

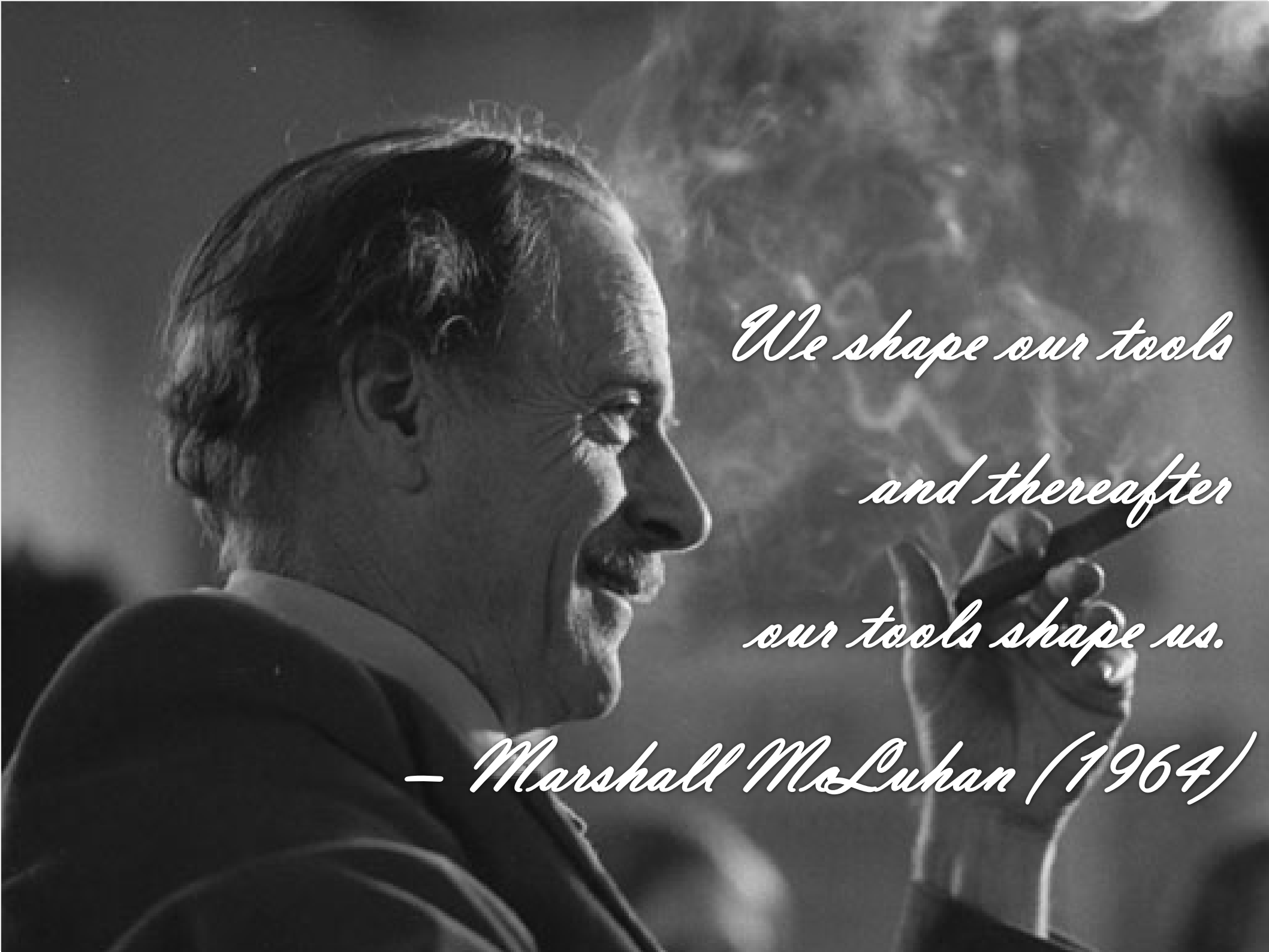


First, we shape our tools

How to build a better speech recognizer

Frank Rudzicz
The University of Toronto

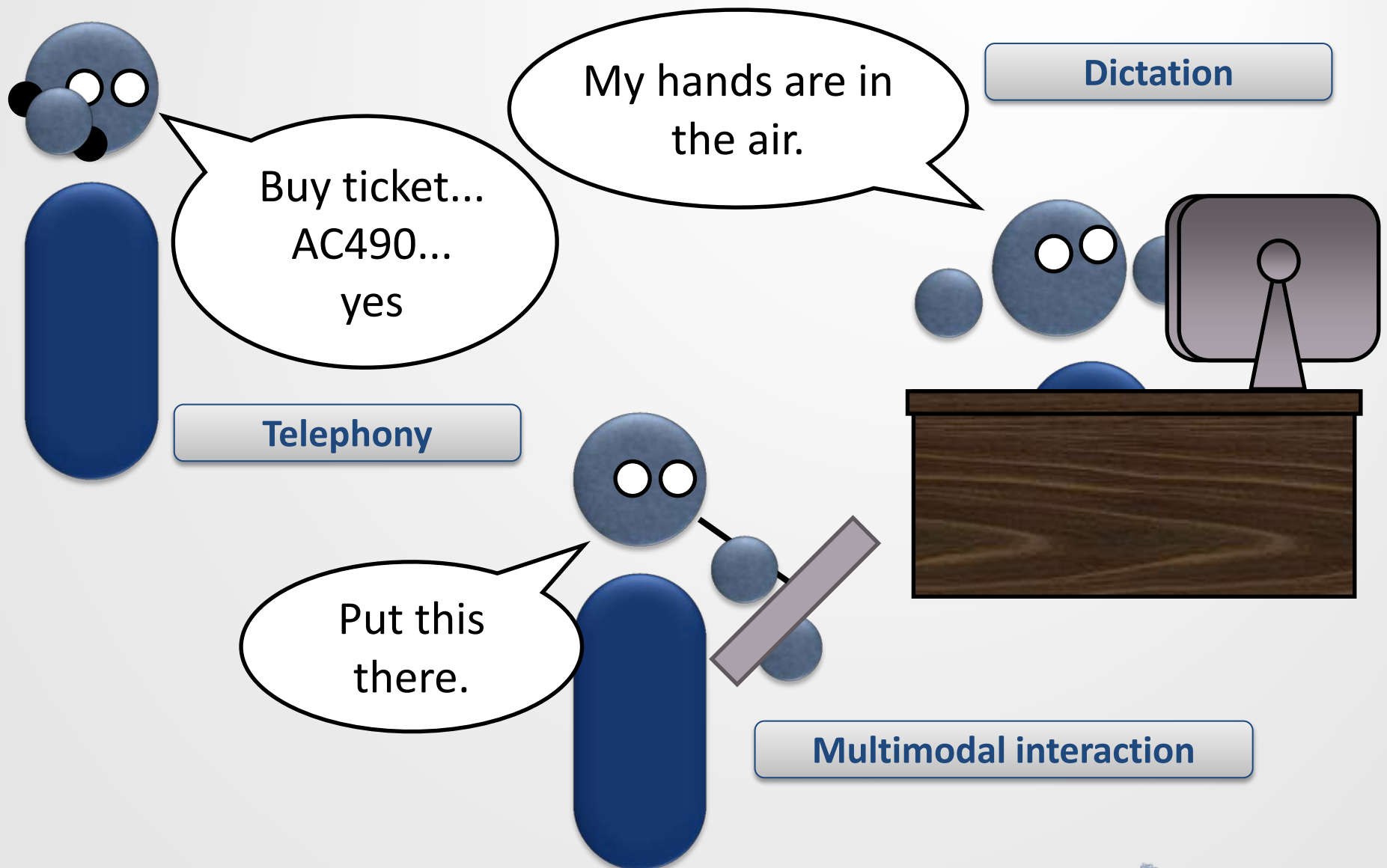
Concordia University, 2 December 2011, Montréal QC



*We shape our tools
and thereafter
our tools shape us.*

— Marshall McLuhan (1964)

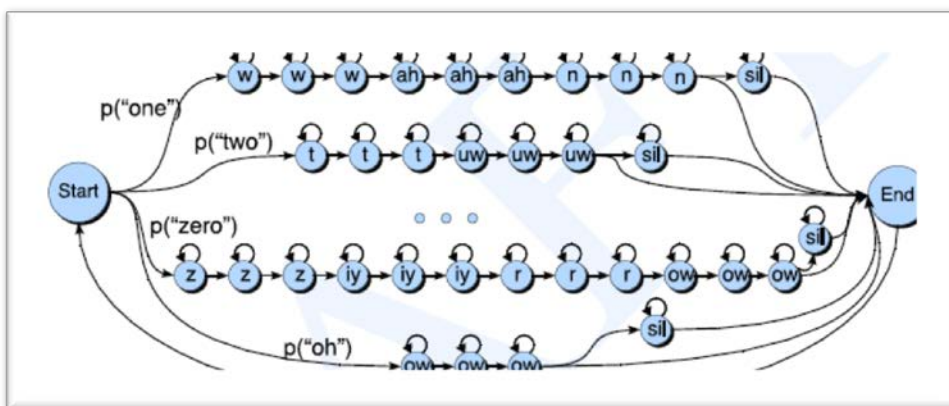
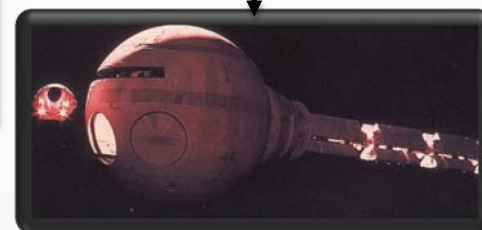
Applications of speech



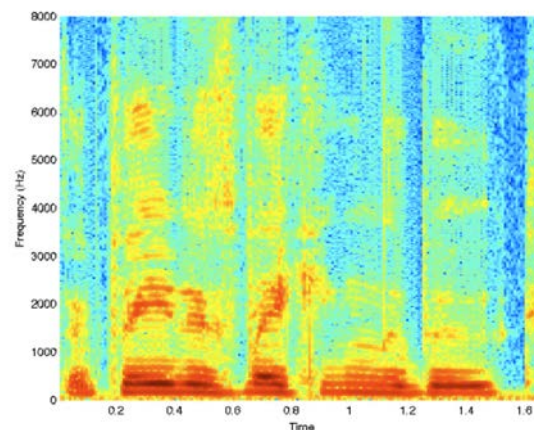
Automatic speech recognition (ASR)



“open the pod bay doors”



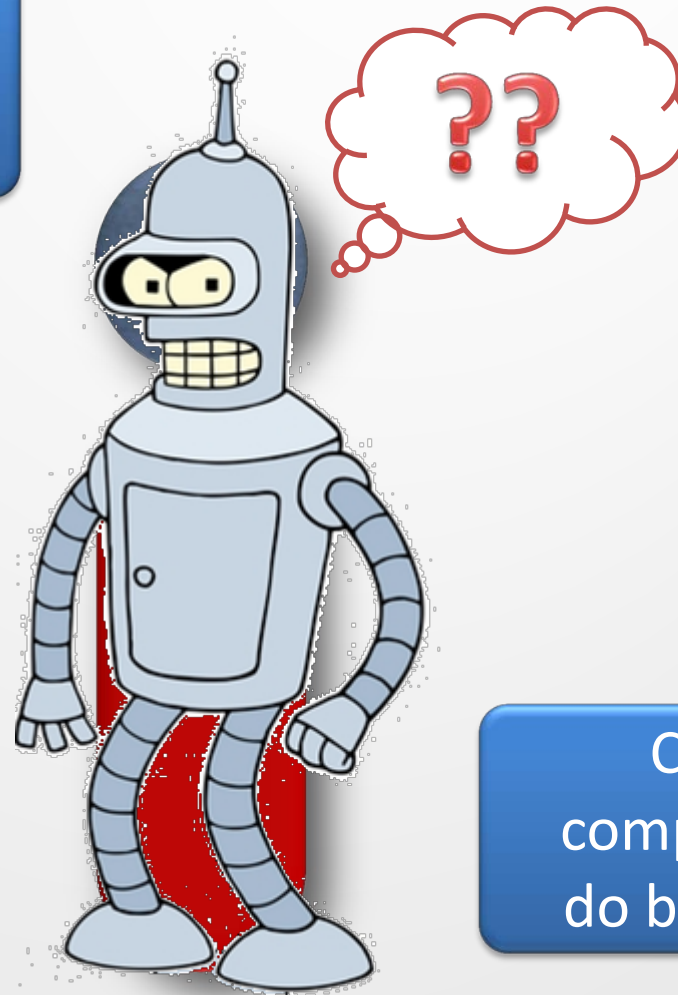
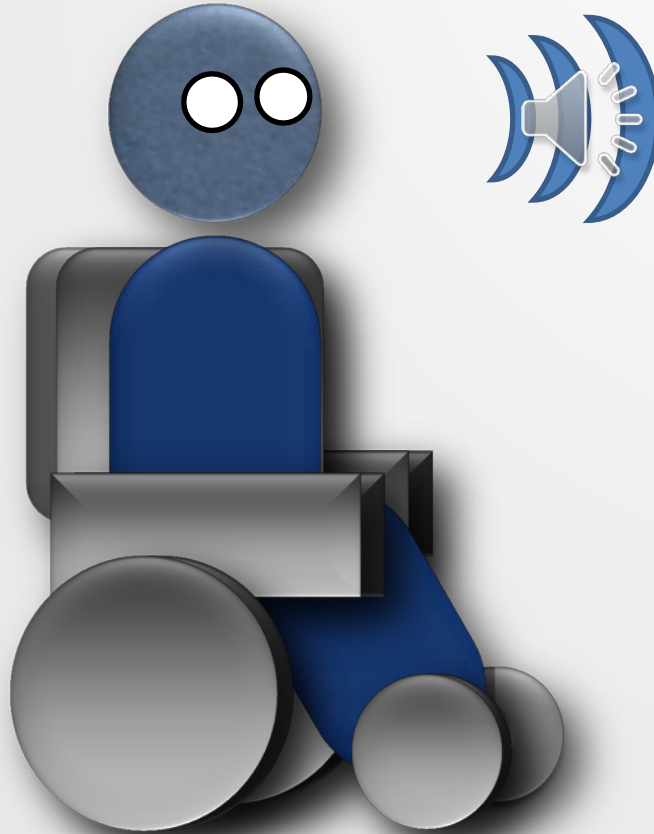
Language model



Acoustic model

Dysarthria

Neuro-motor articulatory difficulties resulting in unintelligible speech.



Can
computers
do better?

Dysarthria and ASR word accuracy

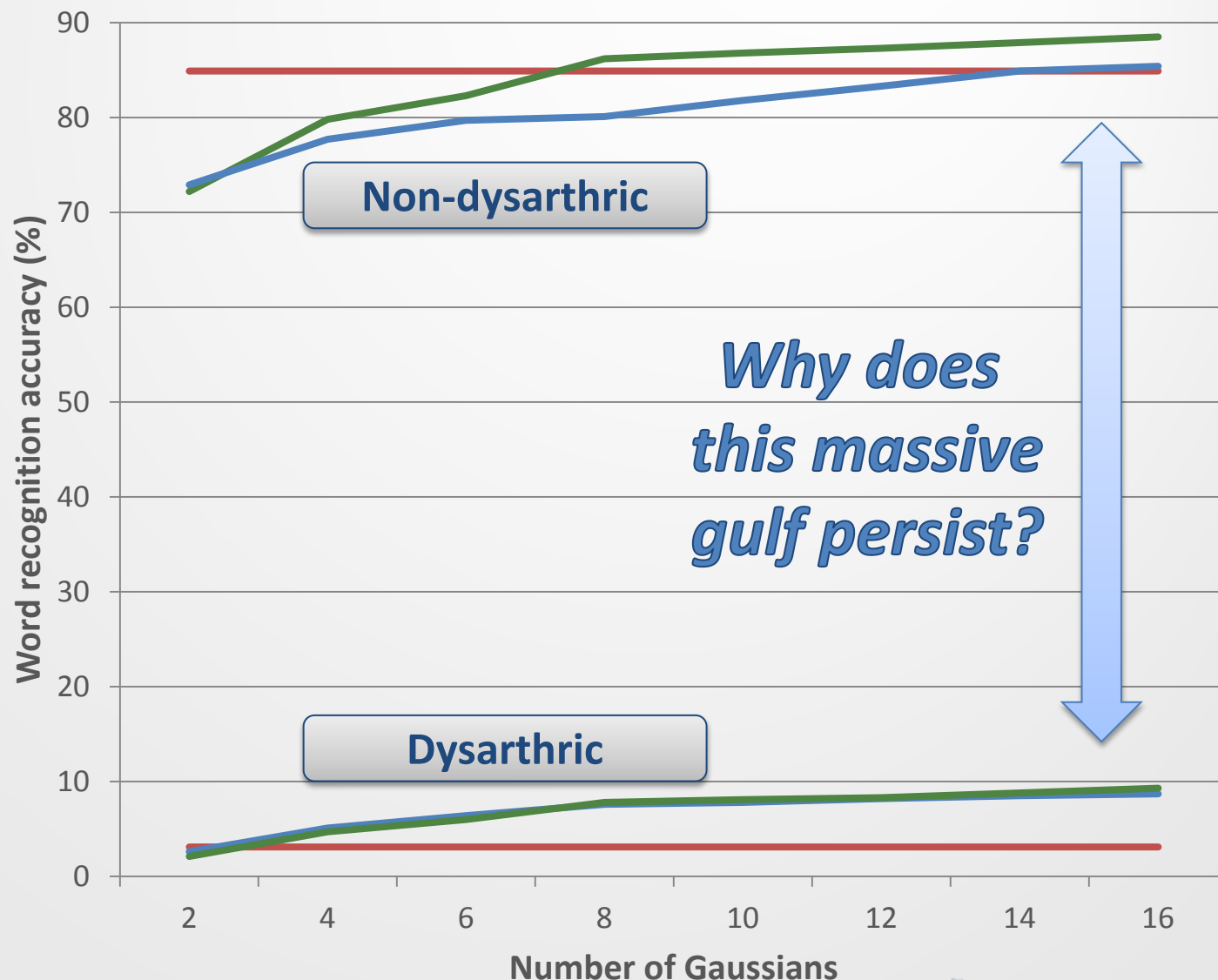
84.9% →

Traditional ASR

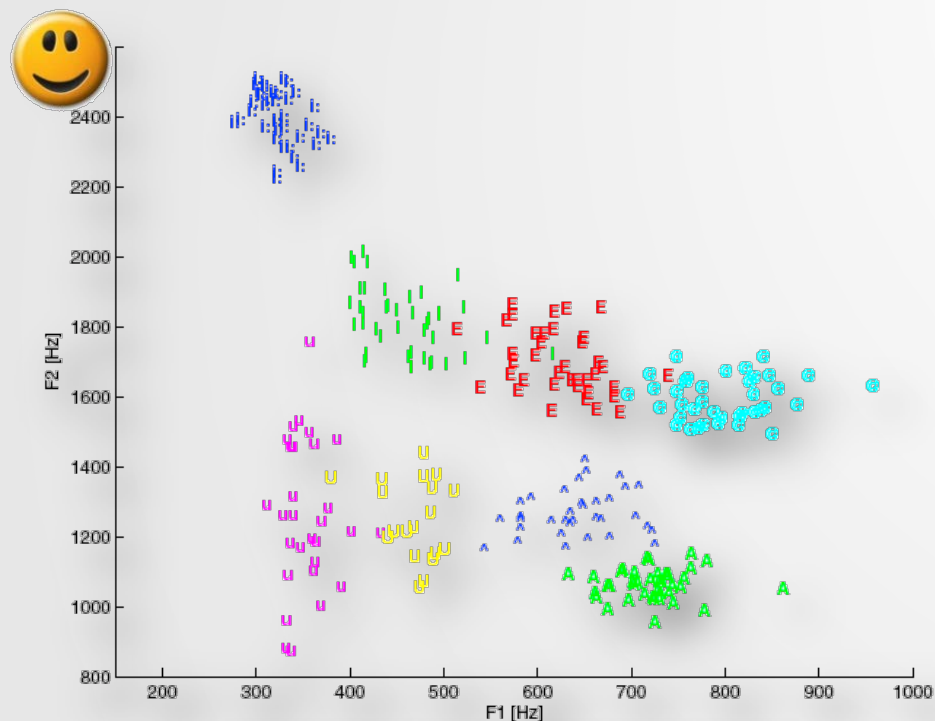
Speaker-
dependent

Speaker-
adaptive

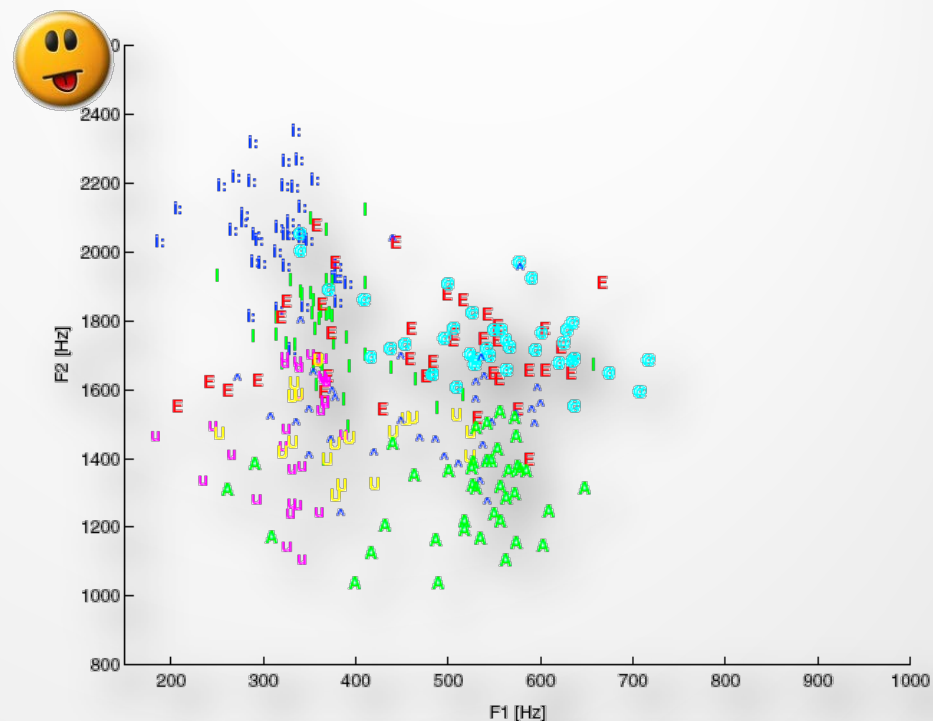
3.1% →



Acoustic ambiguity



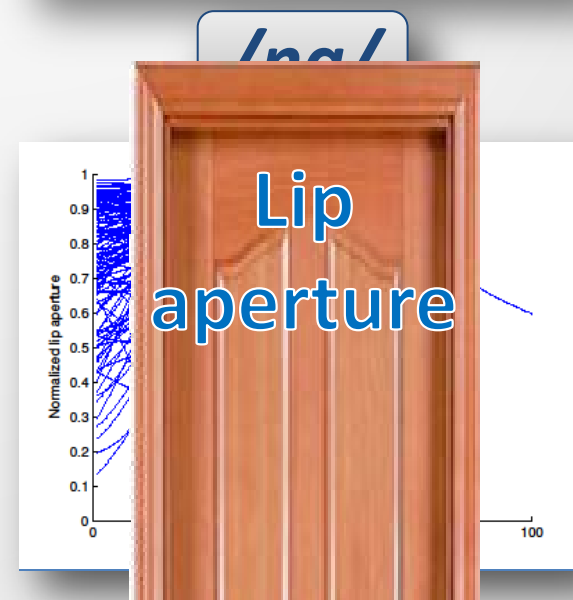
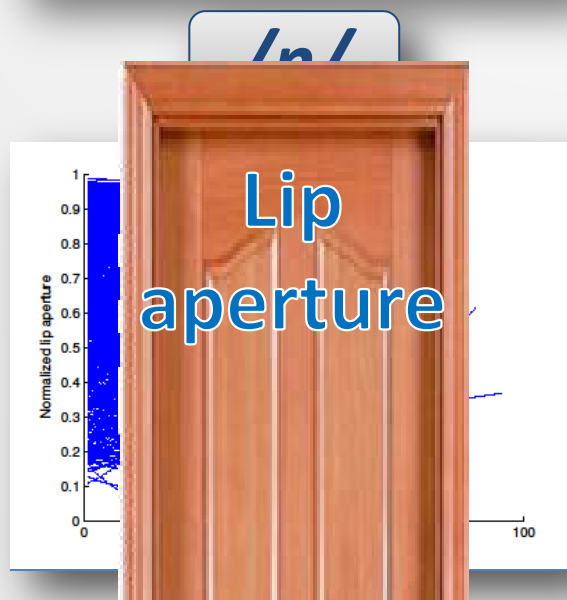
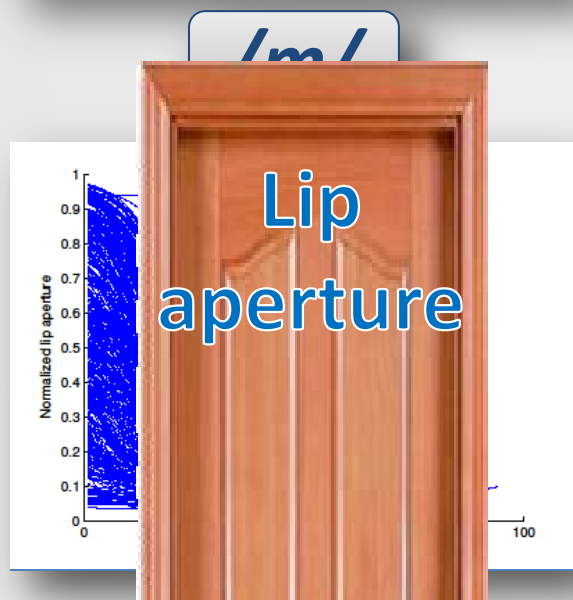
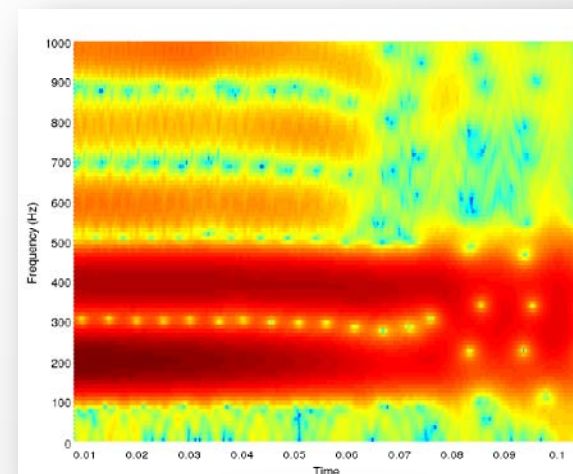
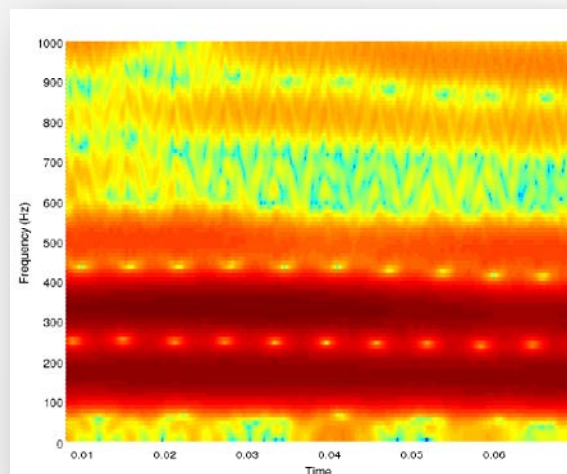
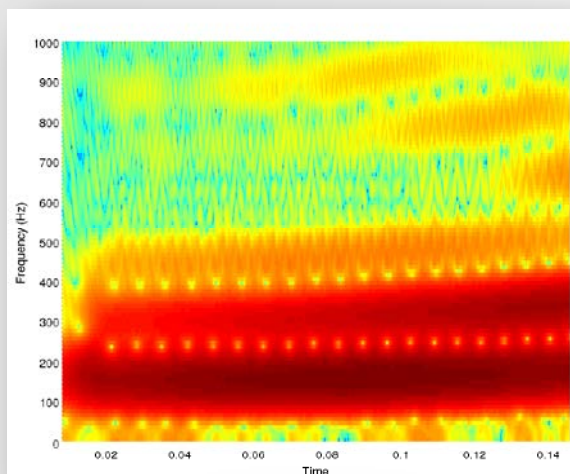
Non-dysarthric



Dysarthric

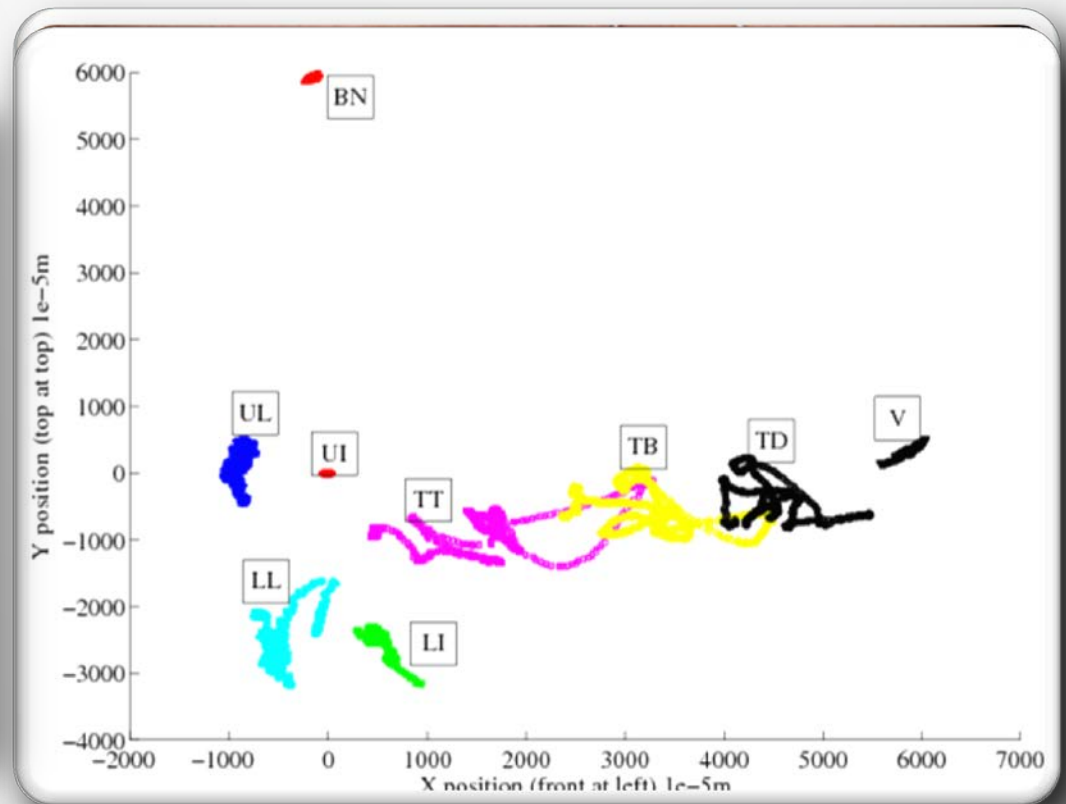
Is this acoustic behaviour indicative of underlying articulatory behaviour?

Articulatory knowledge



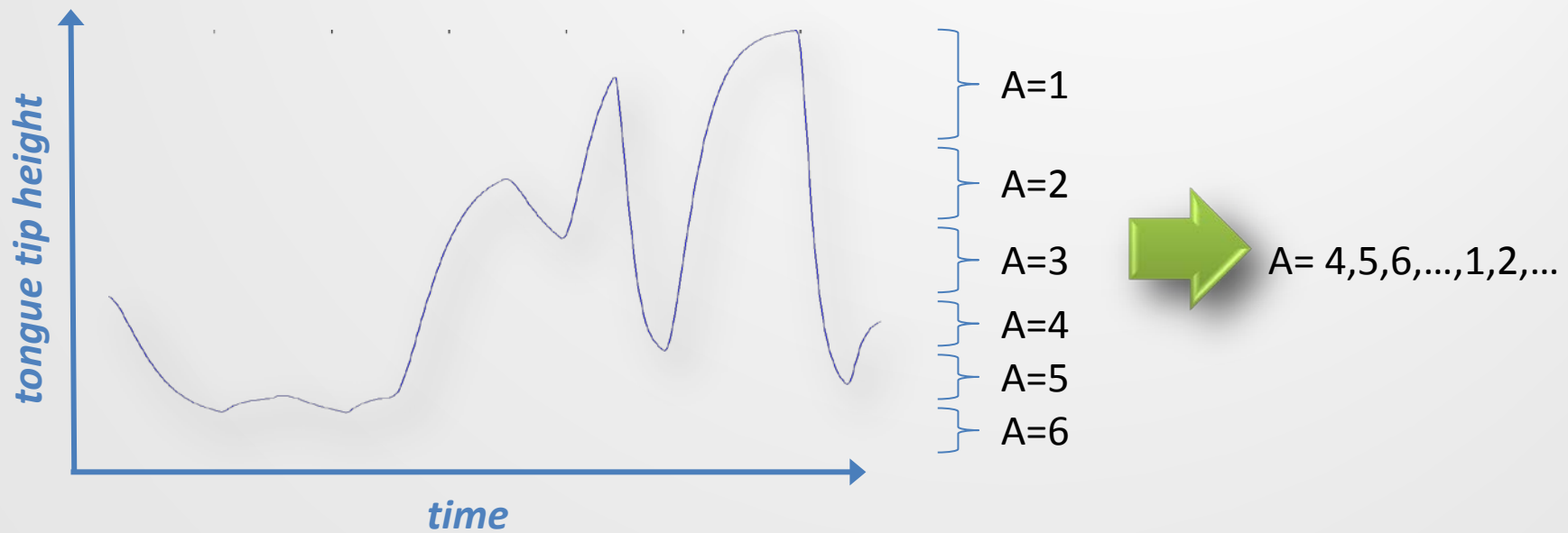
The TORGO database

- **Electromagnetic articulography (EMA):** *n.* 3D recordings (position and velocity) of sensors on the lips and tongue to $\leq 1\text{mm}$ error.

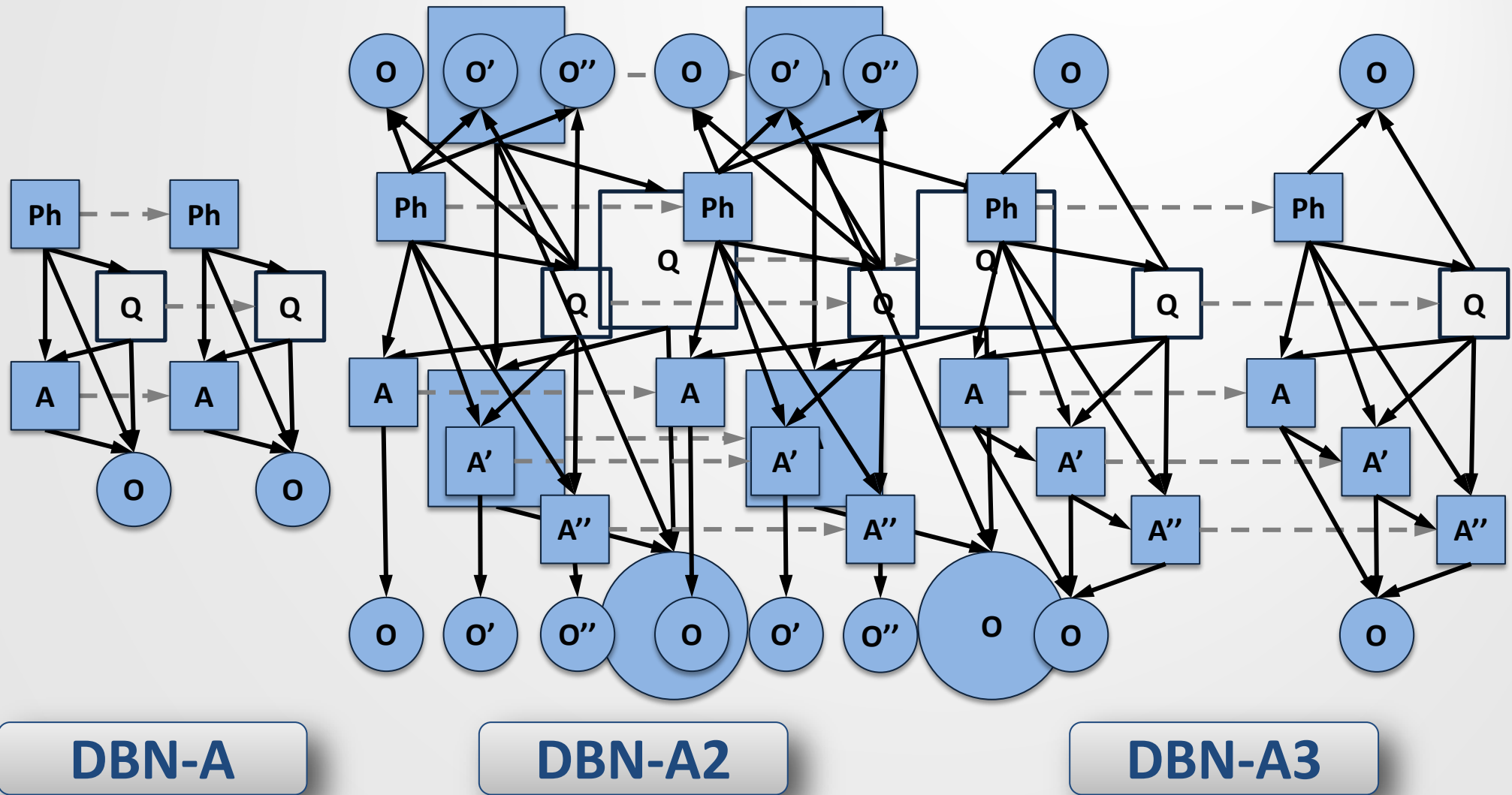


Discretizing articulation

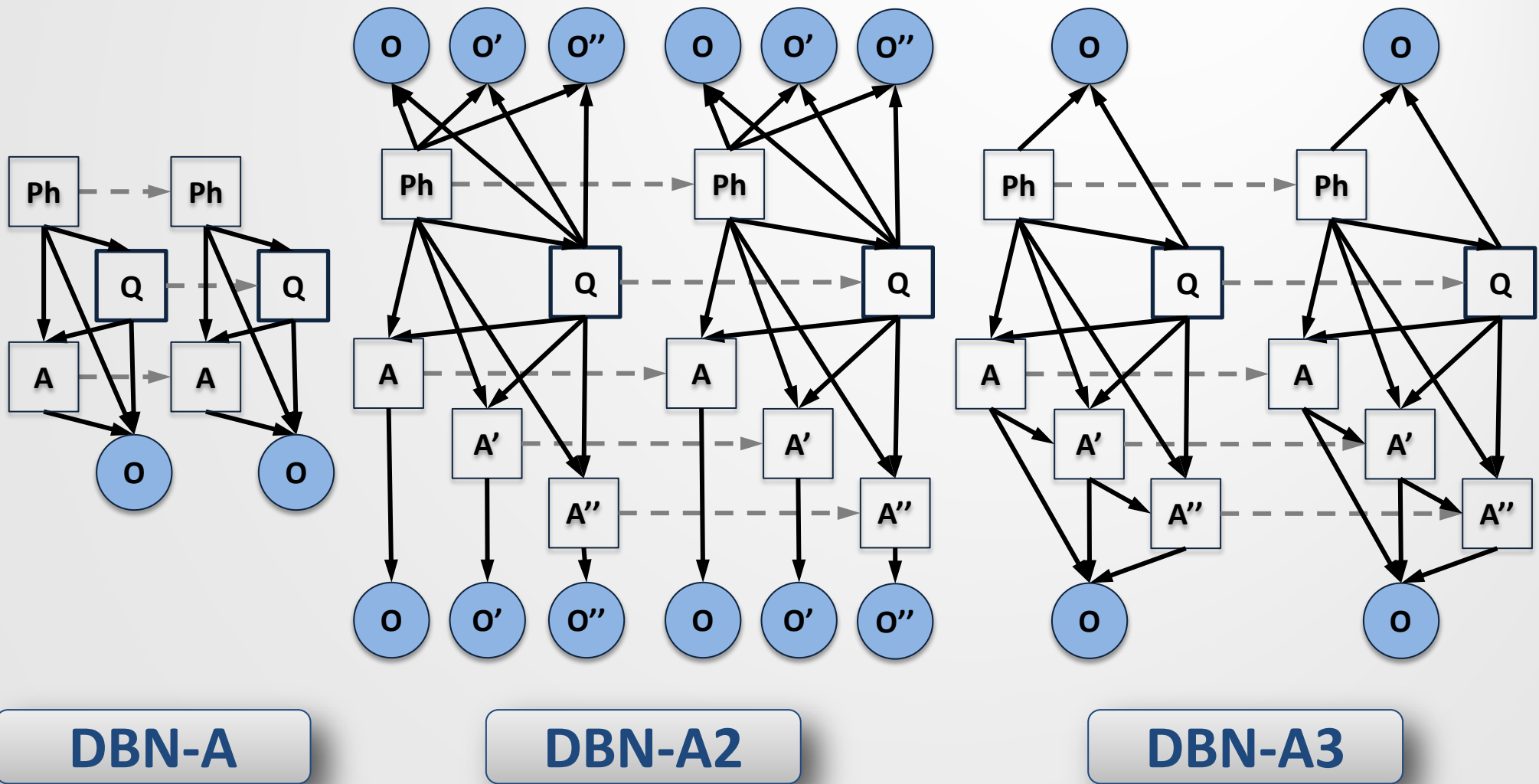
- We convert continuous articulatory motion to discrete **configurations**.
 - The actual configurations are determined automatically by machine learning.
 - This will make learning speech models easier.



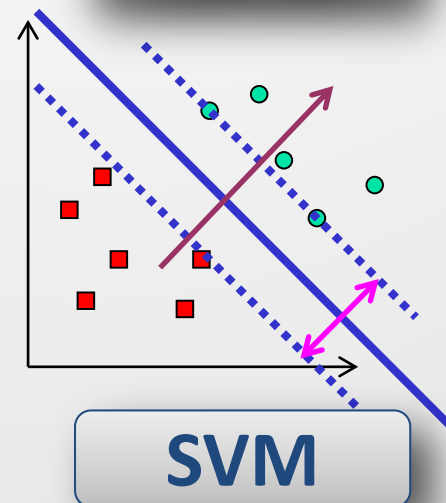
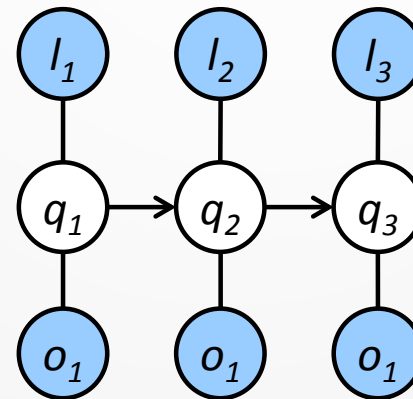
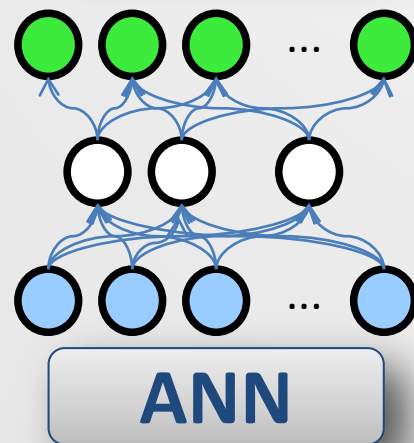
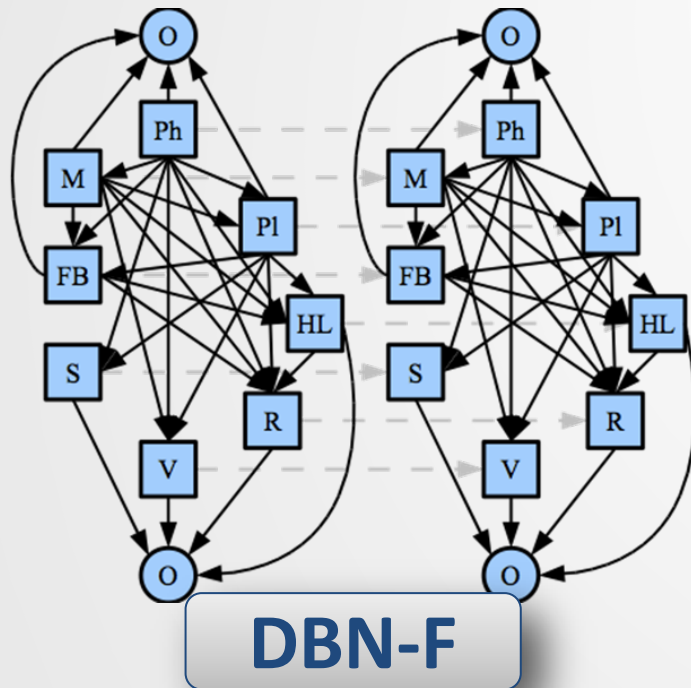
Dynamic Bayes nets with EMA data



Dynamic Bayes nets with EMA data



Other classifiers

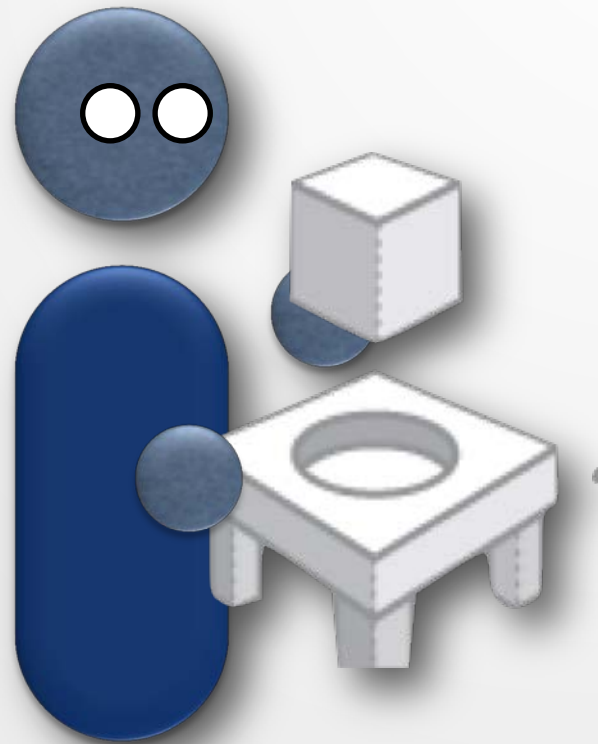


Phoneme recognition

Severity	HMM	LDCRF	DBN		NN	
			DBN-F	DBN-A	MLP	Elman
Severe	14.1	15.2	15.0	16.4	15.5	15.6
Moderate	27.8	28.0	28.0	31.1	28.6	30.5
Mild	51.6	51.8	51.6	54.2	51.4	51.2
Control	72.8	73.5	73.3	73.6	72.6	72.7

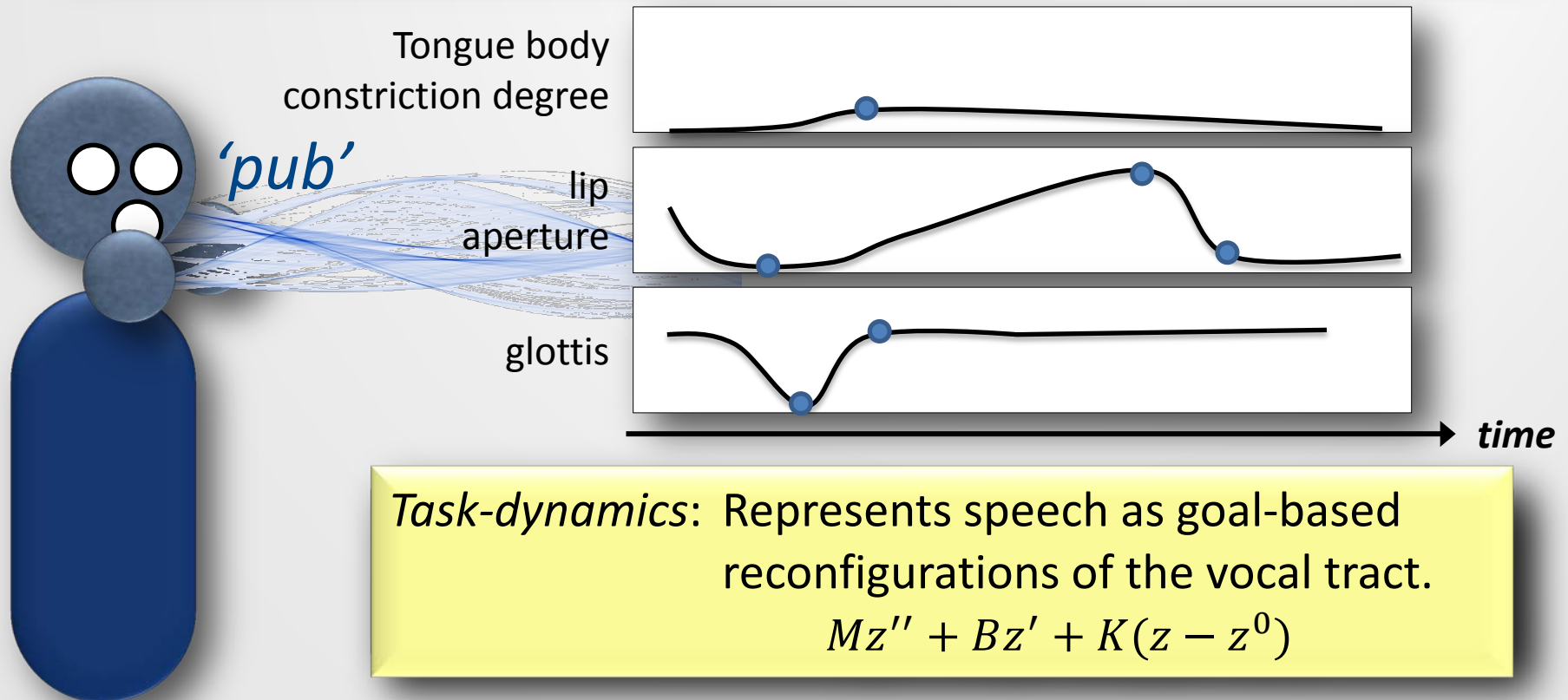
Average % phoneme accuracy (frame-level) with speaker-dependent training

Beyond discretized articulation

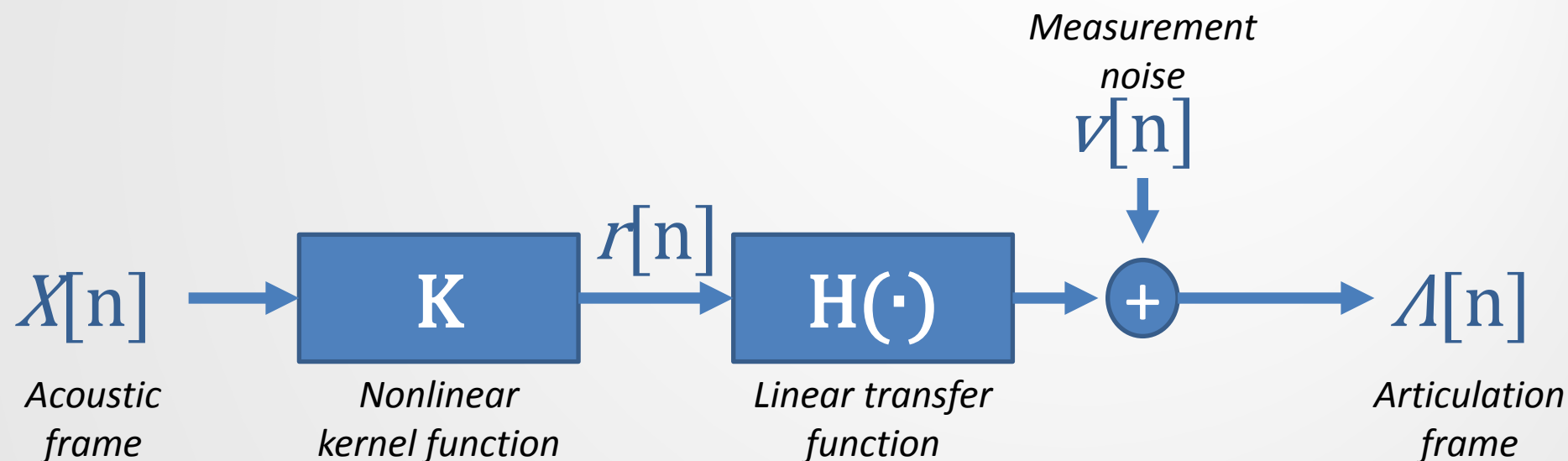


Acoustic-articulatory inversion

We wish to **convert** observed acoustics into an articulatory form amenable to the **identification of linguistic intentions**.

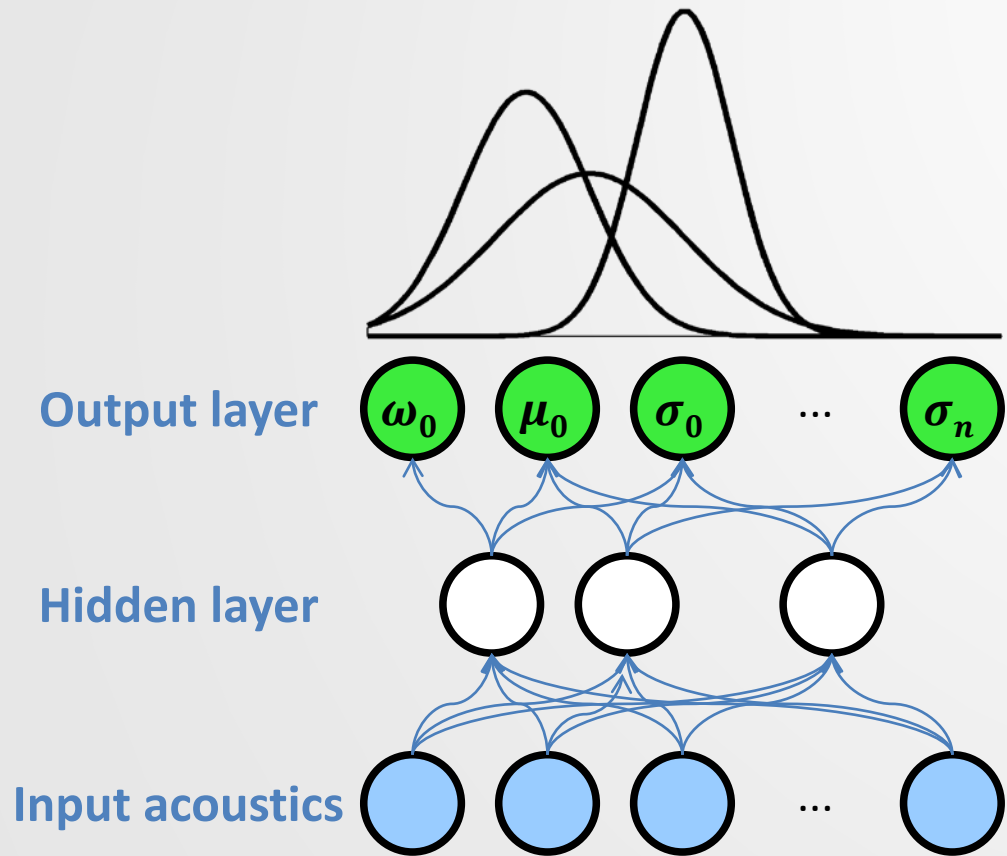


Acoustic-articulatory inversion

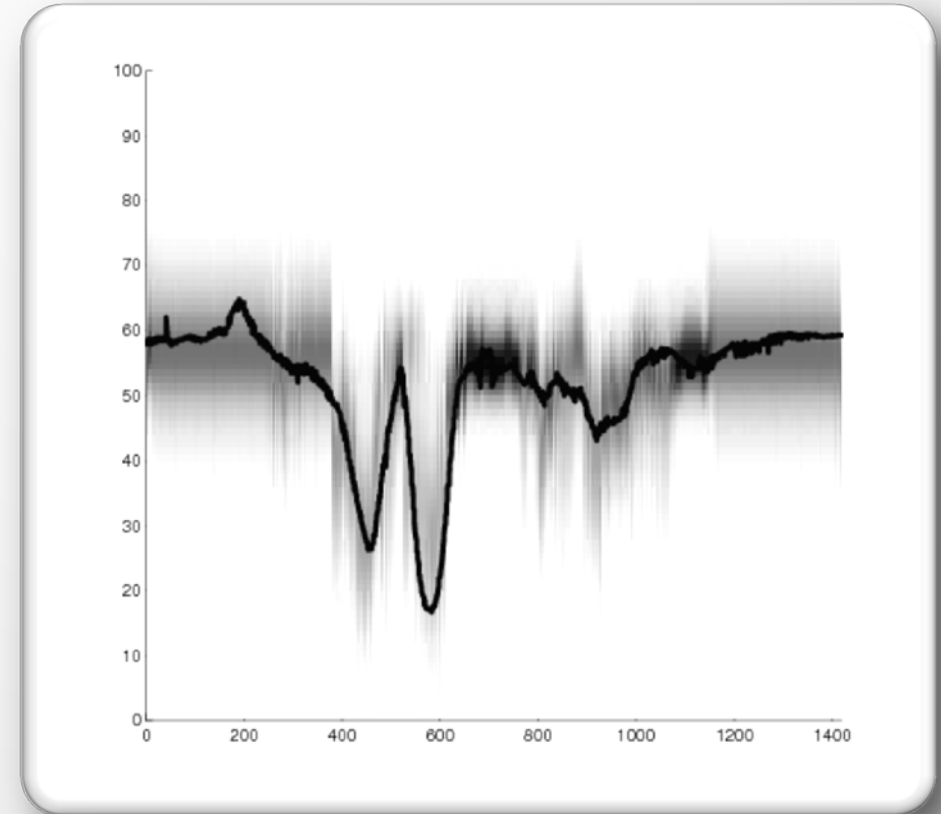


**Inversion system using
kernel-canonical correlation analysis (KCCA)**

State of the art: MDNs



**Mixture density
network**



**Intensity map of estimated
tongue tip constriction over
time**

Comparison of MDN and KCCA

TV	MDN	KCCA	TV	MDN	KCCA
VEL	-0.28	-0.23	TTCD	-1.60	-1.60
LTH	-0.18	-0.18	TTCL	-1.62	-1.57
LA	-0.32	-0.28	TBCD	-0.79	-0.80
LP	-0.44	-0.41	TBCL	-0.20	-0.18
GLO	-1.30	-1.14			

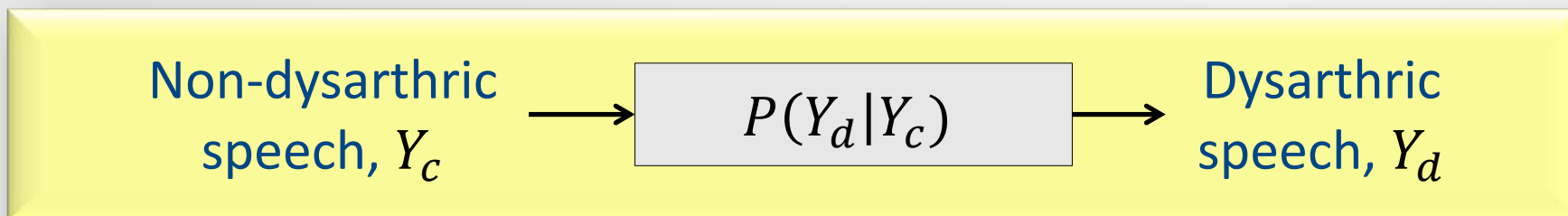
Log likelihoods of true tract variable positions in test data, under distributions produced by MDN and KCCA. Higher values are better.

95% confidence

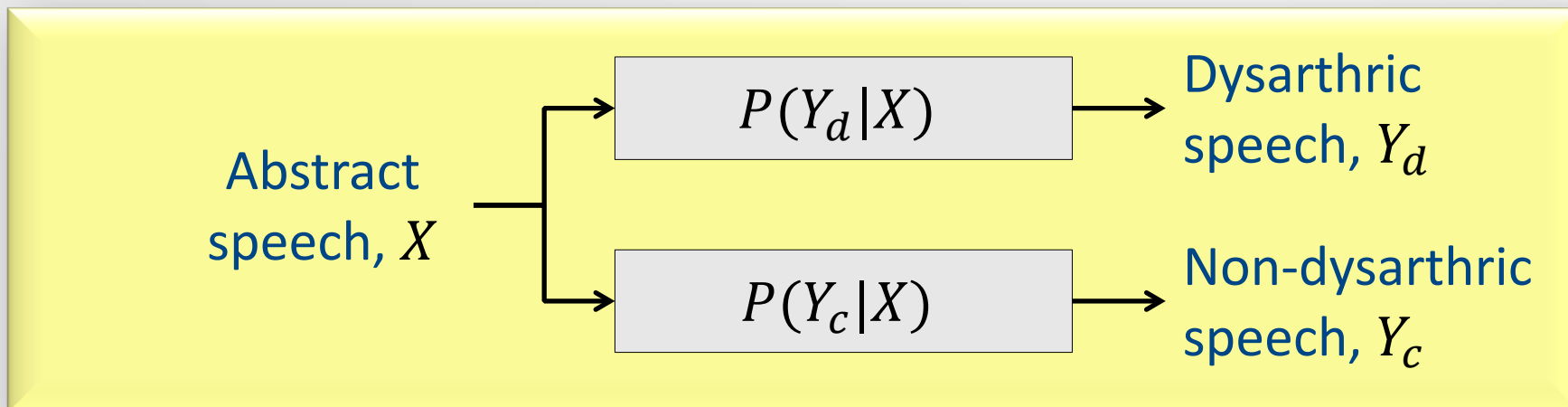
99% confidence

The noisy channel

Is dysarthria a distortion of non-dysarthric speech?



... or are they both distortions of a common abstraction?



Noisy channel results

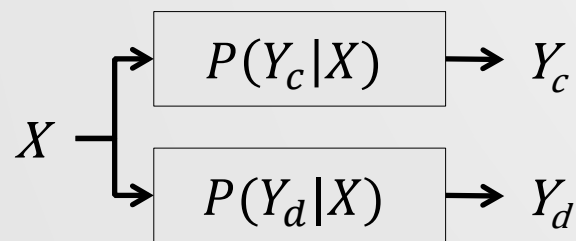
		Source distribution	Target distribution	KL divergence (10^{-2} nats)	
				Acoustics	Articulation
		Ctrl	Dys	25.36	3.23
		Ctrl $\rightarrow$ Dys	Dys	17.78	2.11
		TD $\rightarrow$ Ctrl	Ctrl	N/A	1.69
		TD $\rightarrow$ Dys	Dys	N/A	1.84

$Y_c \rightarrow P(Y_d|Y_c) \rightarrow Y_d$
 $X \rightarrow P(Y_c|X) \rightarrow Y_c$
 $X \rightarrow P(Y_d|X) \rightarrow Y_d$

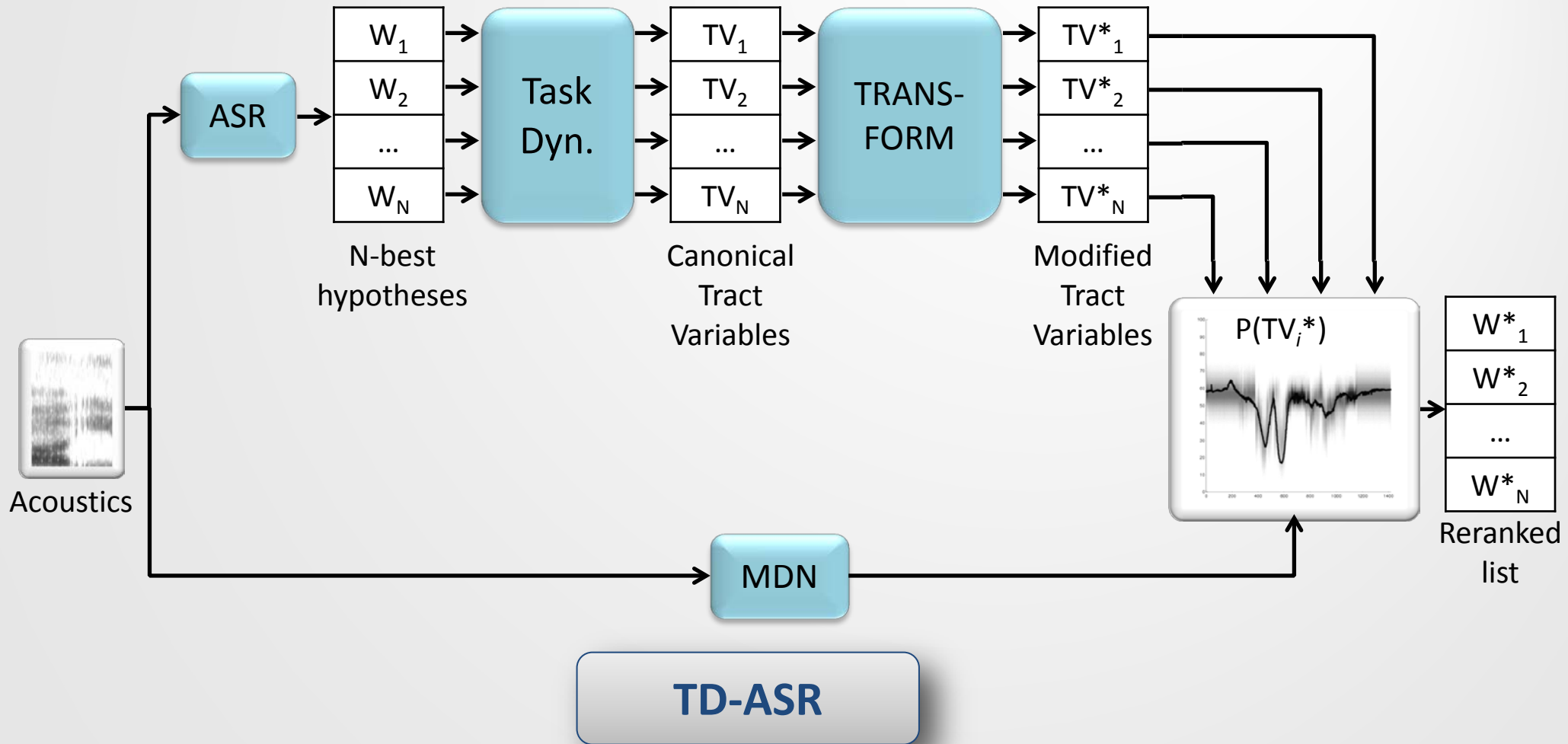
Dysarthric articulation is more accurately predicted given **task-dynamics** as a source model

Noisy channel in ASR

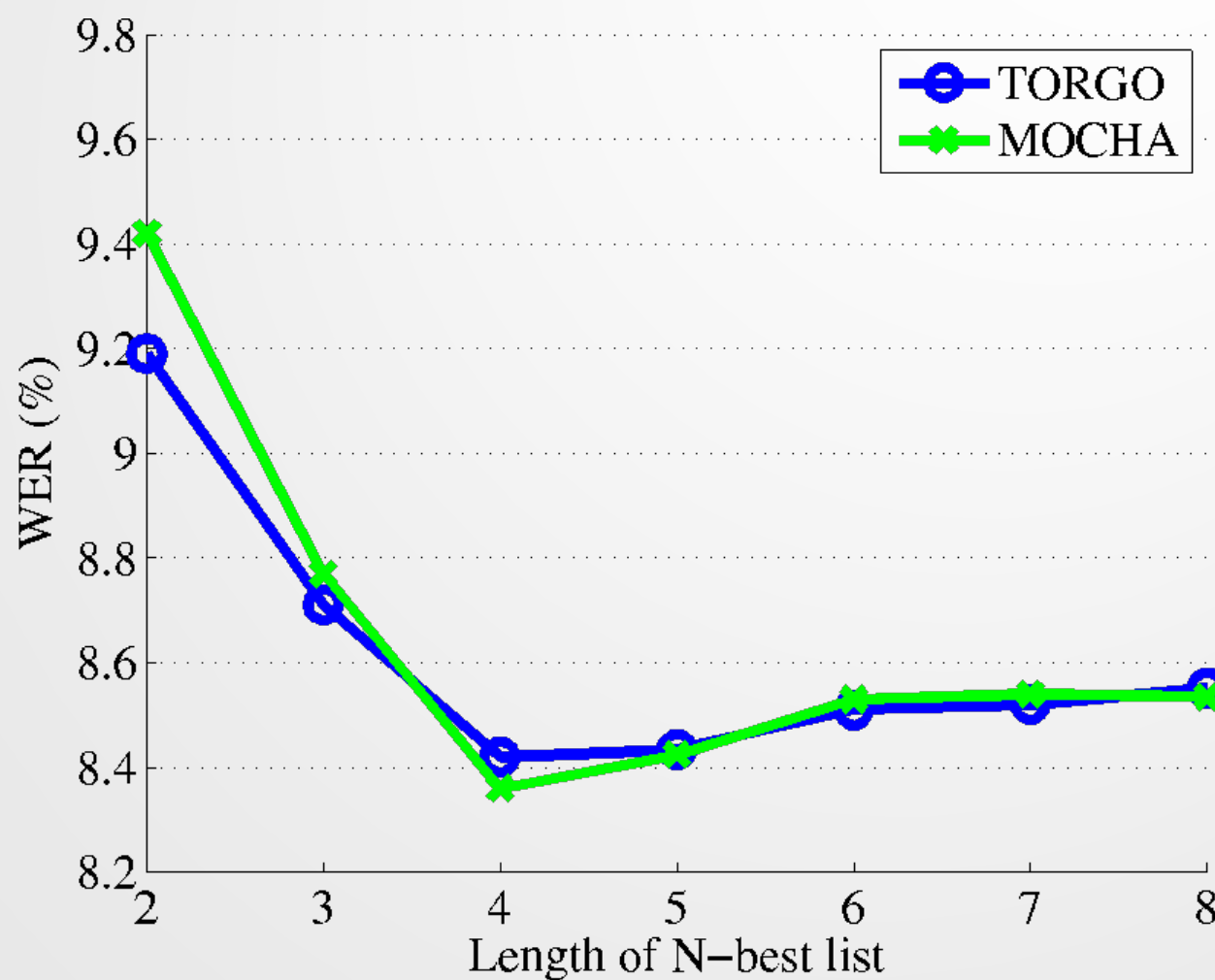
How might we **combine** the insight of this **noisy channel model** and **acoustic-articulatory inversion** within a speech recognition system?

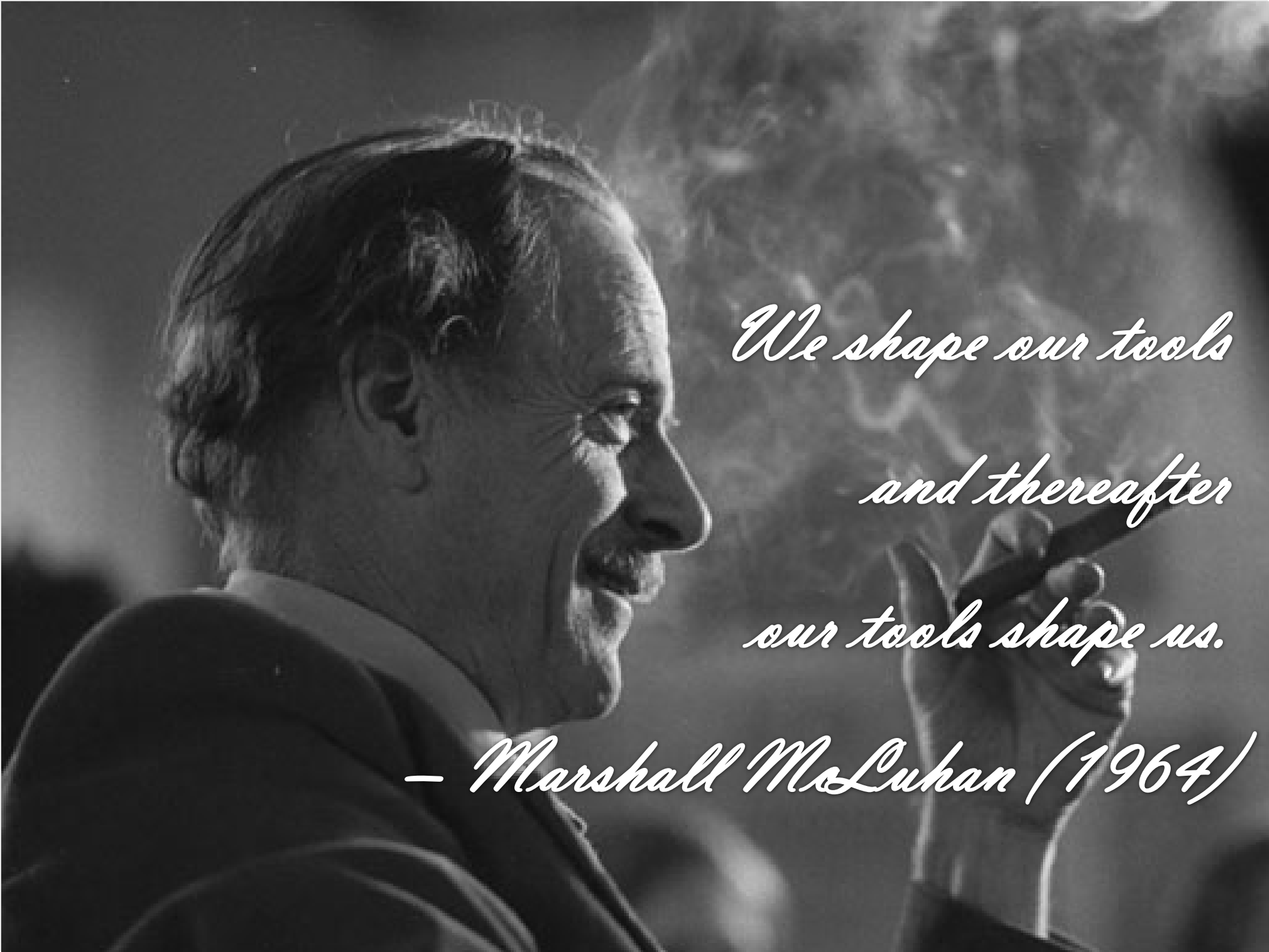


Correcting ASR with task dynamics



TD-ASR results

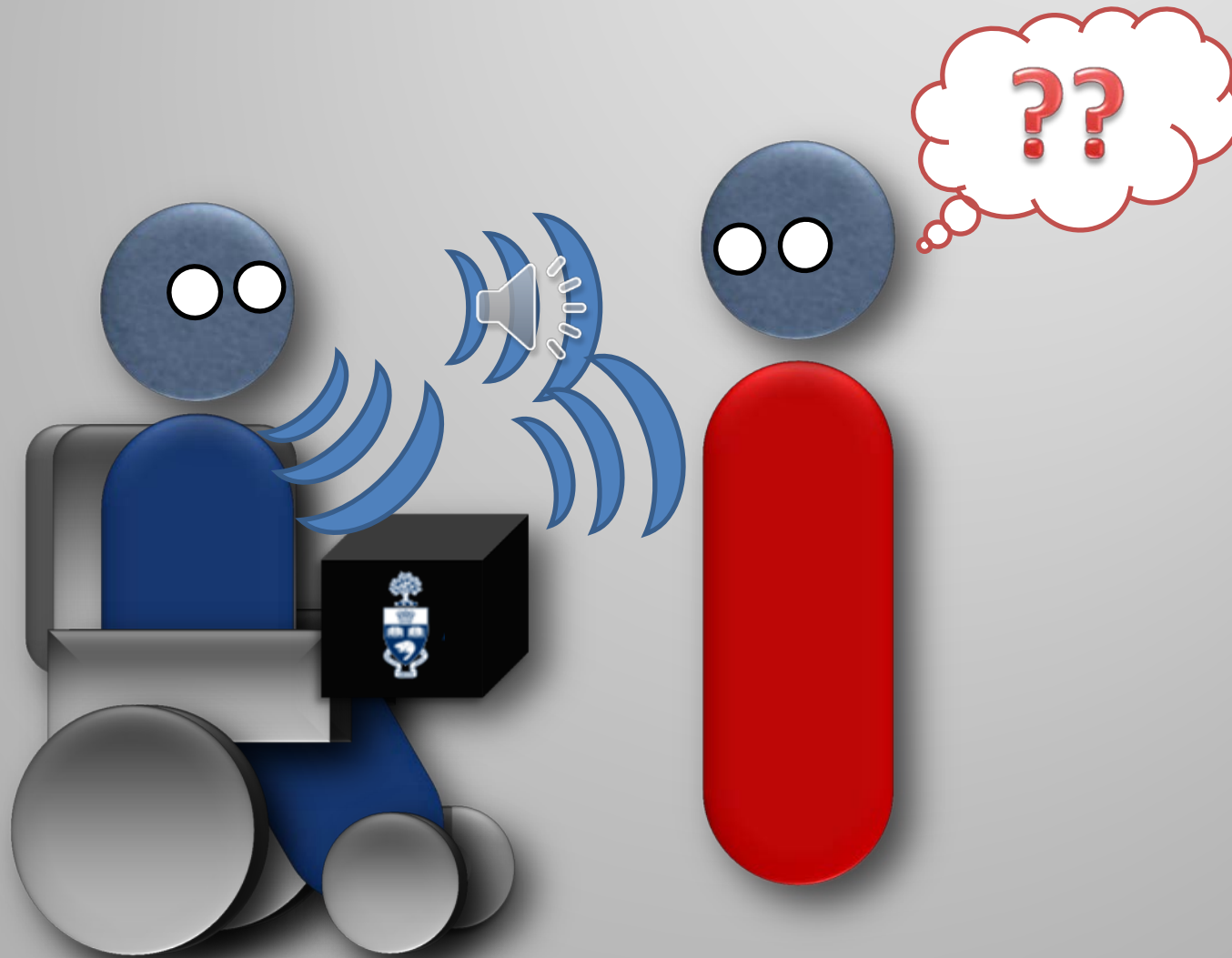


A black and white photograph of Marshall McLuhan in profile, facing right. He is wearing a dark suit jacket and a light-colored shirt. He has a mustache and is holding a pipe in his right hand, which is raised towards his face. The background is dark and textured, possibly smoke or a cloudy sky.

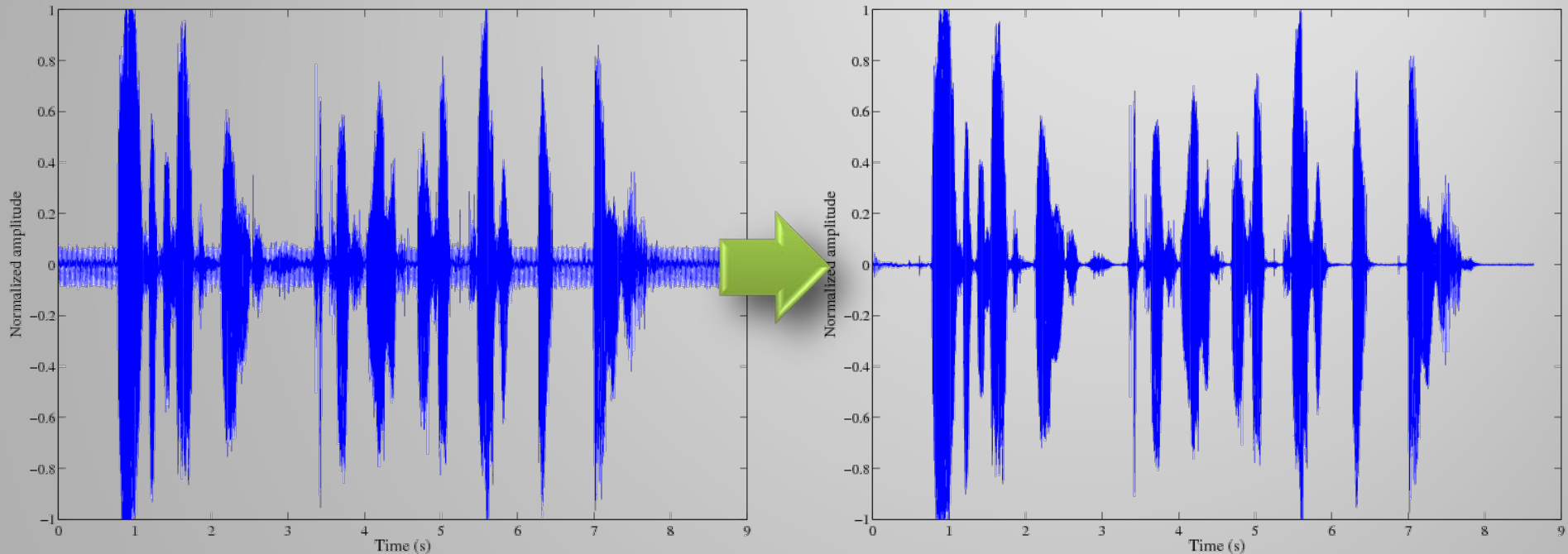
*We shape our tools
and thereafter
our tools shape us.*

— Marshall McLuhan (1964)

Speech transformation



1. Noise reduction

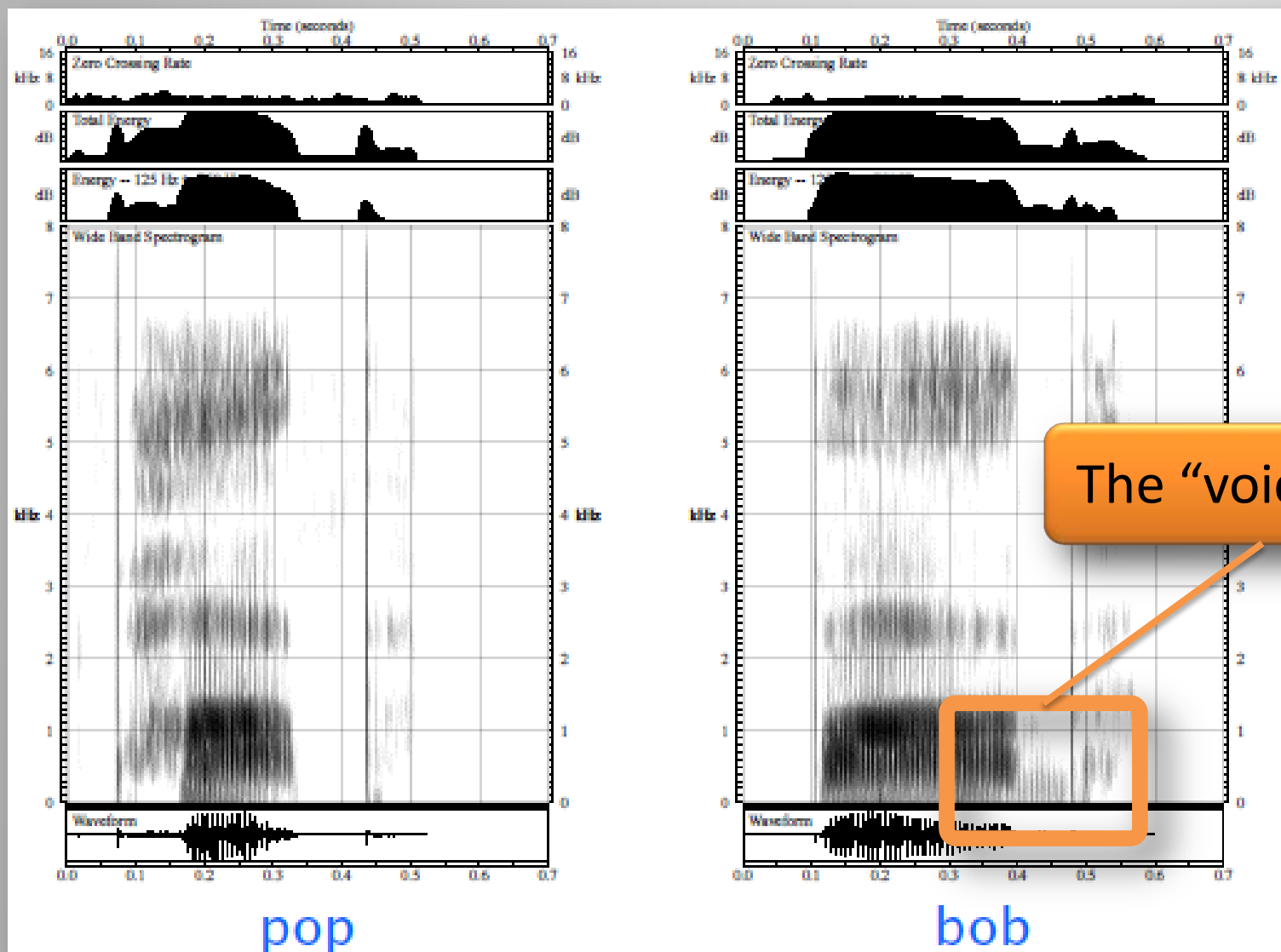


Before

After

Spectral subtraction removes environmental signal noise.

2. Unvoiced and voiced consonants



3. 'Splicing': Deletions and insertions

- Deleted sounds are patched with synthetic equivalents.

feelin



feeling

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

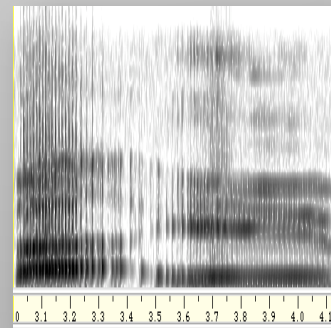
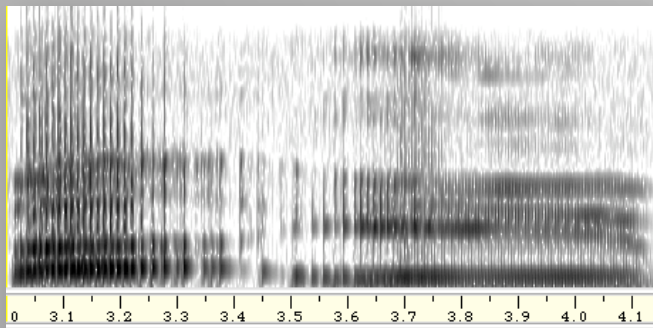
pr-pr-pr-pronounced



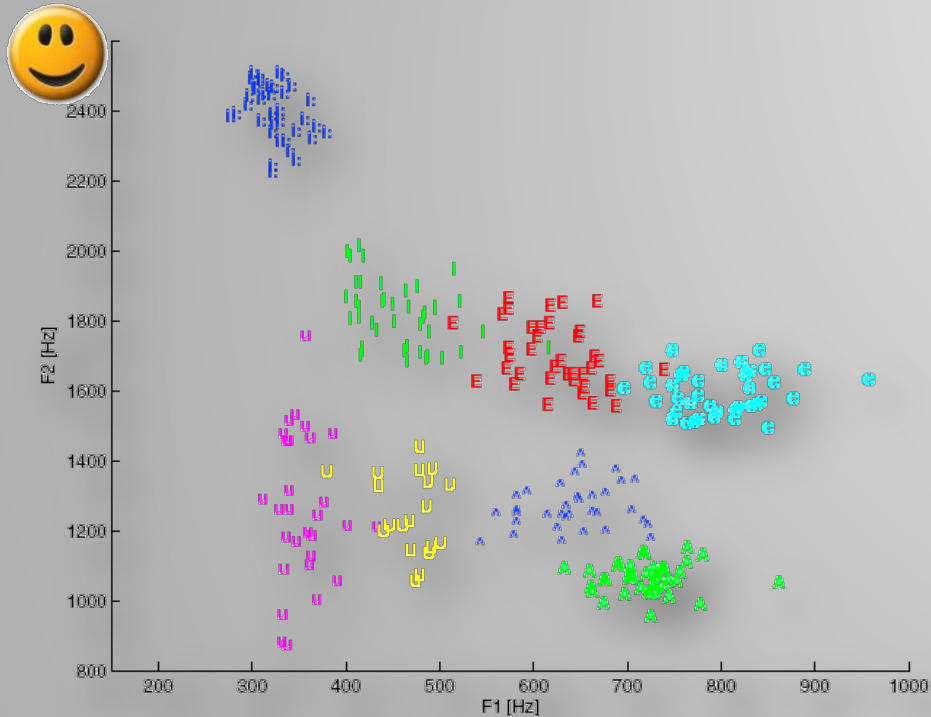
pronounced

4. Tempo morphing

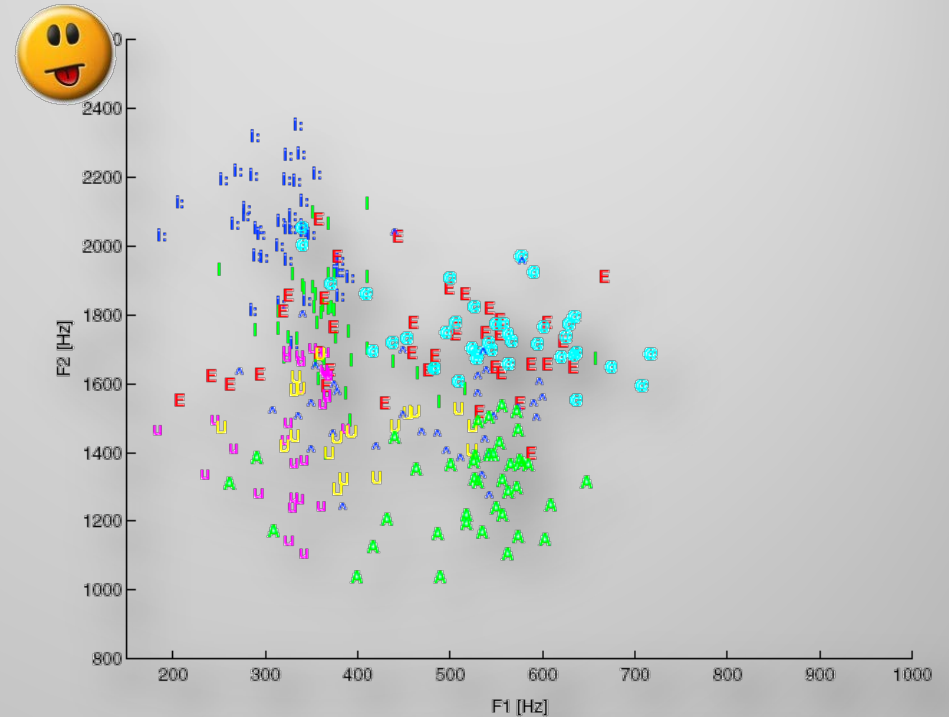
- Dysarthric speech tends to be a lot (often 3x) **slower** than typical speech.
- **Sonorants** are contracted in time to be closer to their **expected** length.
 - A **phase vocoder** contracts the length of a signal **without** affecting its pitch or frequency characteristics.



5. Formant ambiguity



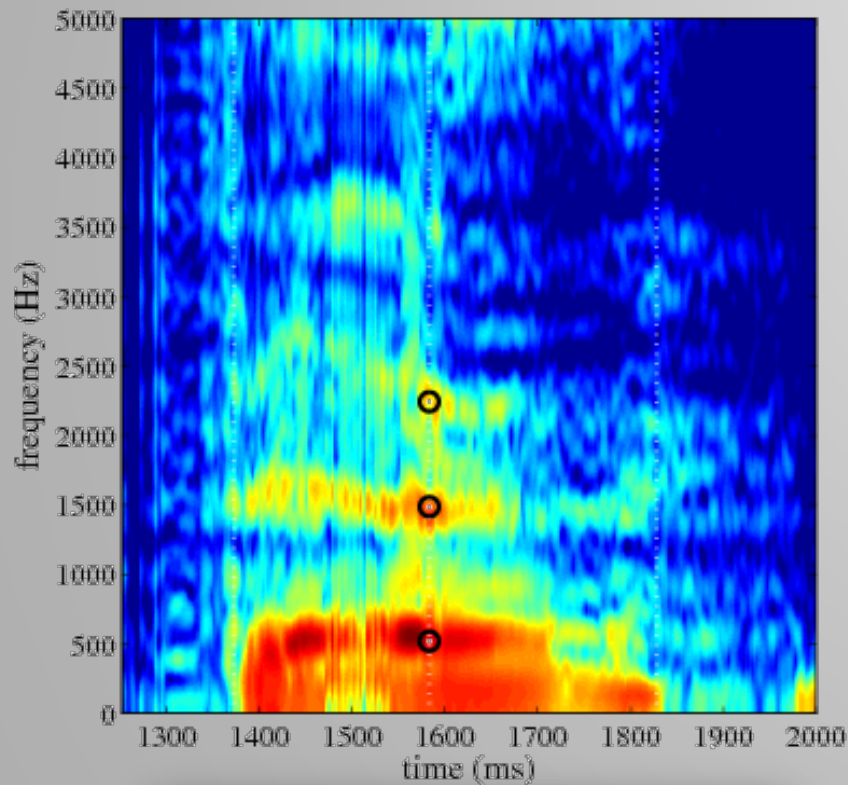
Non-dysarthric



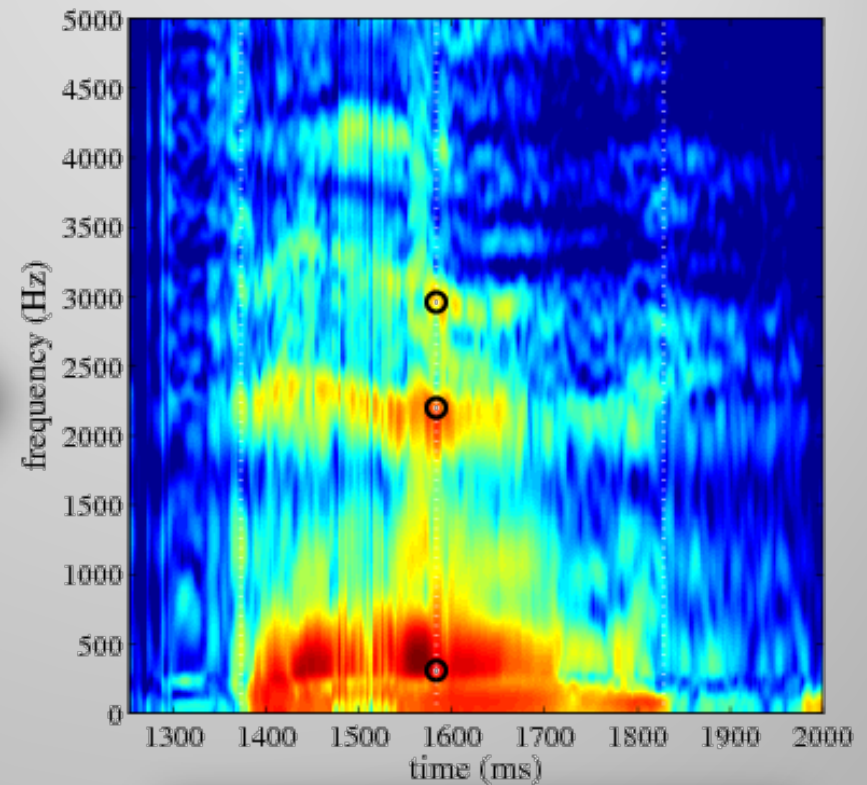
Dysarthric

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

5. Formant morphing



Before



After

Speech transformation



Improvements in intelligibility

Listener	Original	GMM	Synthetic TTS	TORGOMorph		
				Splice	Timing	Frequency
L01	22.1	15.6	82.0	40.2	34.7	35.2
L02	27.8	12.2	75.5	44.9	39.4	33.8
L03	38.3	14.8	76.3	37.5	12.9	21.4
L04	24.7	10.8	72.1	32.6	22.2	18.4
Average	28.2	13.6	76.5	38.8	27.3	27.2



Word recognition % across four listeners given various types of acoustics

thotra.

Multimodal interaction

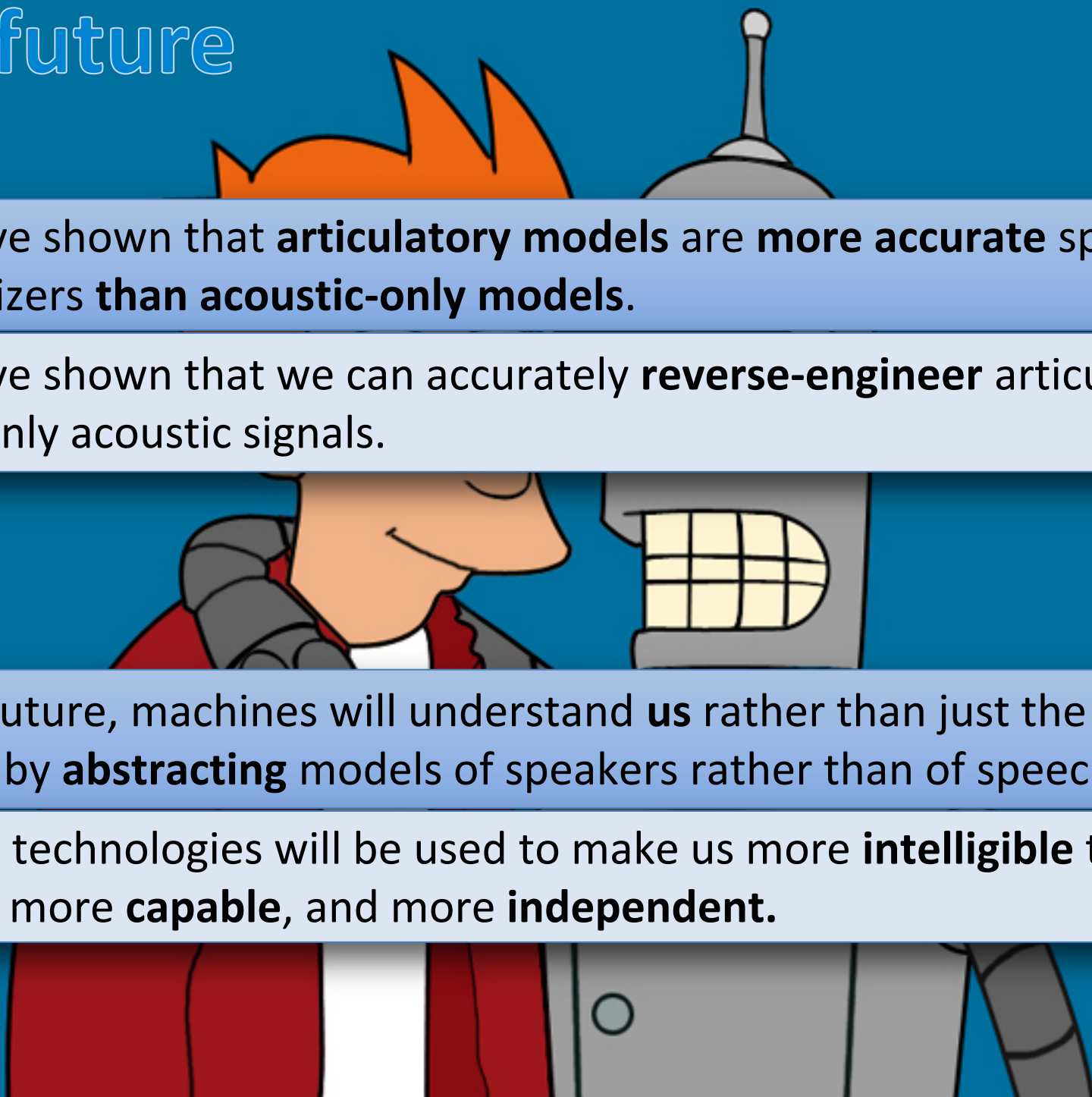
Put this there.

- Integrating concurrent streams of communication can, e.g.:
 - Enable more **natural** and **efficient** expression, and
 - **Reduce ambiguity** in any one of those streams.

Speech for aging gracefully

- Older adults at risk of dementia have special need but want to live at home.
- Speech interfaces in the environment.
 - Can be used in emergency situations (e.g., reacting to falls)
 - e.g., HomeLab: *"do you want me to call your son?"*
 - Can be used to guide an individual through daily tasks.
 - e.g., Homelab: *"don't forget to turn the kettle off."*

The future

A cartoon illustration of a person with bright orange, spiky hair and a robot with a grey dome-shaped head and a single antenna. The person is wearing a red jacket over a white shirt. The robot is wearing a grey suit. They are both smiling and looking at each other. The background is a solid blue color.

We have shown that **articulatory models** are **more accurate** speech recognizers **than acoustic-only models**.

We have shown that we can accurately **reverse-engineer** articulation given only acoustic signals.

In the future, machines will understand **us** rather than just the **words** we say by **abstracting** models of speakers rather than of speech.

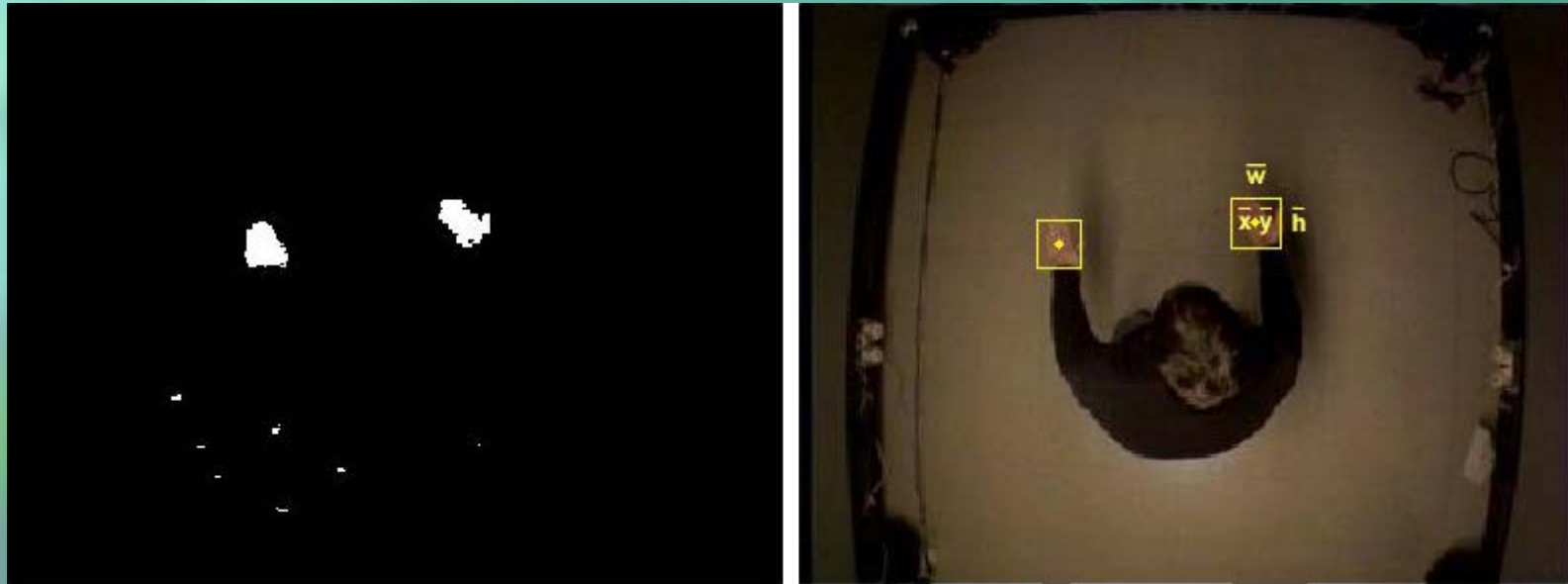
Speech technologies will be used to make us more **intelligible** to others, more **capable**, and more **independent**.

A person in an orange spacesuit is walking away from the viewer down a long, brightly lit corridor. The corridor has a yellowish-orange floor and walls, with several black rectangular panels or doors on the sides. The perspective is from behind the person, looking down the length of the corridor.

Thank you!

`frank@cs.toronto.edu`

Speech and gesture

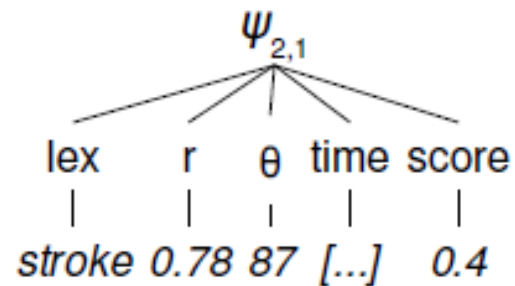
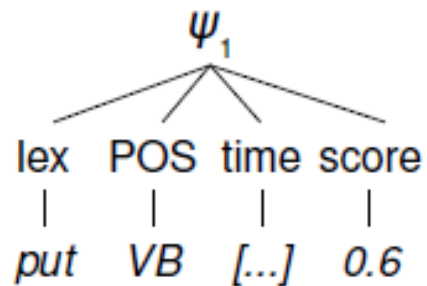


Ex.

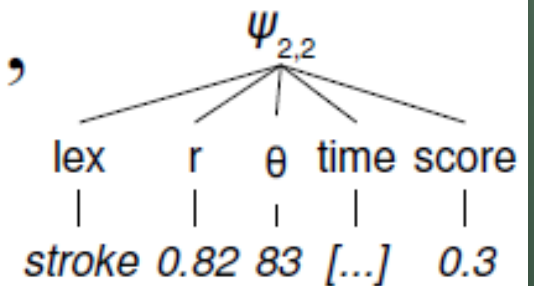
"put"

...

<stroke $r=0.80$ $\theta=85$ $\phi=75$ >

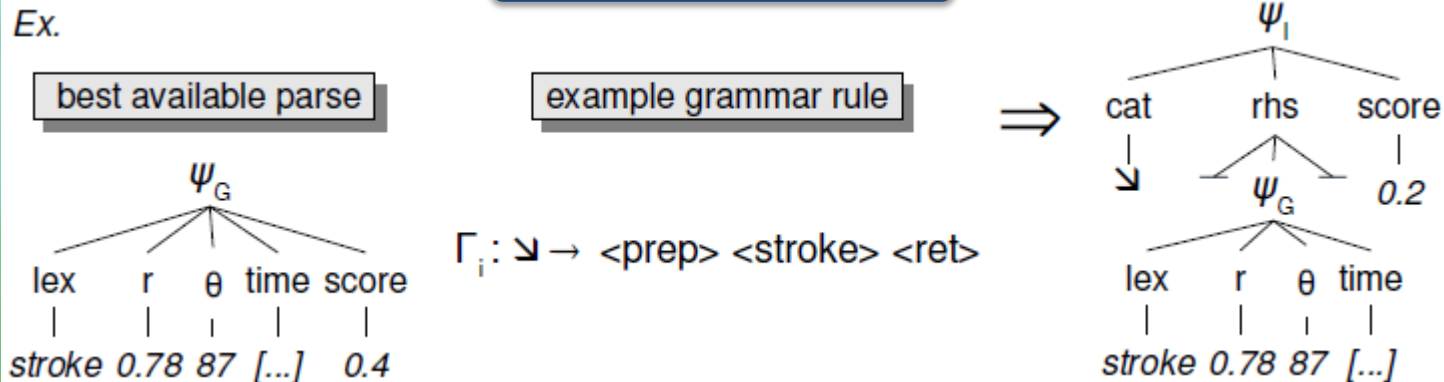


,

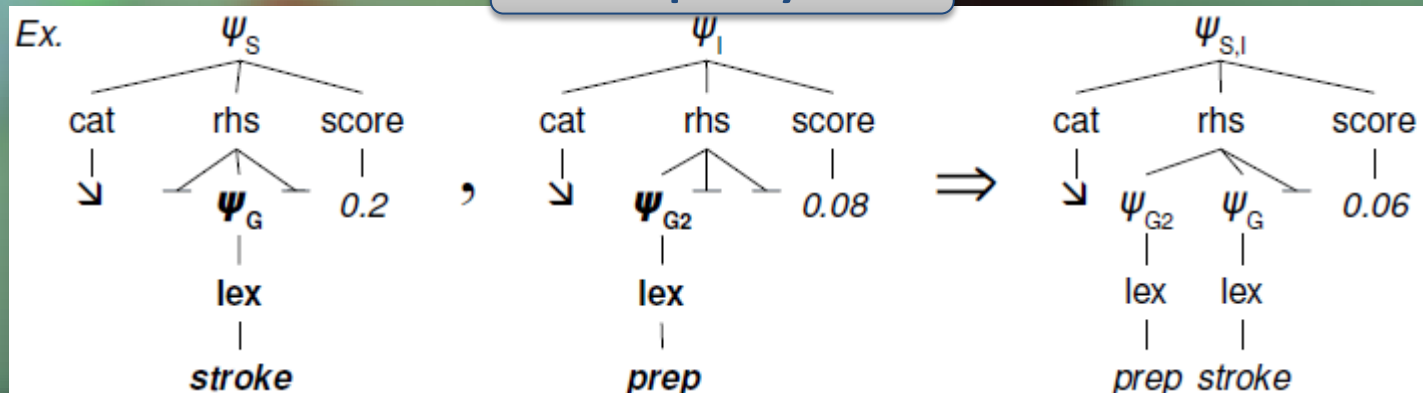


Bidirectional parsing

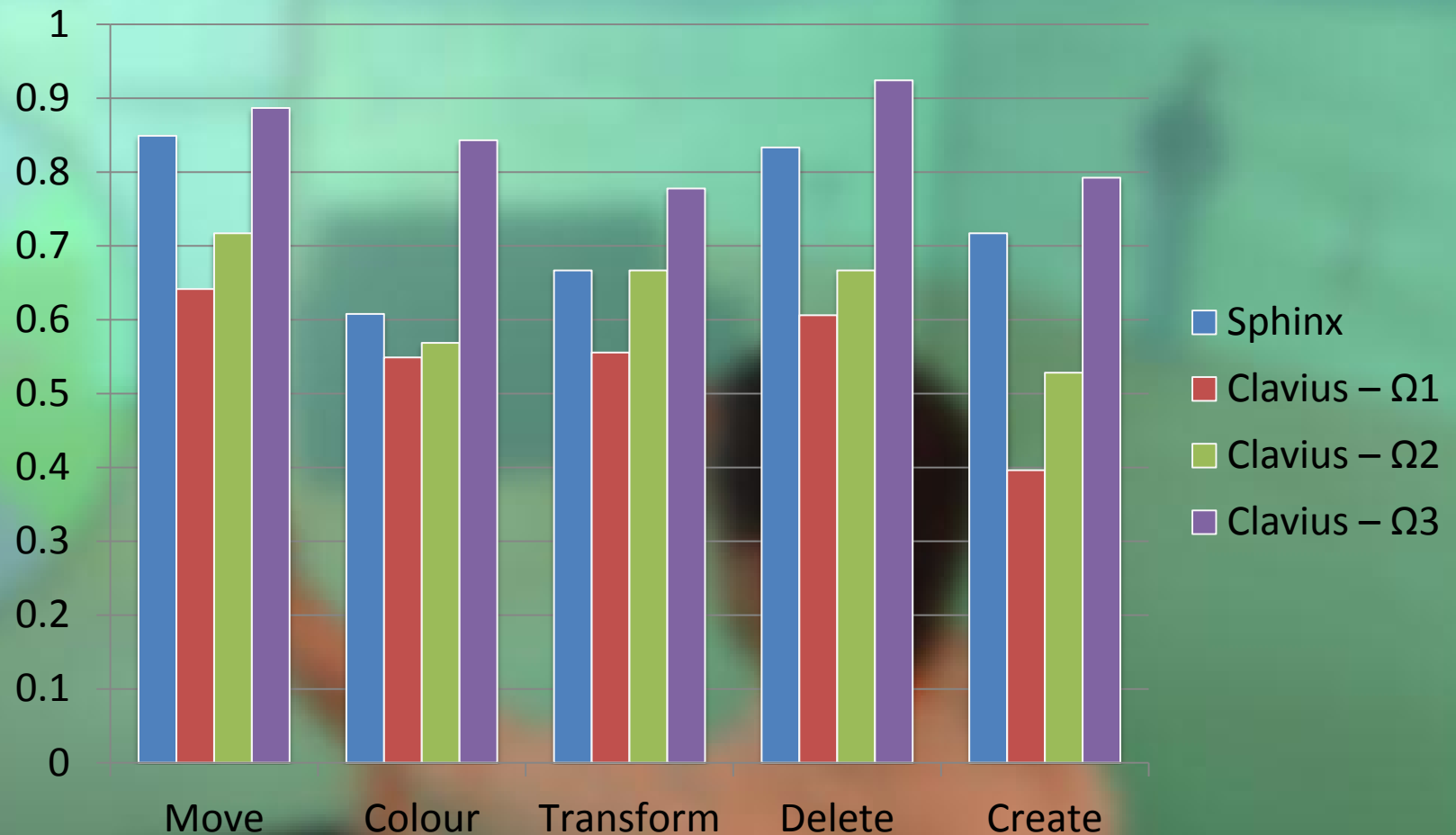
Generalize



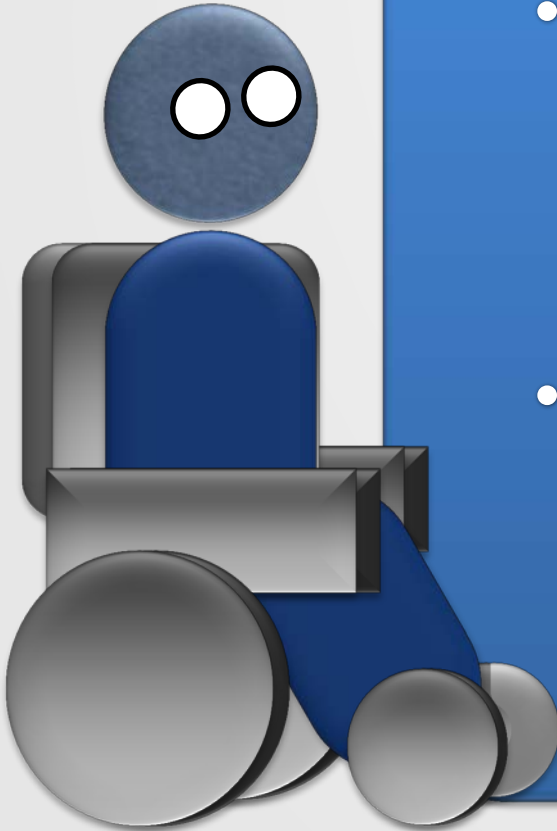
Specify



Bidirectional parsing



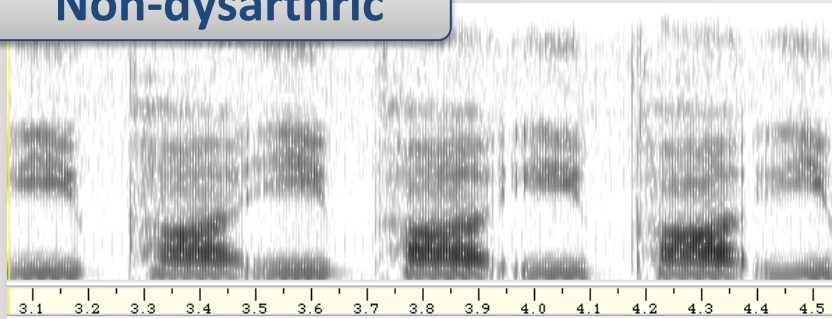
Dysarthria



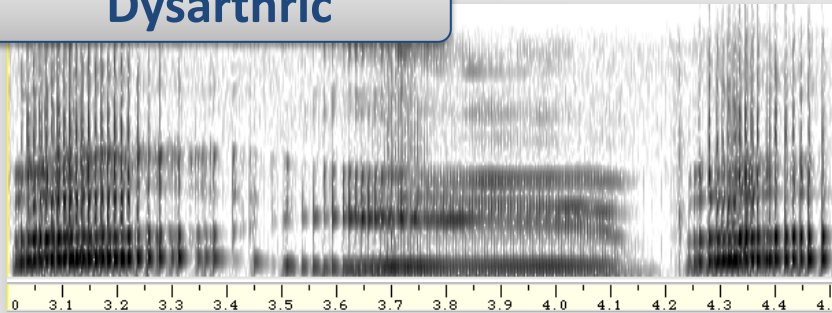
- Articulatory difficulties resulting in unintelligible speech.
 - *Causes* (e.g.)
 - Cerebral palsy,
 - Parkinson's disease,
 - Amyotrophic lateral sclerosis.
 - *Effects* (e.g.)
 - Poor respiration and phonation,
 - Inaccurate timing and speed,
 - Distorted or unintelligible vowels,
 - Atypical control of volume and pitch.

TORGO – articulatory speed

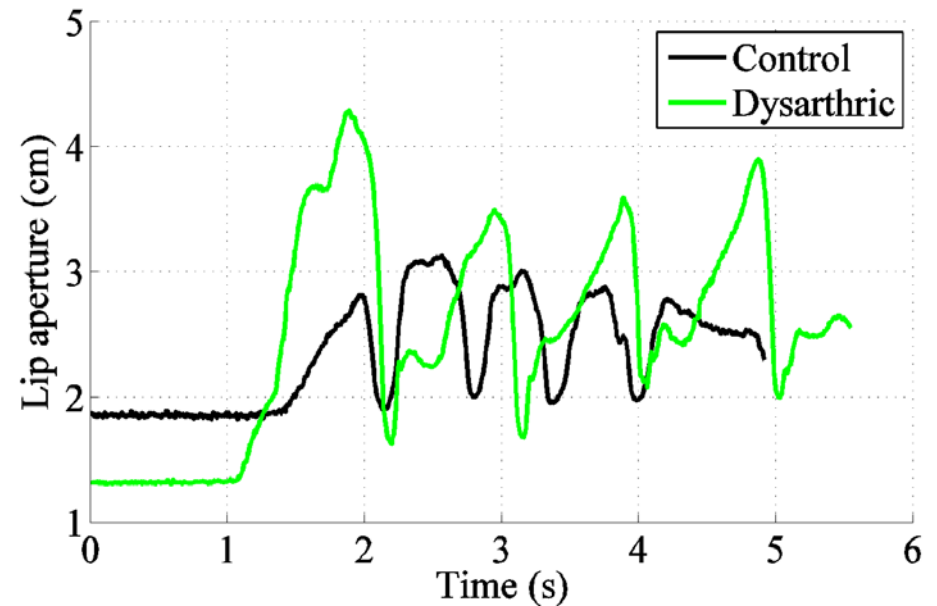
Non-dysarthric



Dysarthric



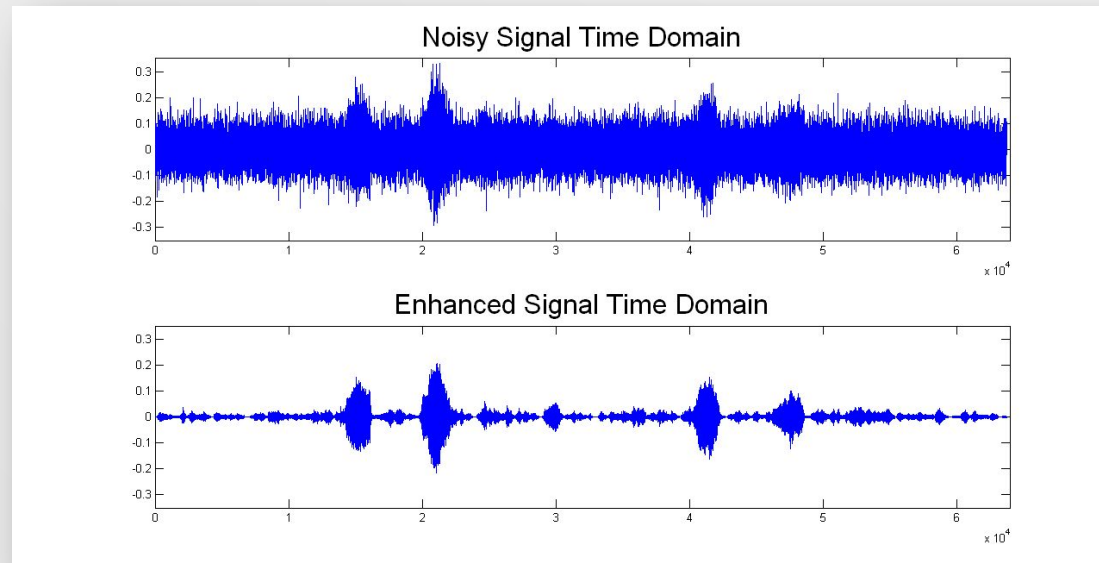
“iypah, iypah, iypah, ...”



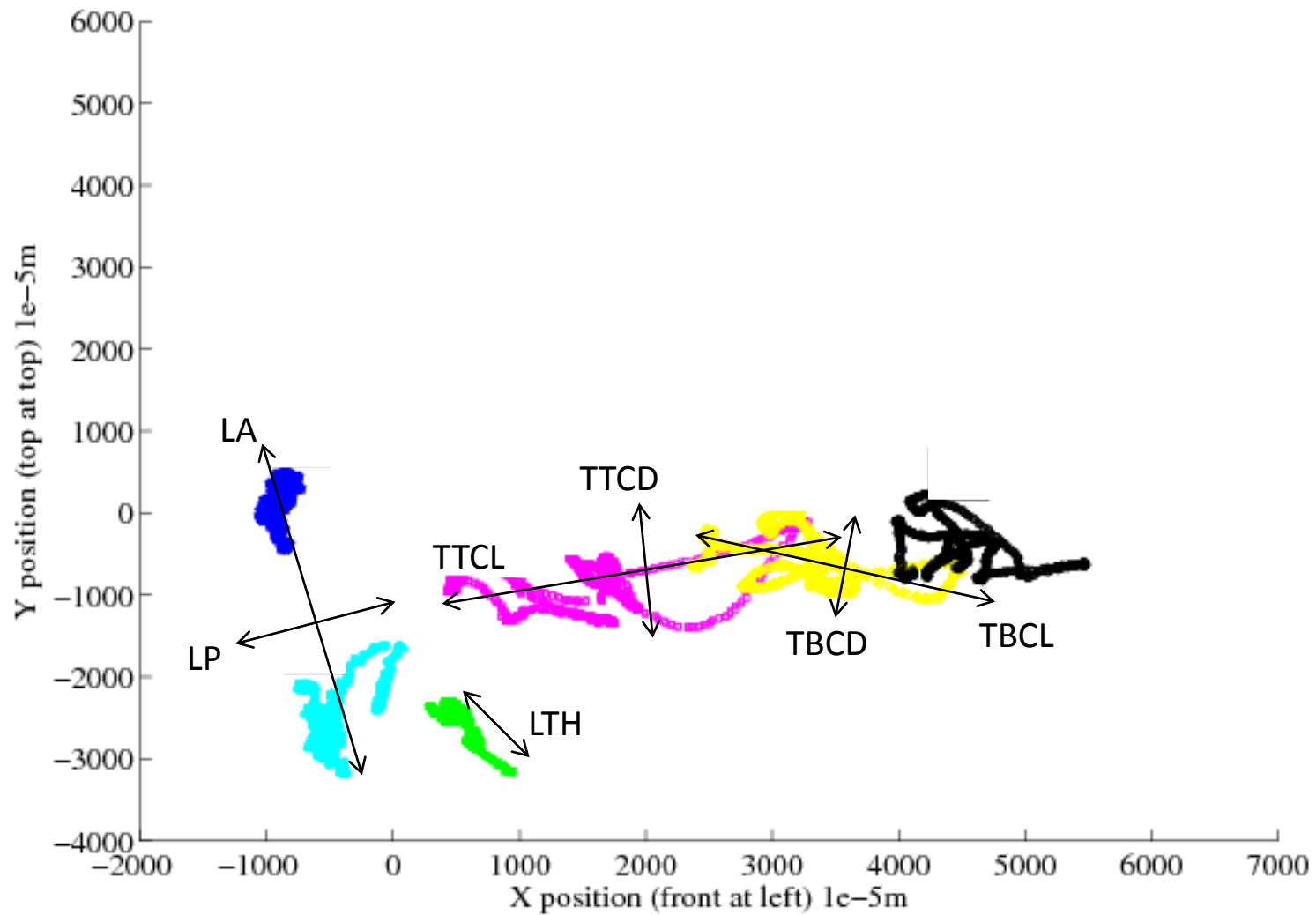
“ahpi, ahpi, ahpi, ...”

Removing acoustic noise

- Especially with older recordings, we should remove additive acoustic noise from the signal.
 - $s(t) = a(t) + v(t)$ signal s , true signal a , noise v
 - $a(t+1) = a(t) + u(t)$ process variation u
- Minimum mean-squared spectral subtraction.

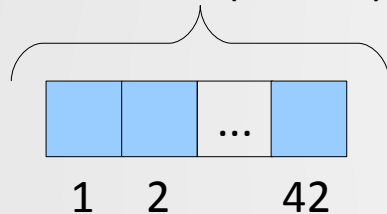


Extracting TVs

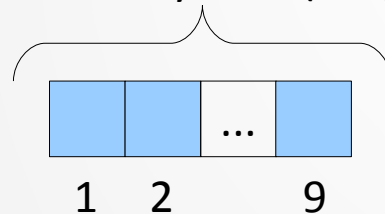


Quantizing articulation data

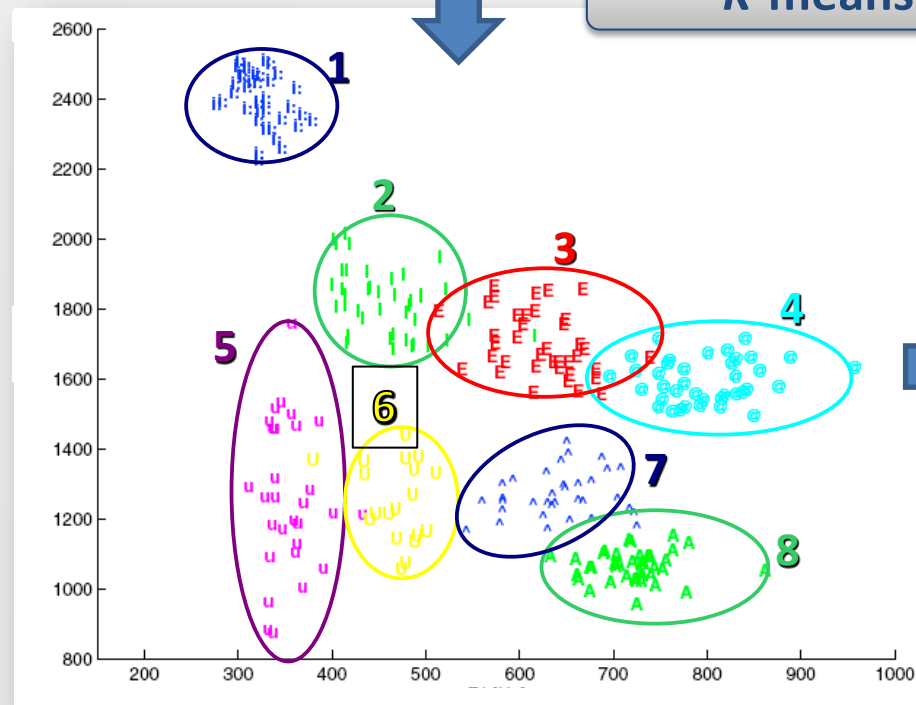
Acoustic data (MFCCs)



Articulatory data (TVs)



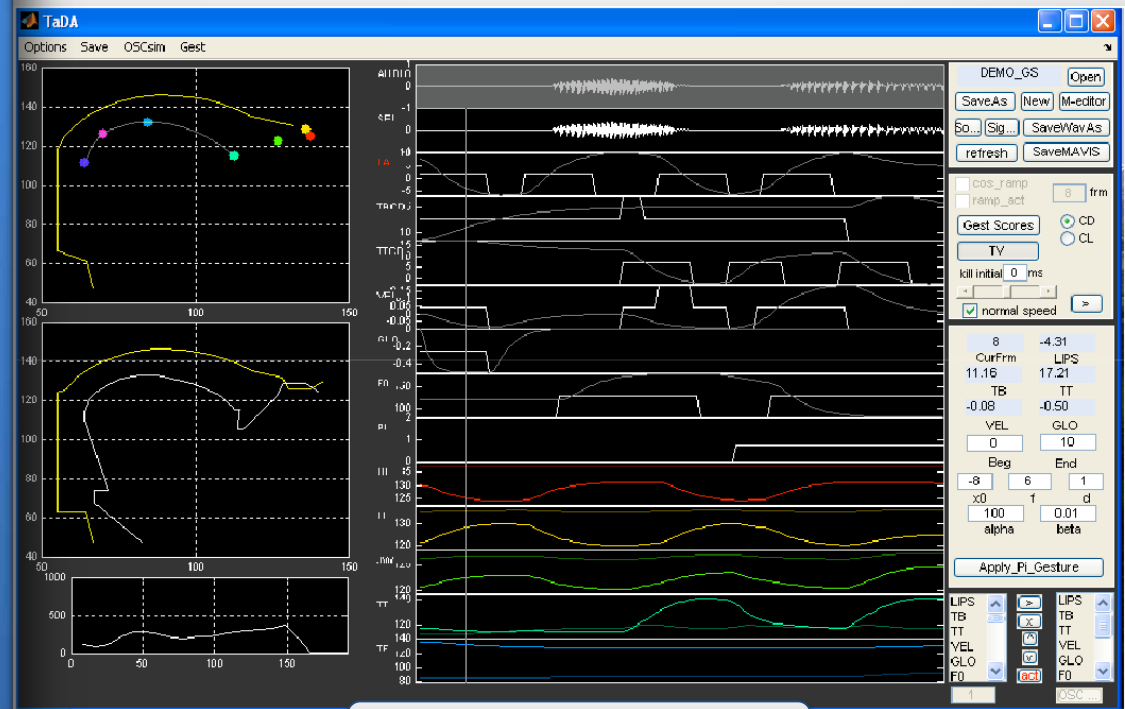
K-means



index

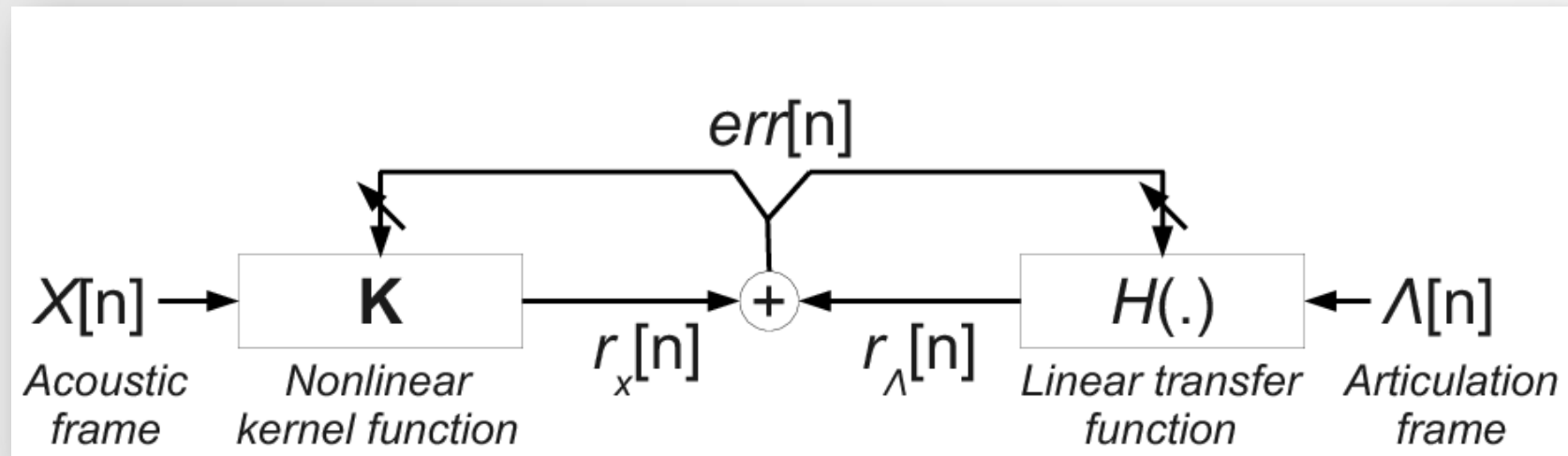
Experiments using TADA

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

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

