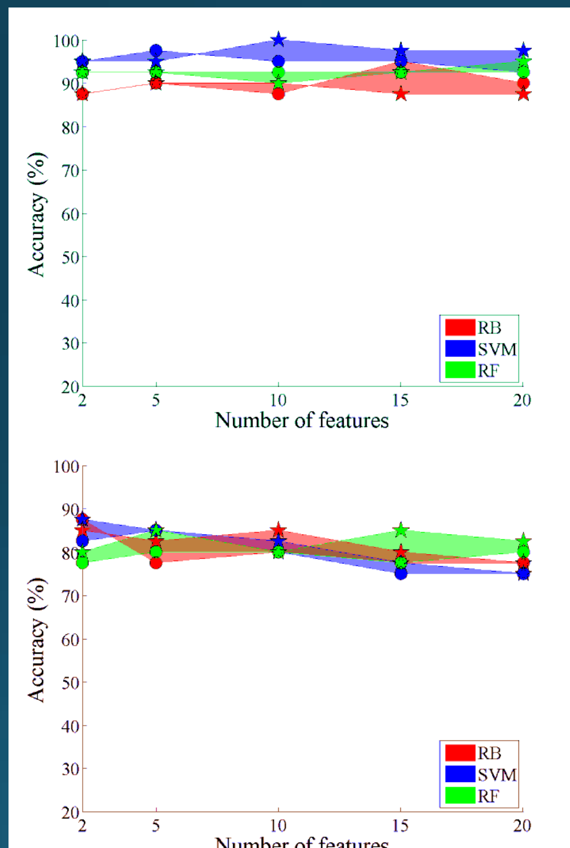
UNIVERSITY OF TORONTO

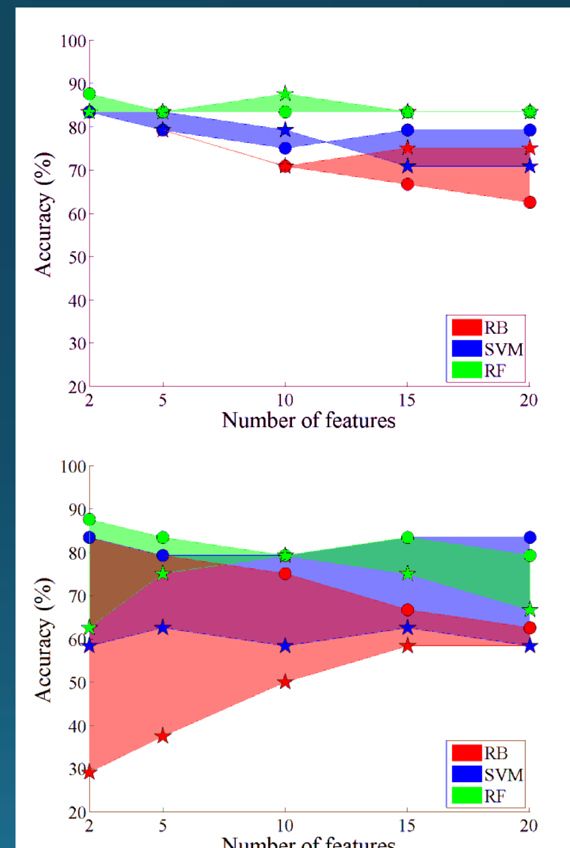
Diagnosing language disorders

PPA vs. CTRL | **SD vs. PNFA**

Lexical/ syntactic

acoustic





UNIVERSITY OF TORONTO

Honourable mentions

- Dyslexia.
- Autism.
- Traumatic brain injury and cardiovascular stroke.

- Brain-computer interfaces.
- Interfaces and coding schemes for the blind.