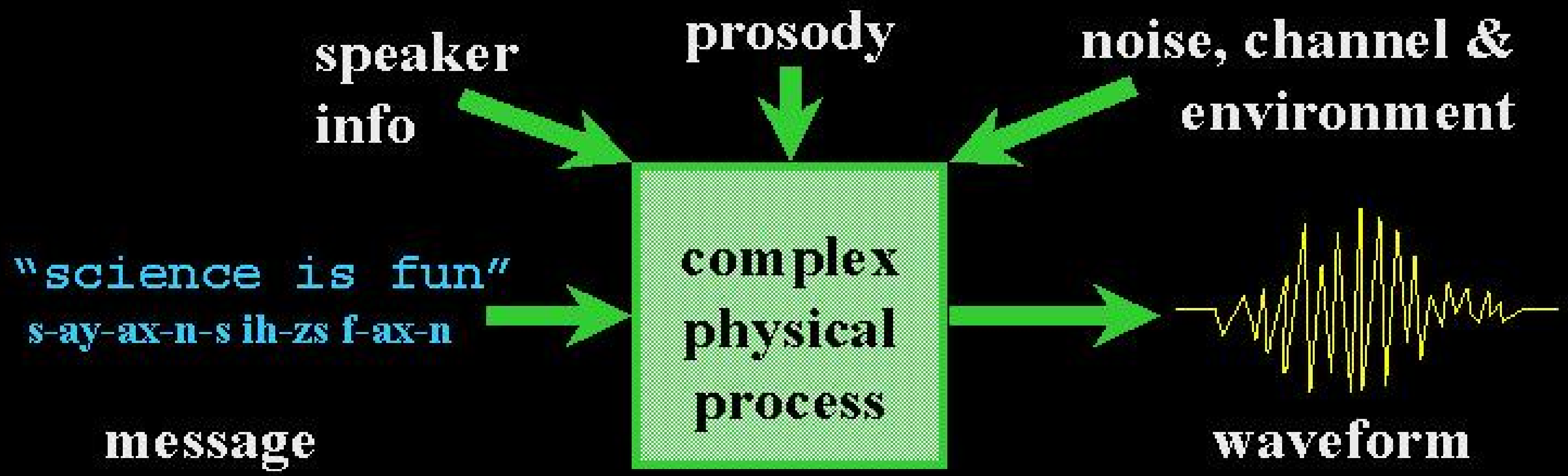
Articulatory Speech Processing

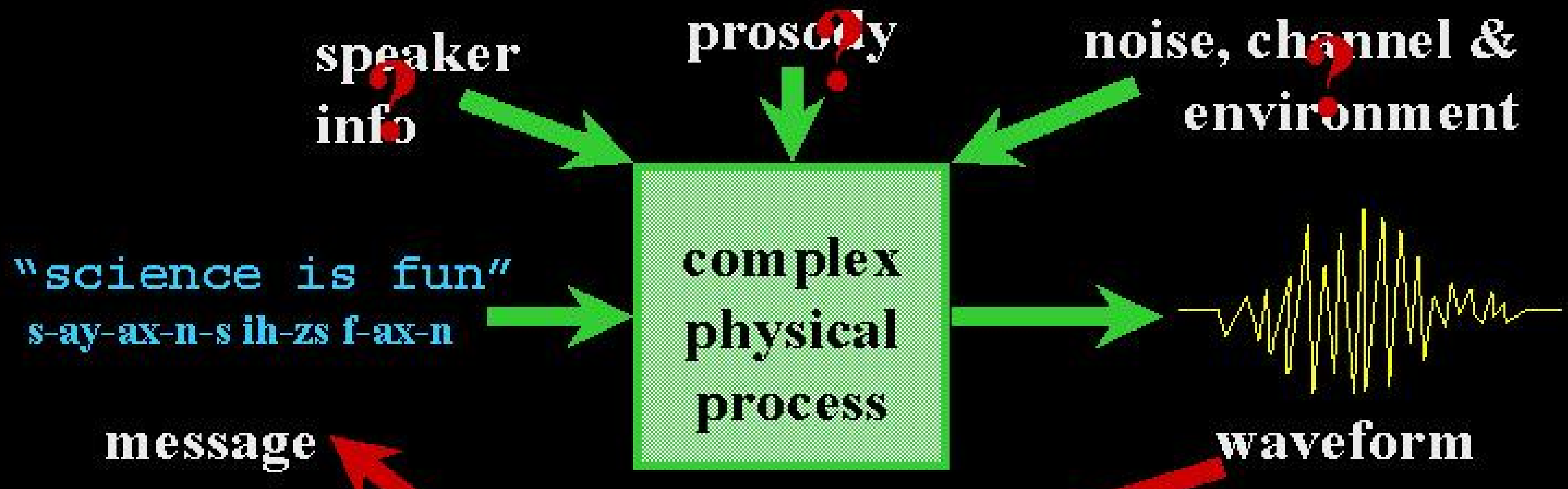
Sam Roweis

Speech processing



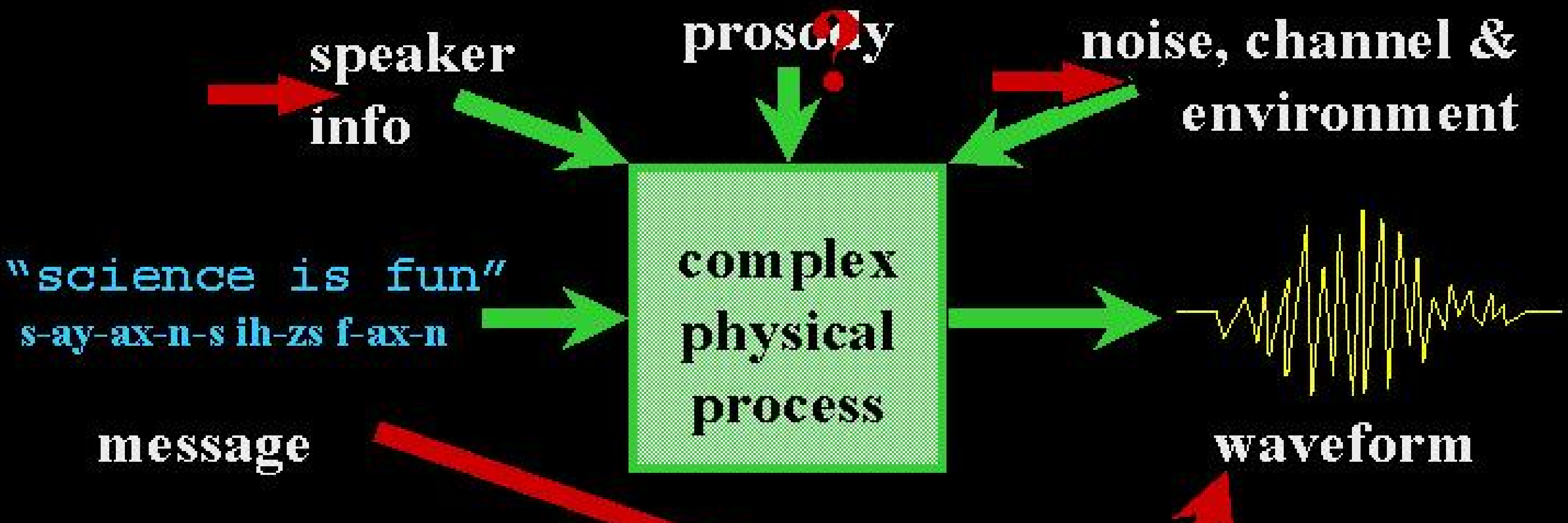
- problems: recognition synthesis speaker id
- How does the human brain solve these problems? How can we build machines which also solve them?

Speech processing



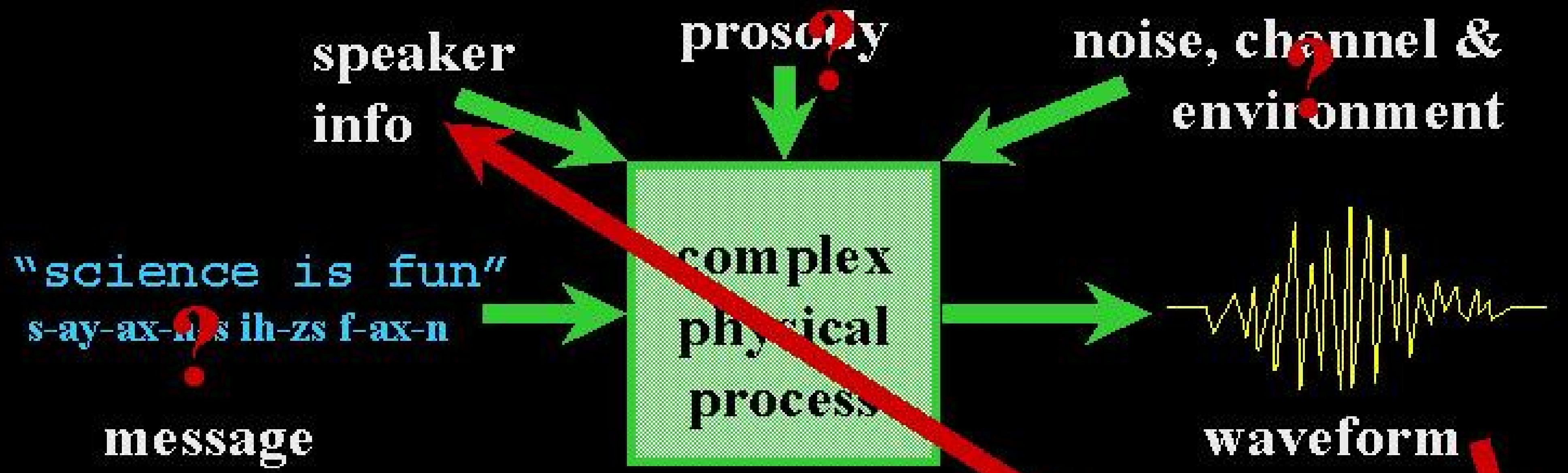
- problems: recognition synthesis speaker id

Speech processing



- problems: recognition synthesis speaker id

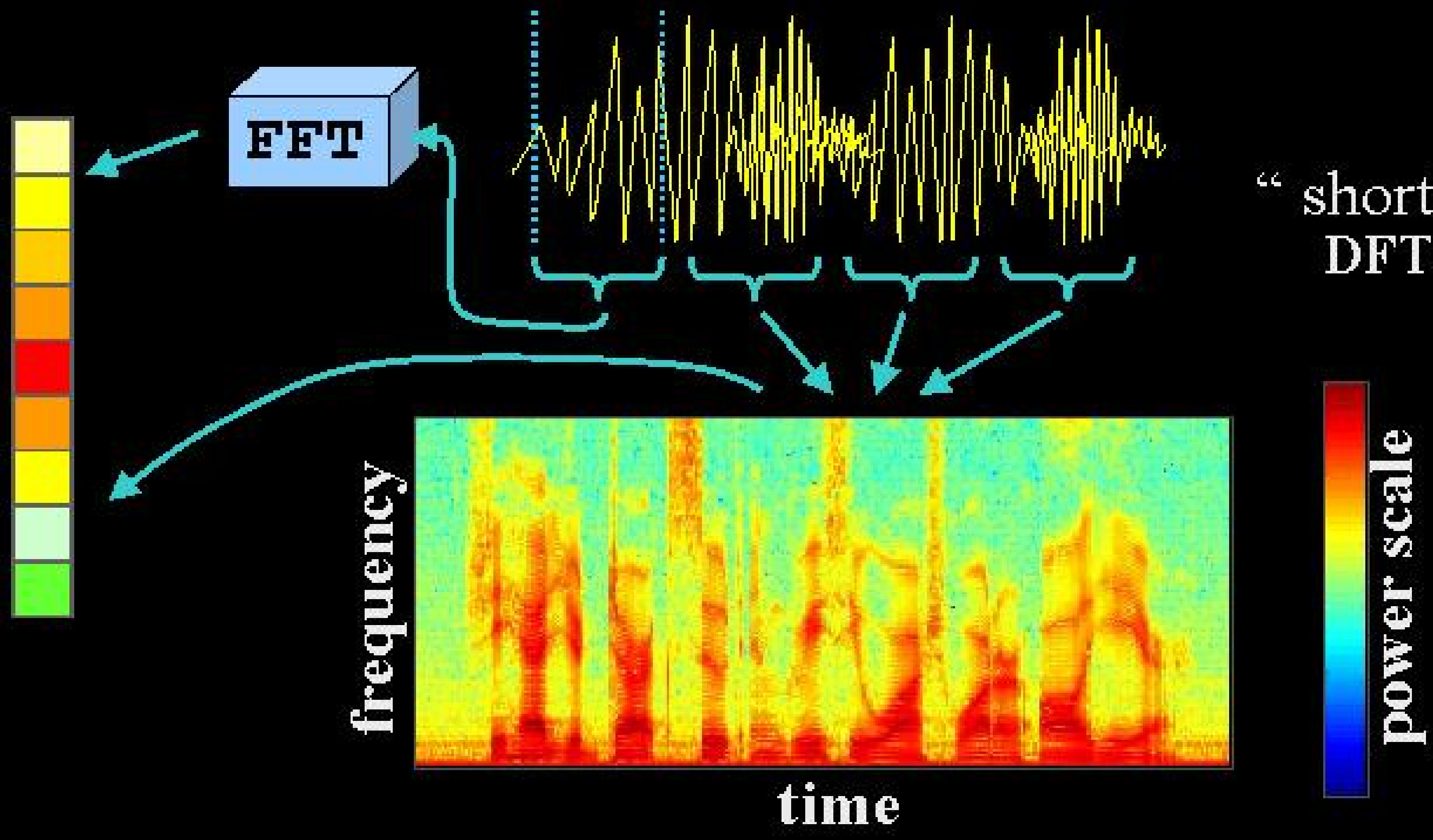
Speech processing



- problems: recognition synthesis speaker id

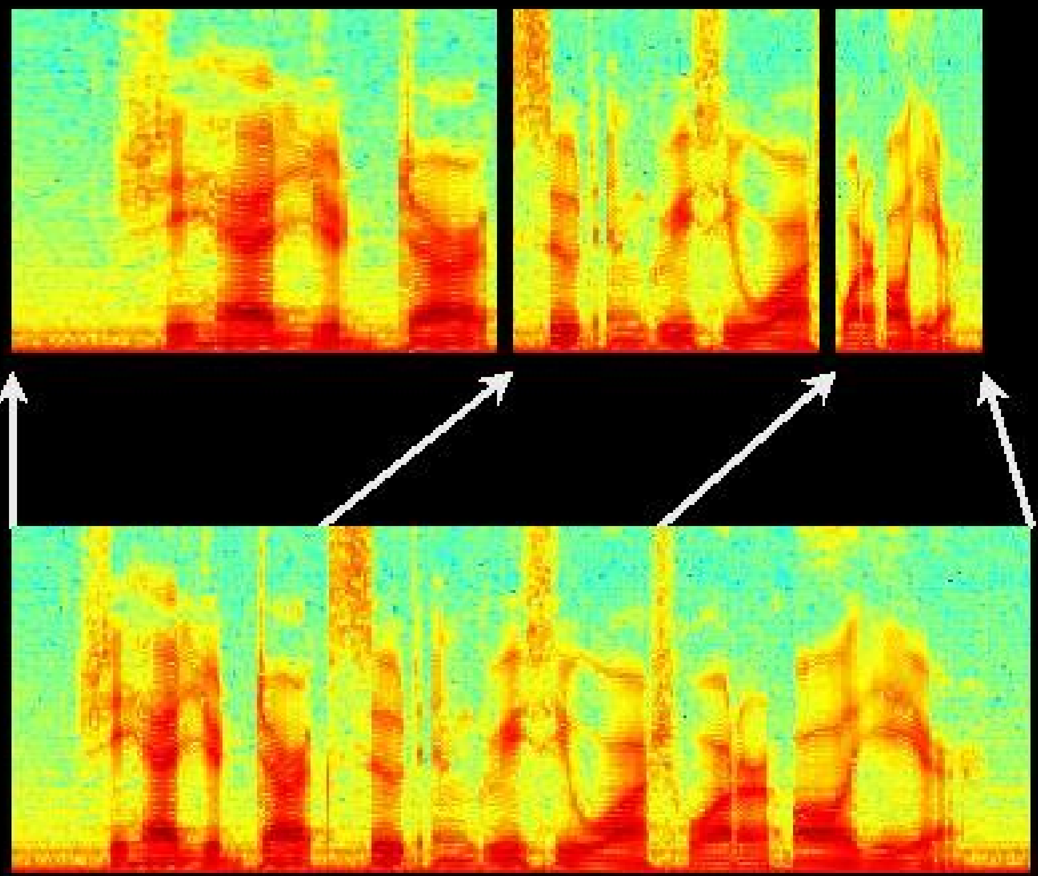
Current approach: feature extraction

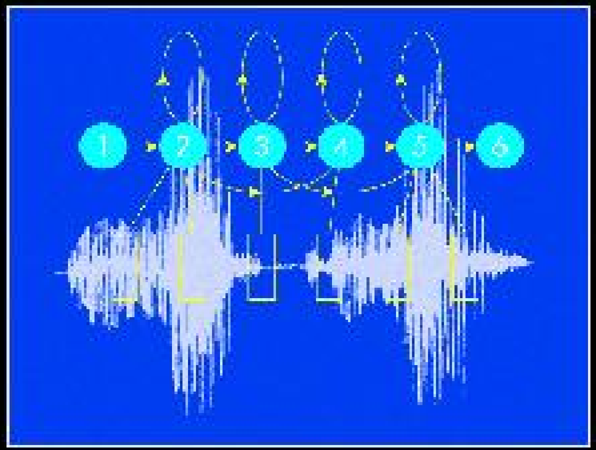
- Spectrograms: energy in time-frequency plane



Current approach: templates

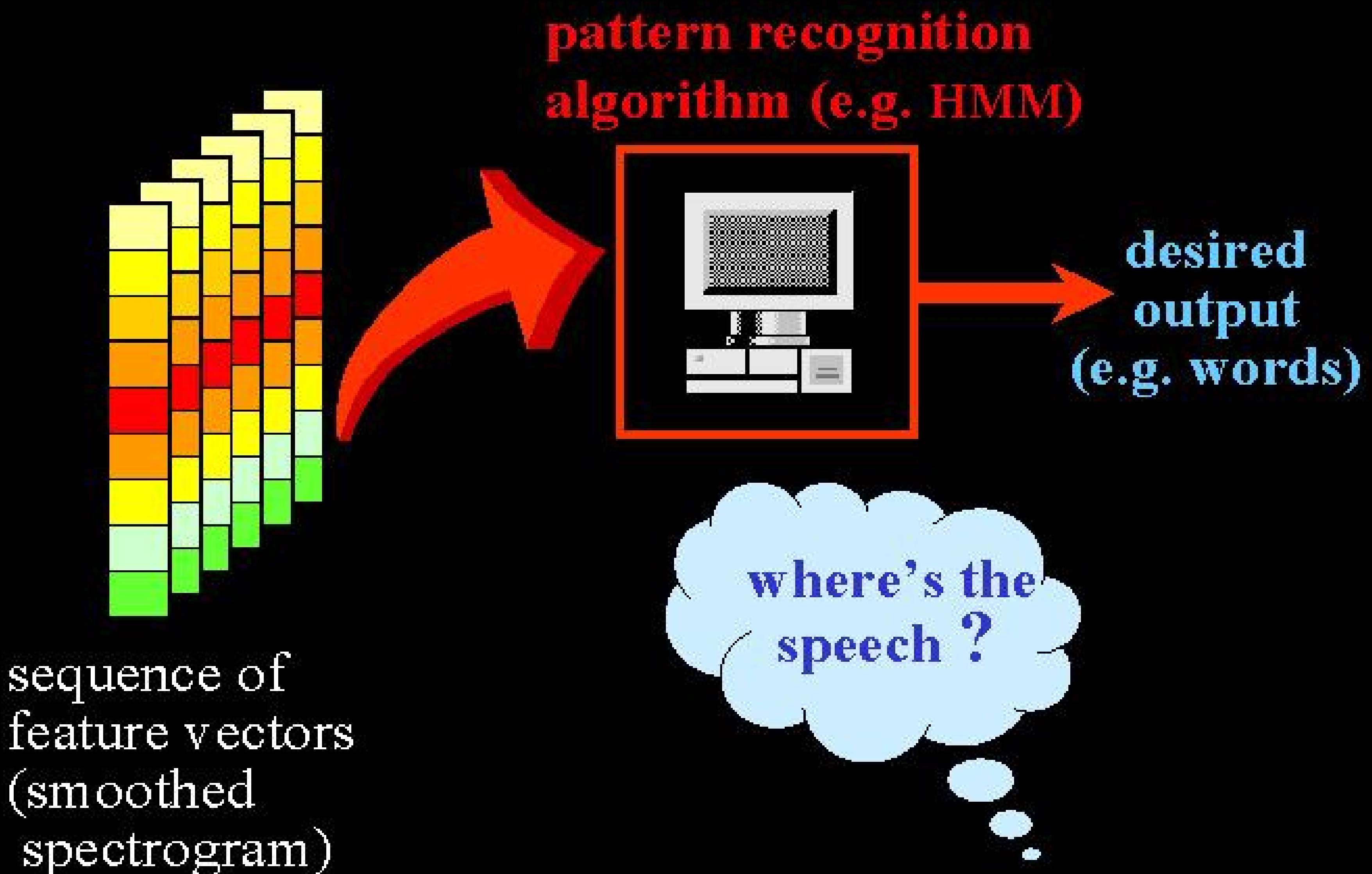
- Dynamic Time Warping (DTW)
  Hidden Markov Models (HMMs)



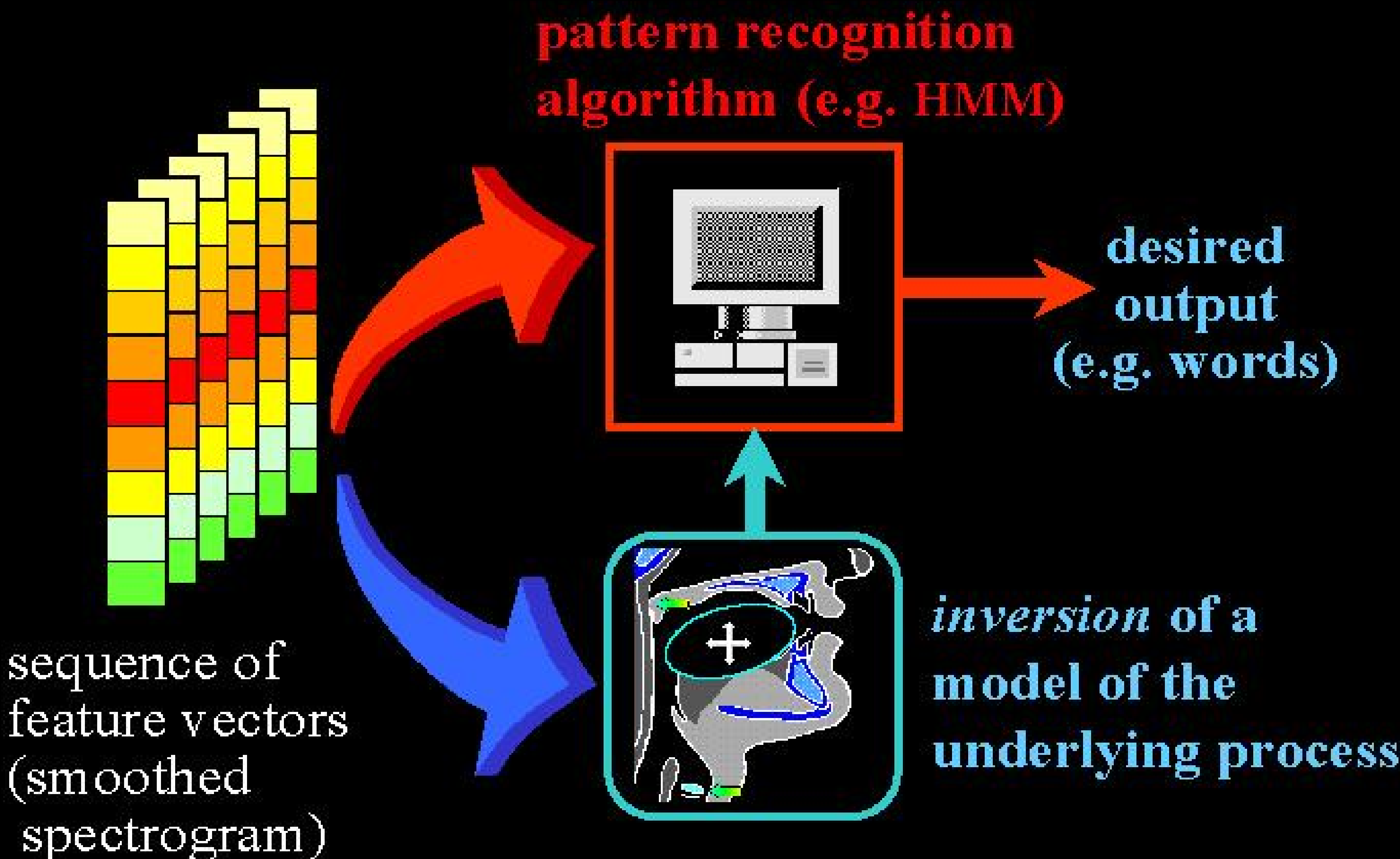


Spectrogram templates with local stretching and squishing plus noise.

An engineering objection

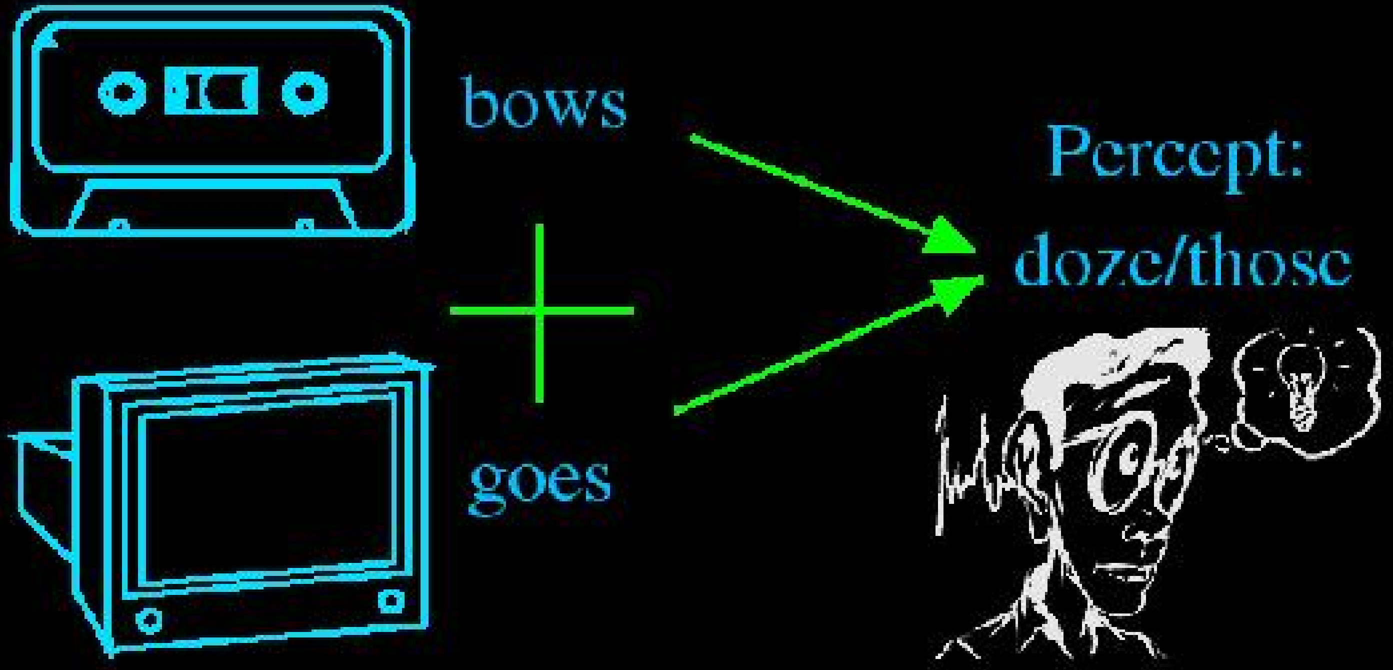


New idea: use a model !



Linguistics & psychophysics

- "Motor theory" of speech (Liberman et al.)
- Lipreading
- McGurk–MacDonald effect

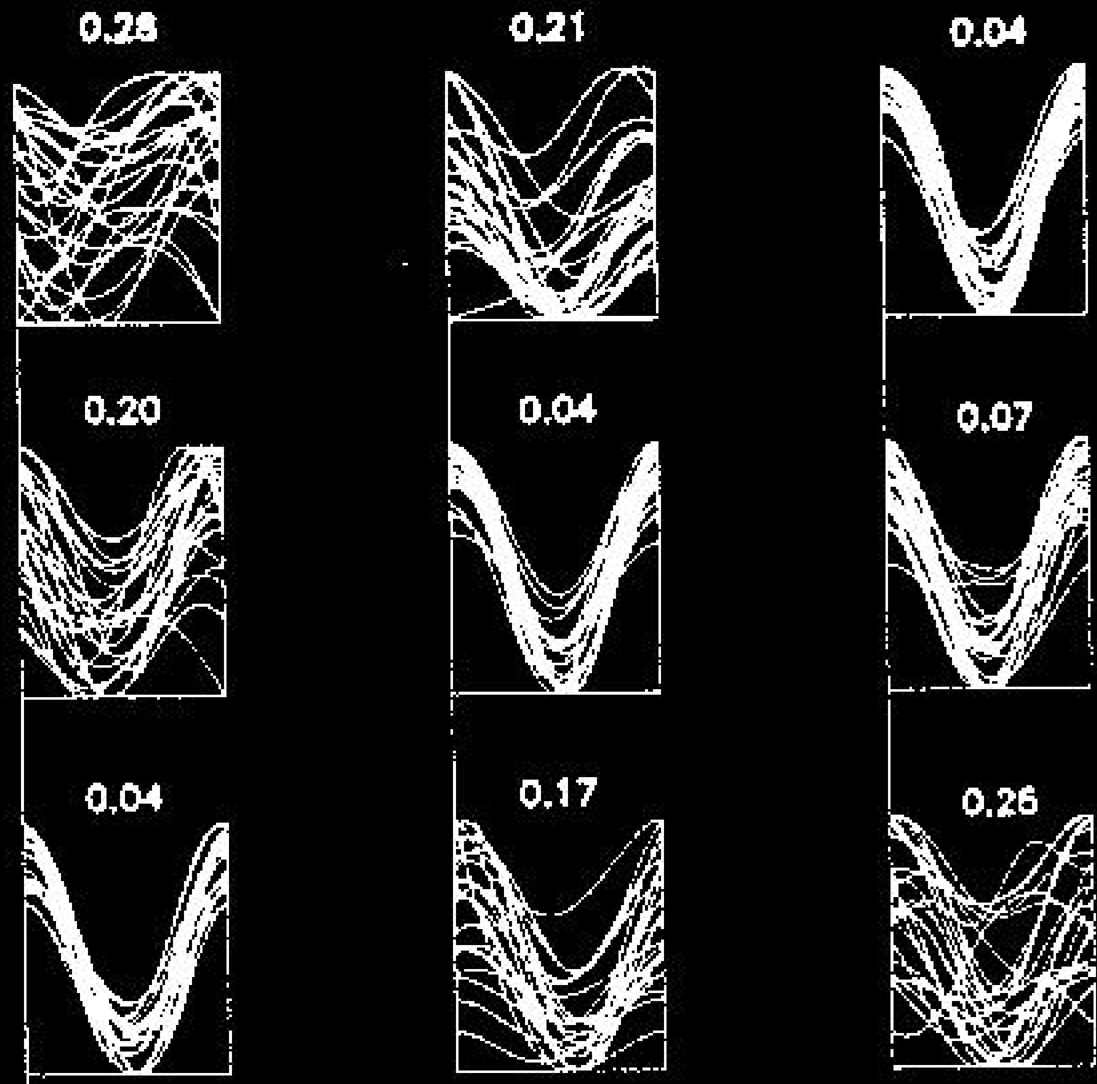


Is variability easier to model in the articulatory domain?

tongue dorsum

tongue tip

lower lip

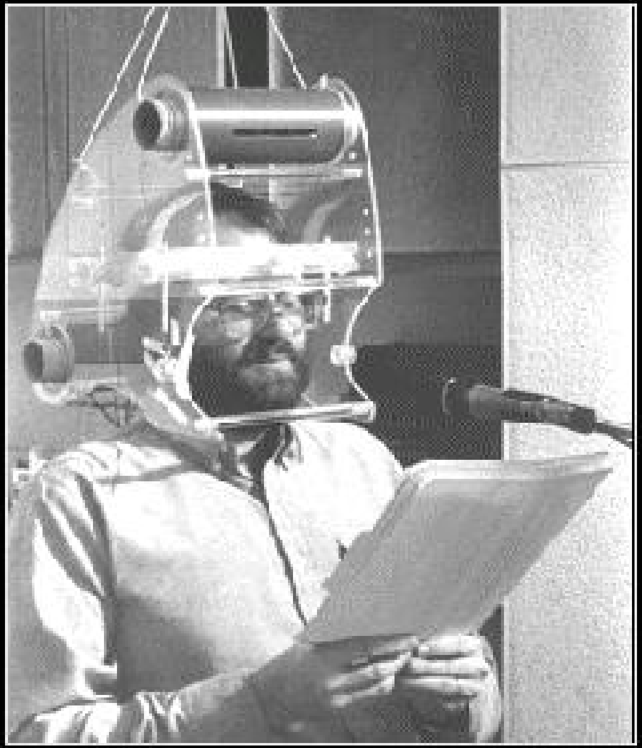


3 subjects,
3 sound types
/p/,/b/
/t/,/d/
/k/,/g/

(from Papcun et. al)

Approach: analysis of real data

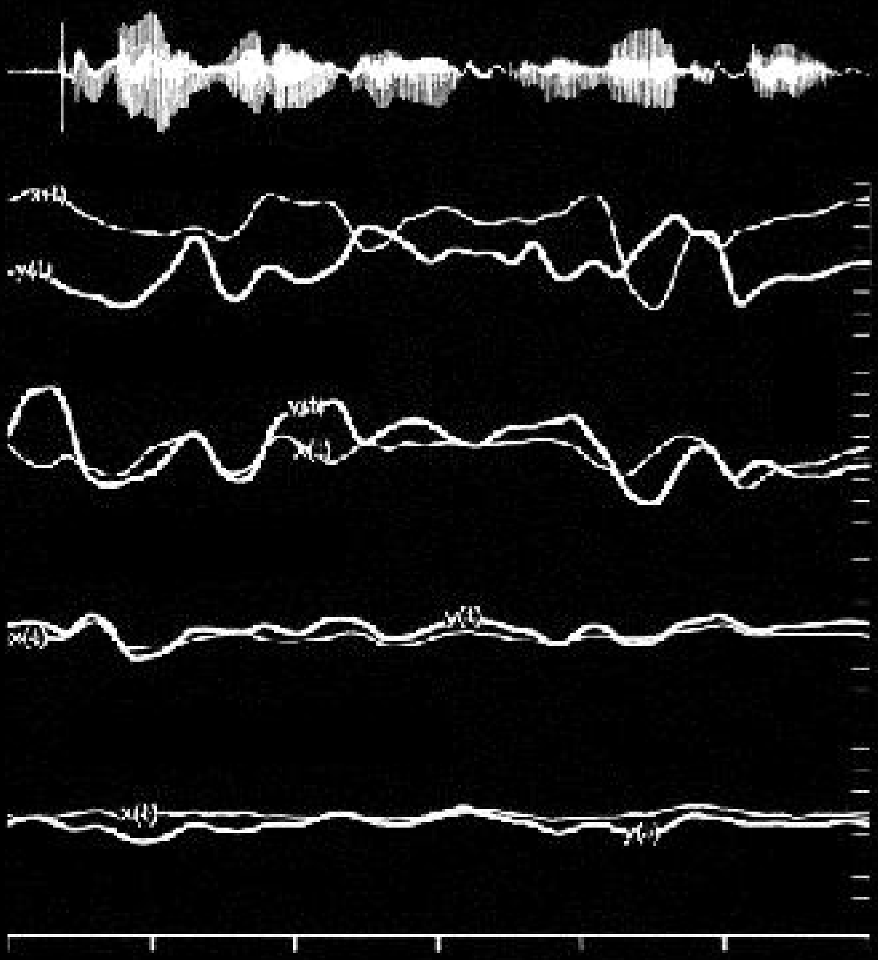
- Look at real speech production data containing simultaneous audio and movement/voicing measurements.
- Advantages:
  - Learning models is easier: supervised problem.
  - Understanding models is easier: we have ground truth.
  - Answer some speech science questions as well as tackle some engineering problems.



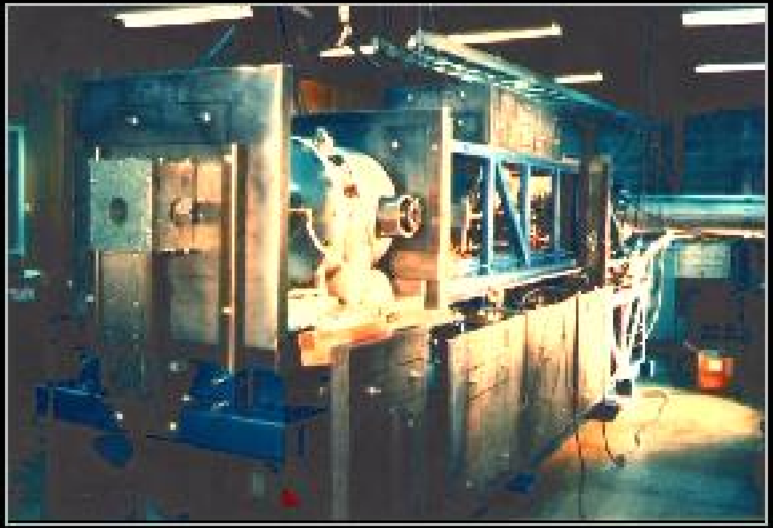
X-ray microbeam database

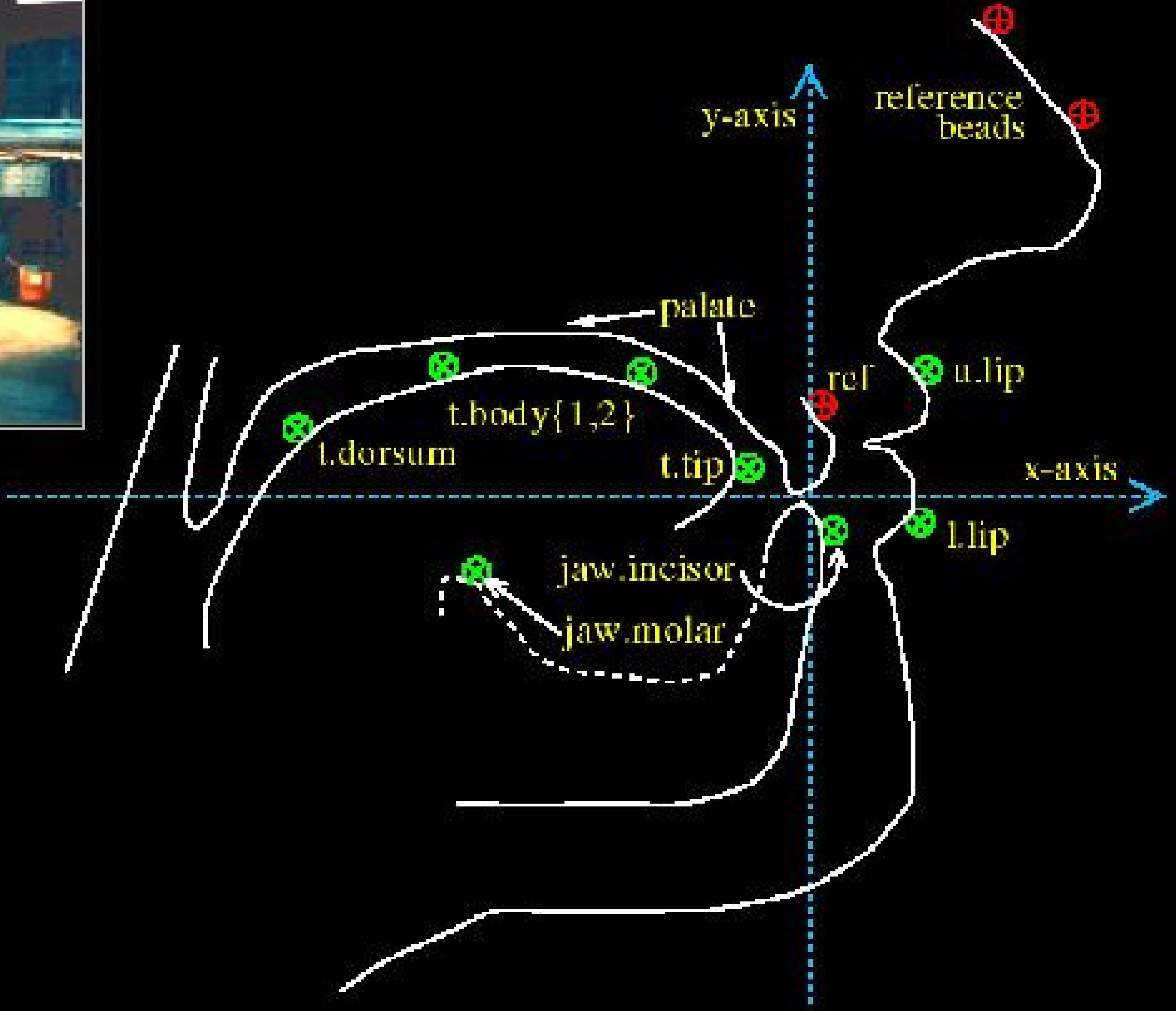
University of Wisconsin

- simultaneous audio + movements
- speech wave (**21 kHz**)
- 8 beads (**146 Hz, 1mm**)
- also video, voicing
- 32 women, 25 men
- midsaggital only



Placement of tracking beads

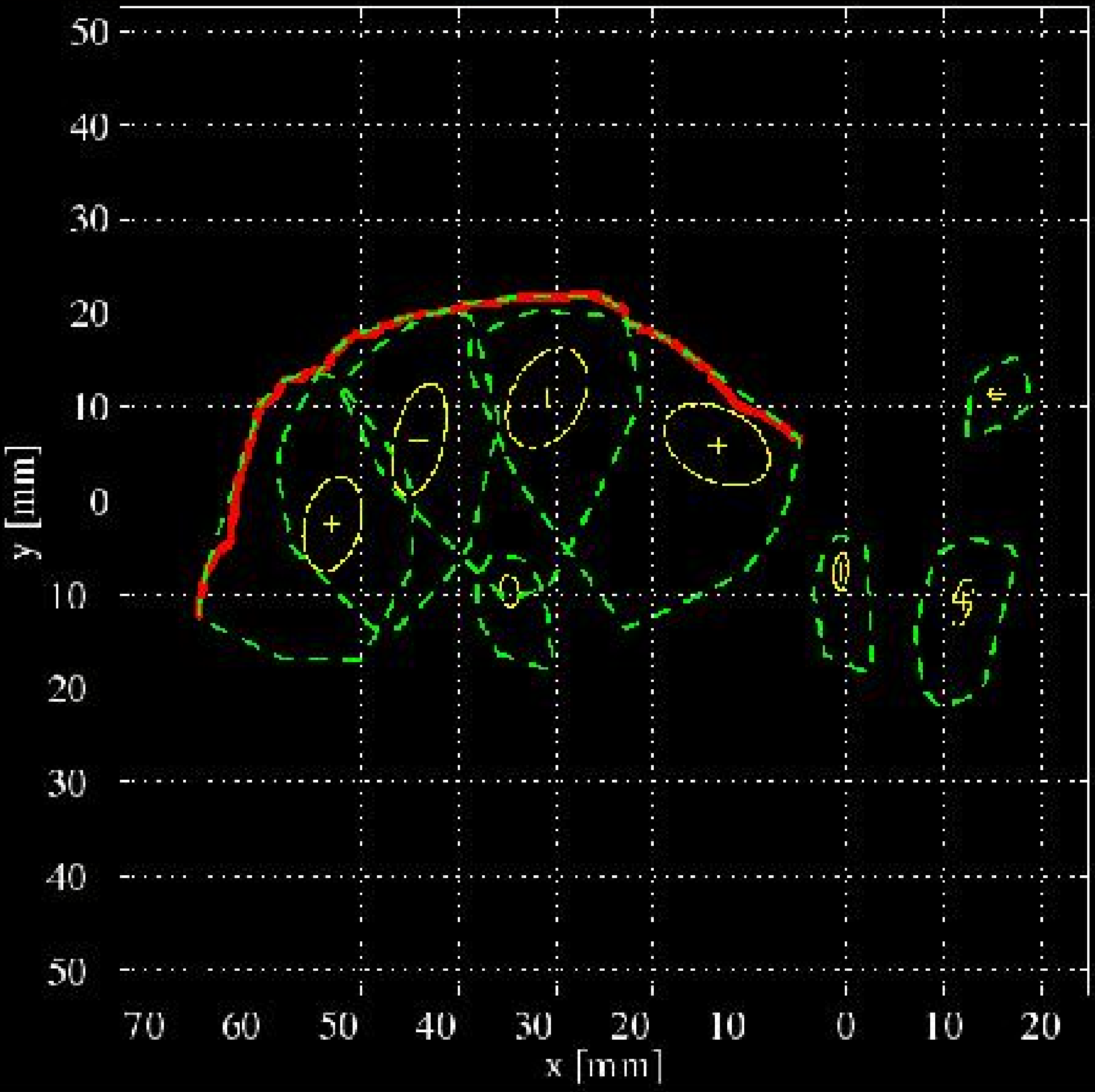




Example data & simple statistics

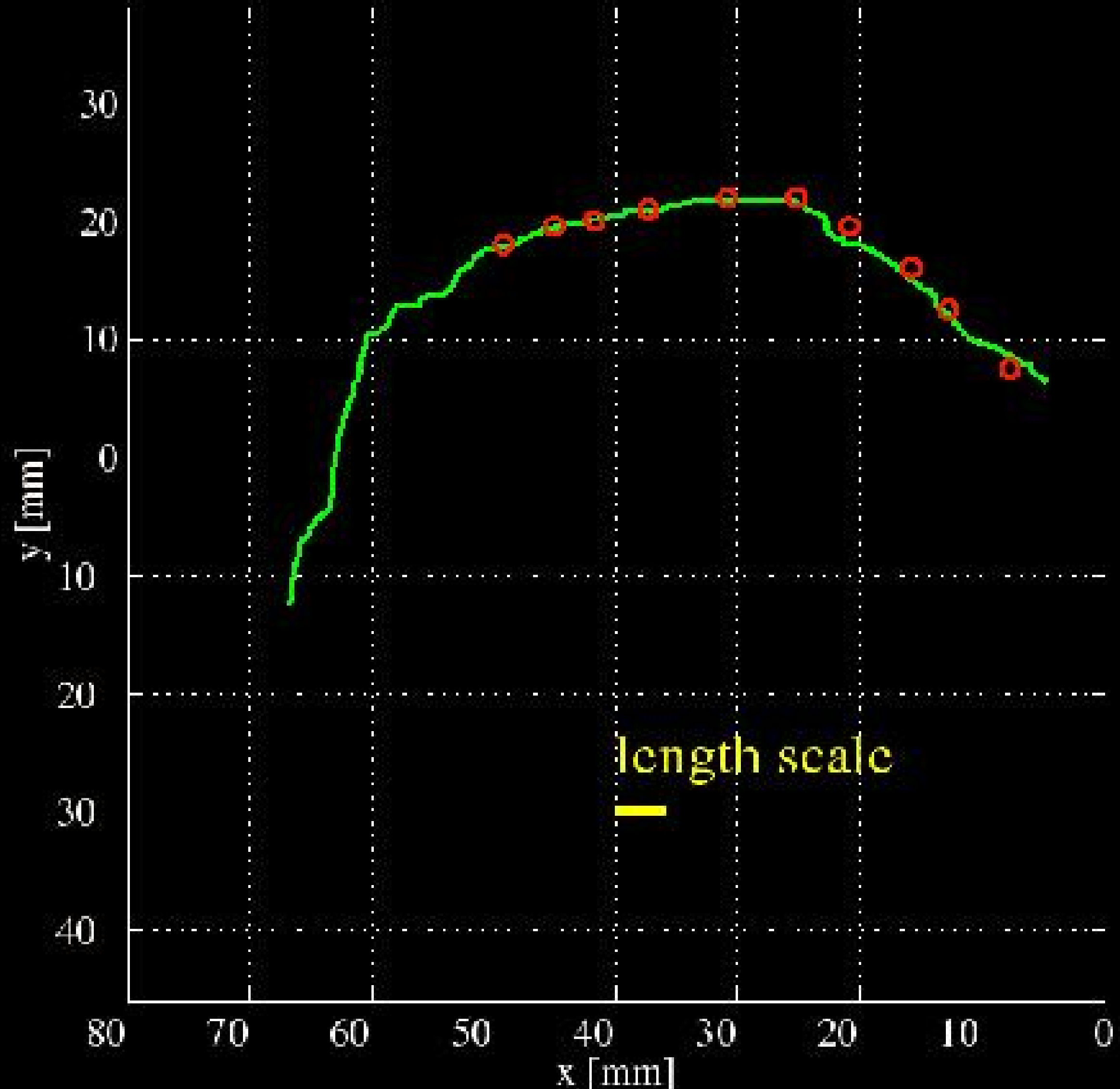
- Audio
- Bead movies

- Means
  Covariances
  Convex hulls



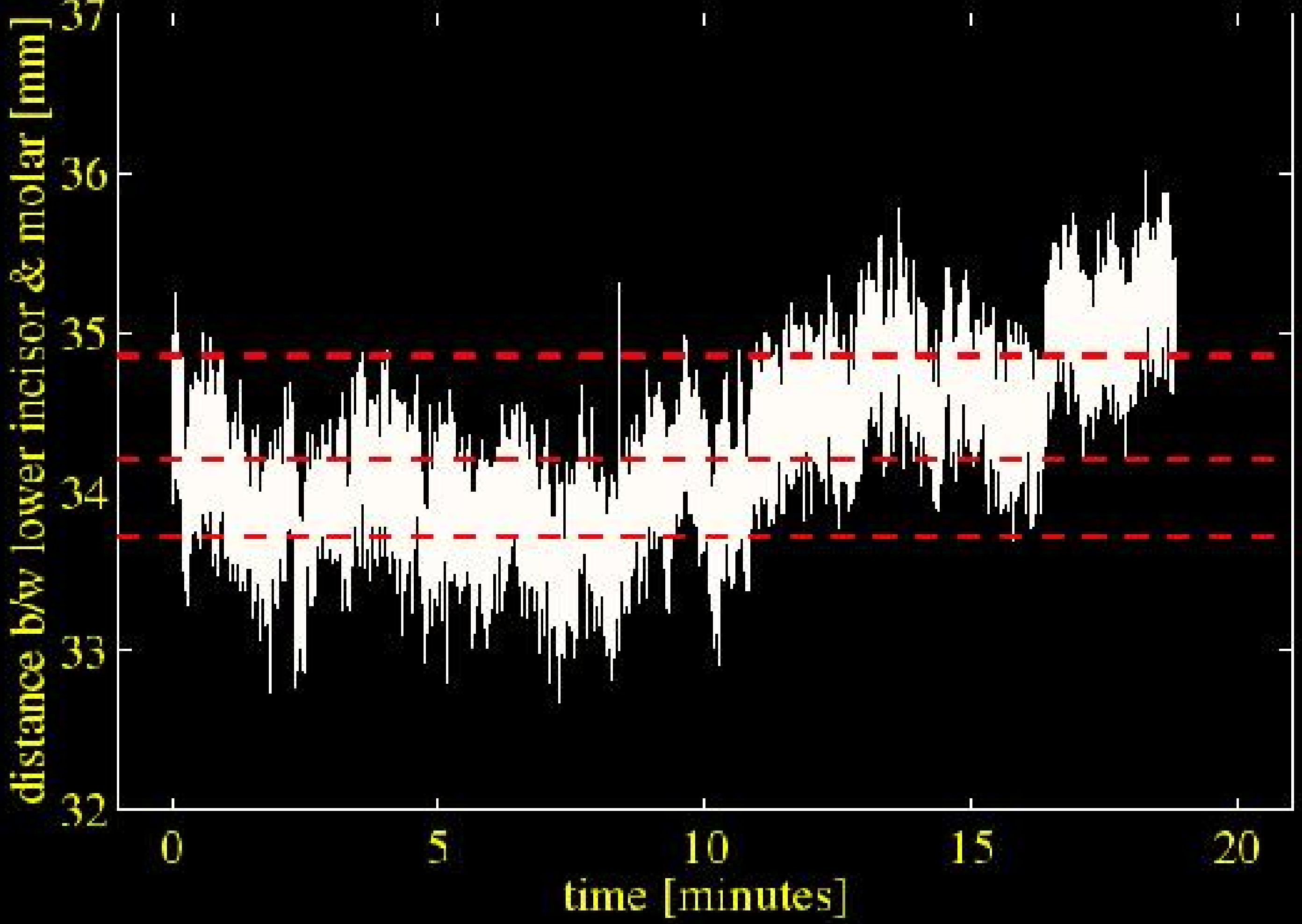
Palate estimation

- An automatic algorithm estimates the palate (line)
- Compares well with the few measured points (circles)



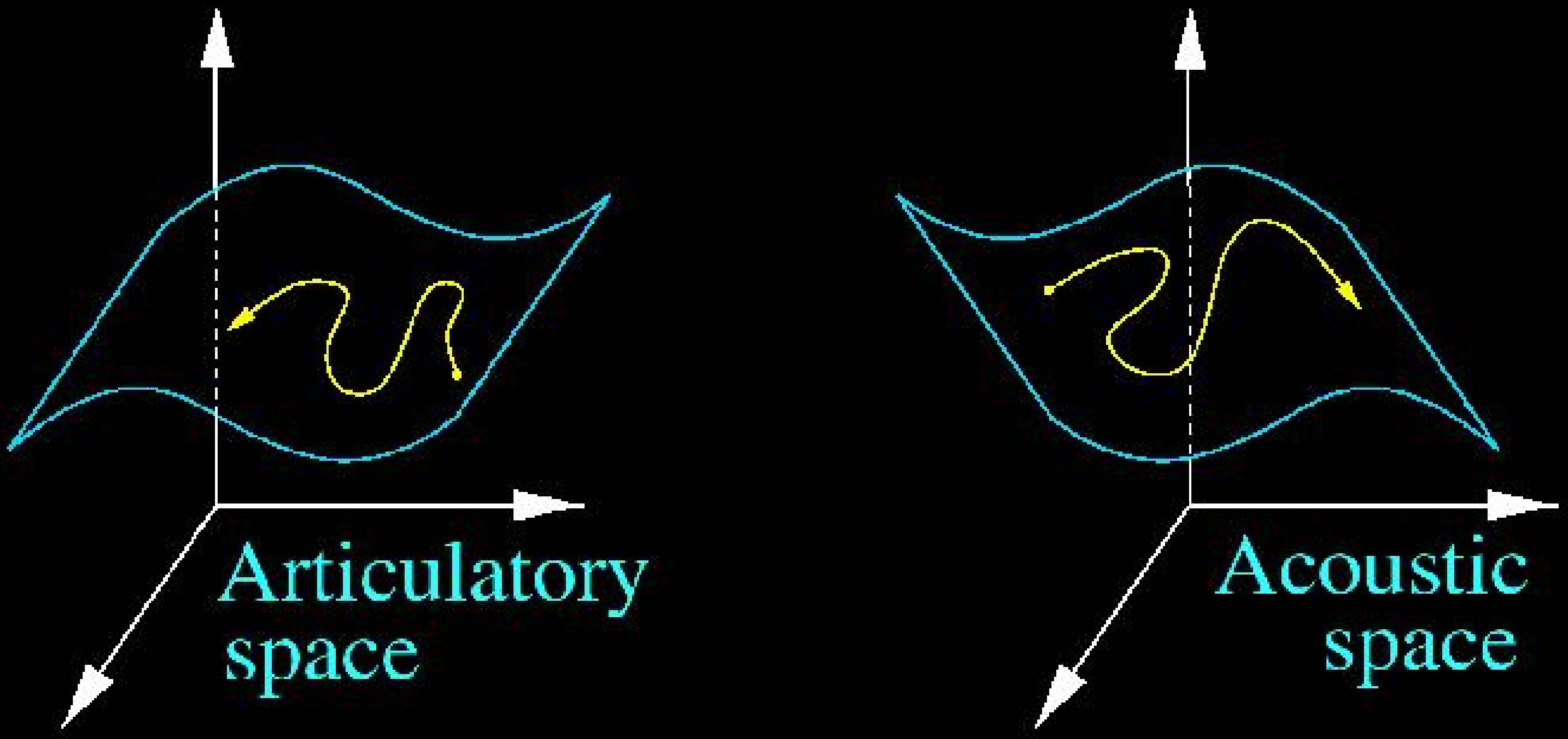
Error level estimation

- Look at a nominally constant value



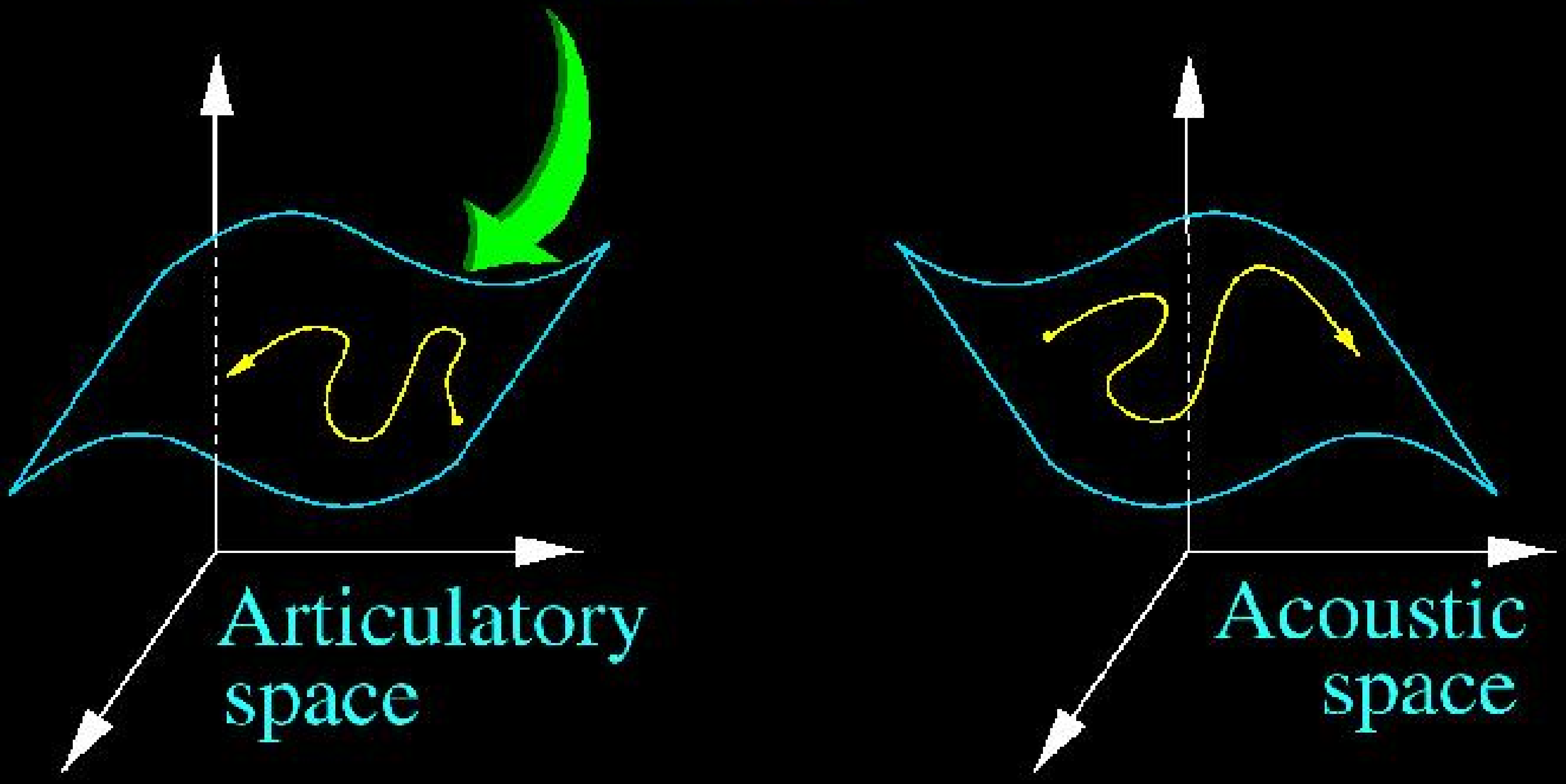
Graphical interpretation

- Each utterance in database can be thought of as a thread in two parallel spaces.



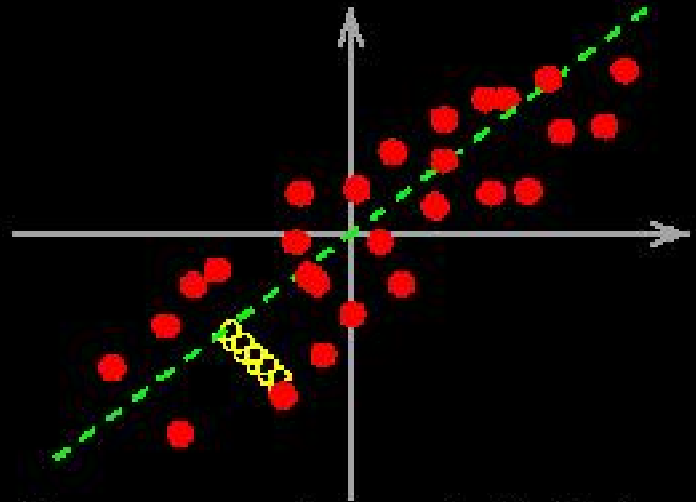
Typical shapes of the mouth

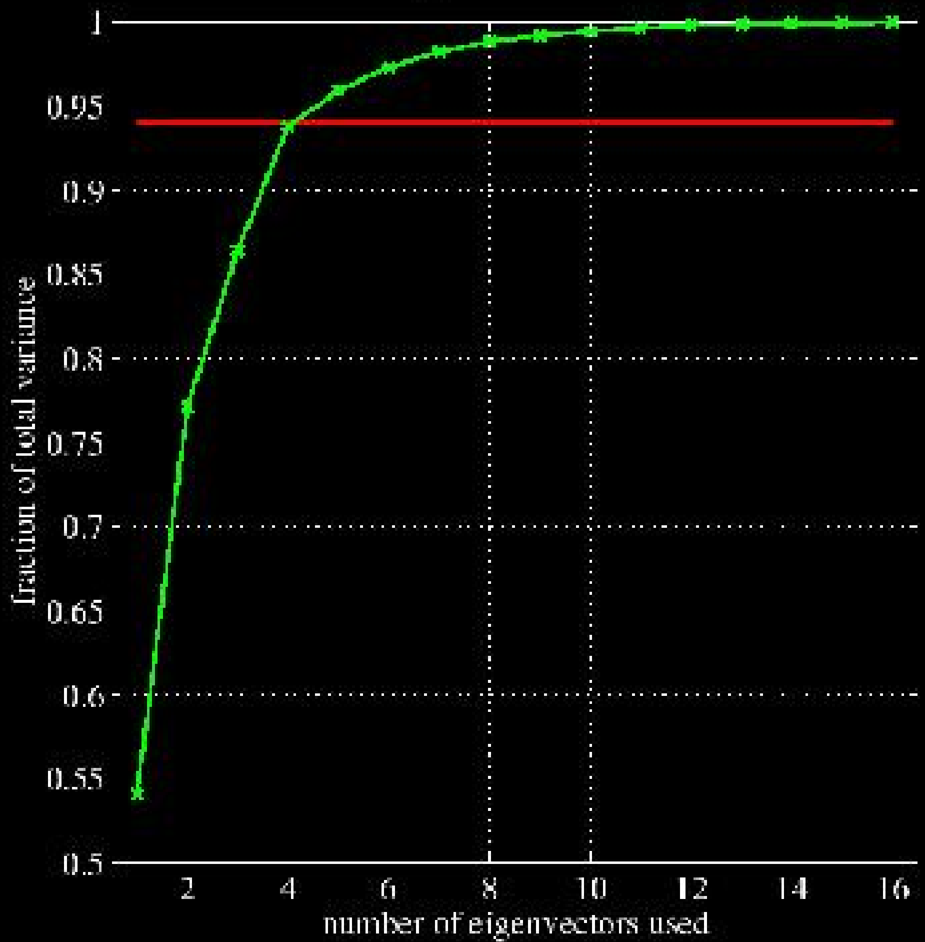
- What does this **manifold** look like?



Most basic manifold model: PCA

- Reduced dimensionality linear model (hyperplane)
- Results: 4–6 dimensions needed for speech data

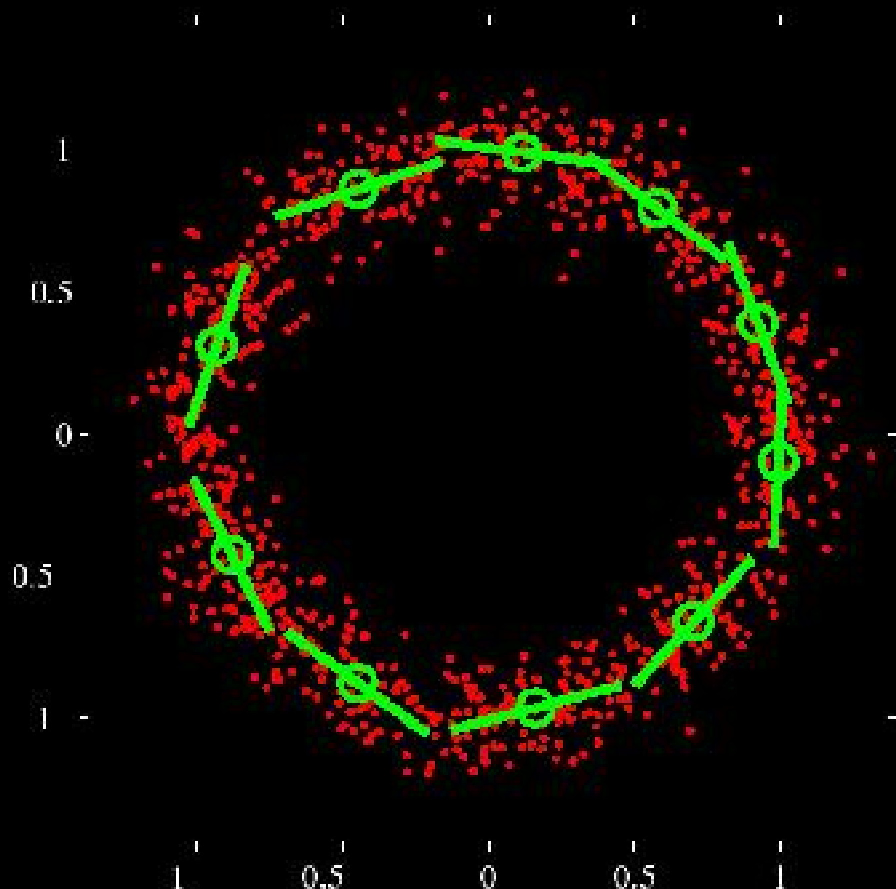


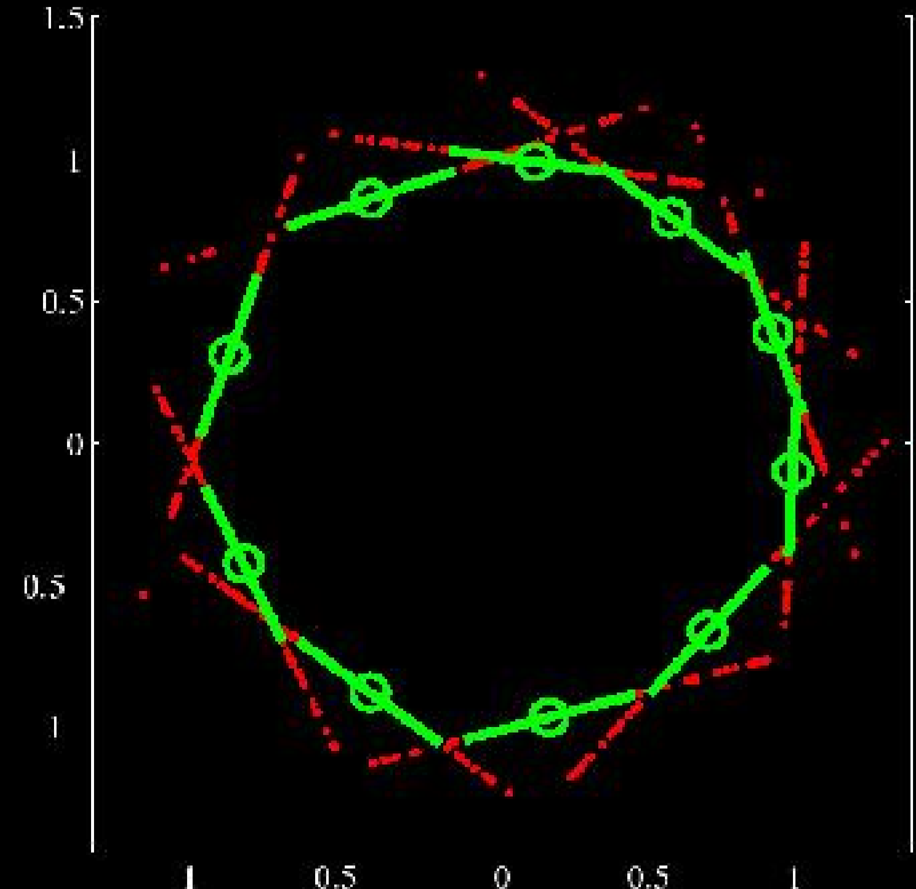


- Conventional PCA has two problems:
  - Lots of data (EMPCA)
  - Need probabilistic model (SPCA)

Mixtures of linear manifolds

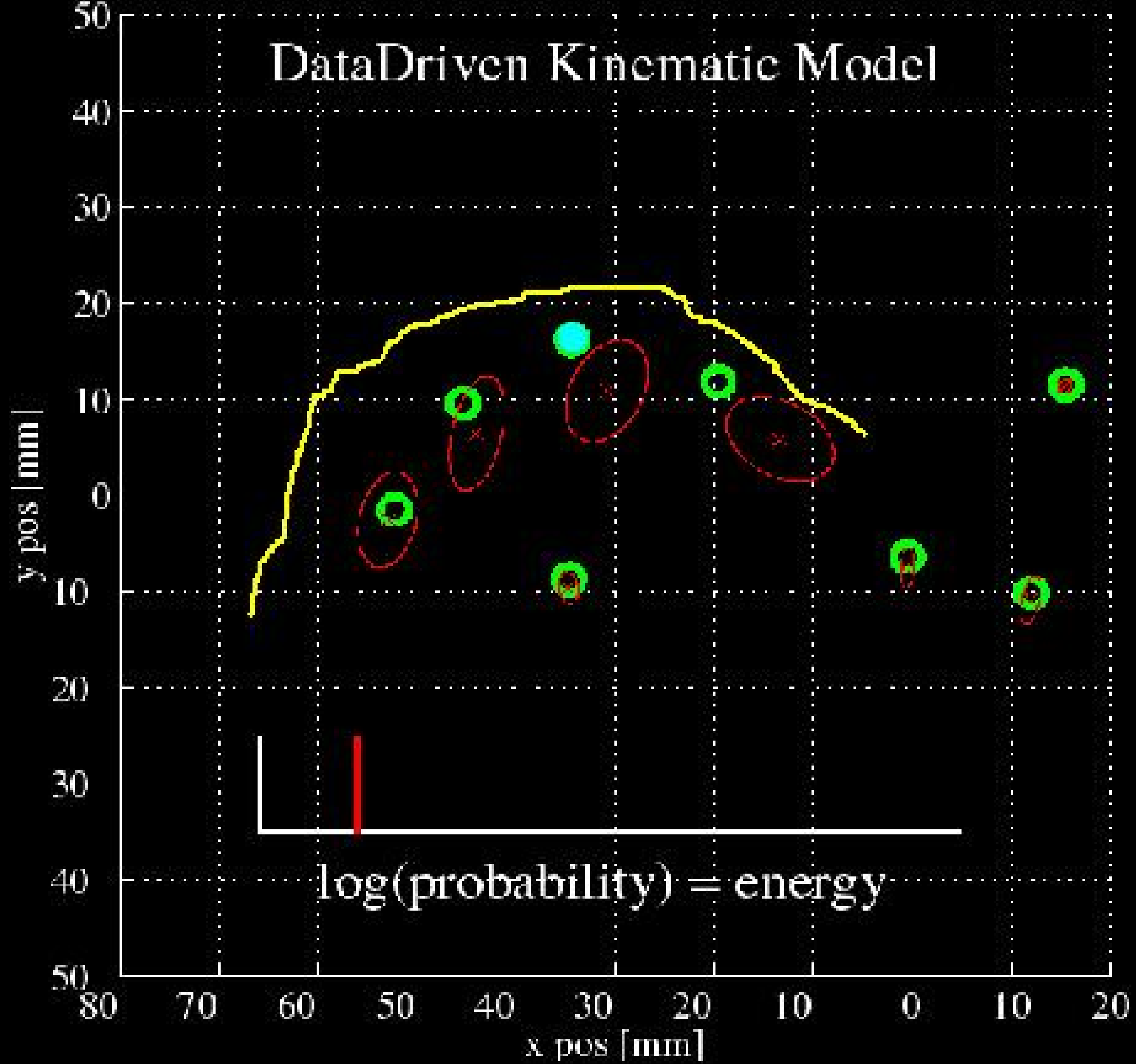
- Mixtures of PCA/SPCA/factor analysis models
- Learn either with EM or with VQ approximation





A pseudo-mechanical model

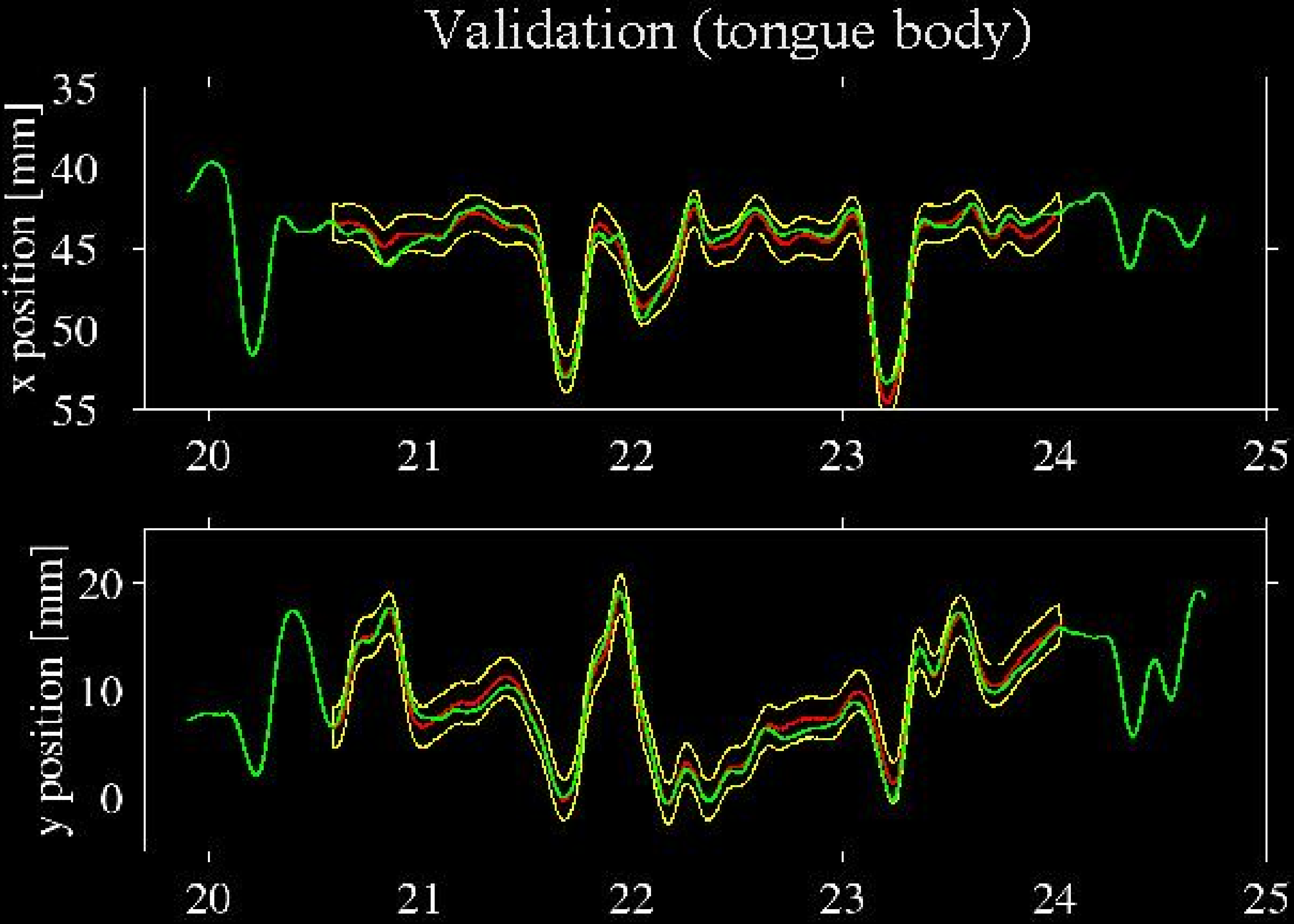
- Given the positions of one or more beads, find the most probable positions of the others (Bayes)
- Force can be defined with derivatives
- Stored energy is log probability



Data de-noising

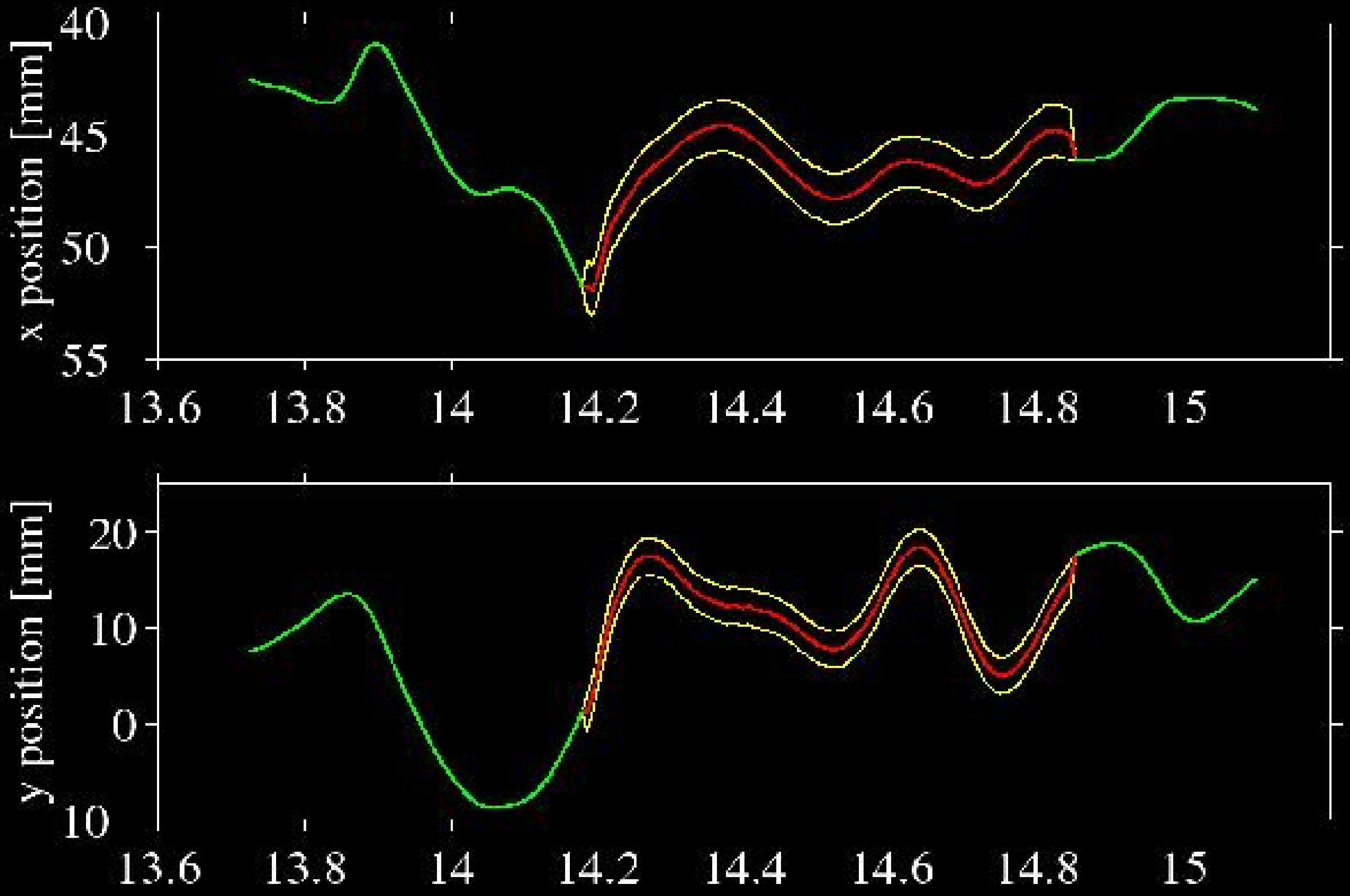
- Several sources of error in original database:
  - numerical mis-encodings (hi-byte/lo-byte)
  - swapped beads (tracking errors)
  - missing data (tracking failure)
- These can be detected & corrected automatically by looking for very low probability configurations.
- Model estimation is not affected because errors are relatively rare (a few percent of frames).

Filling in known movements



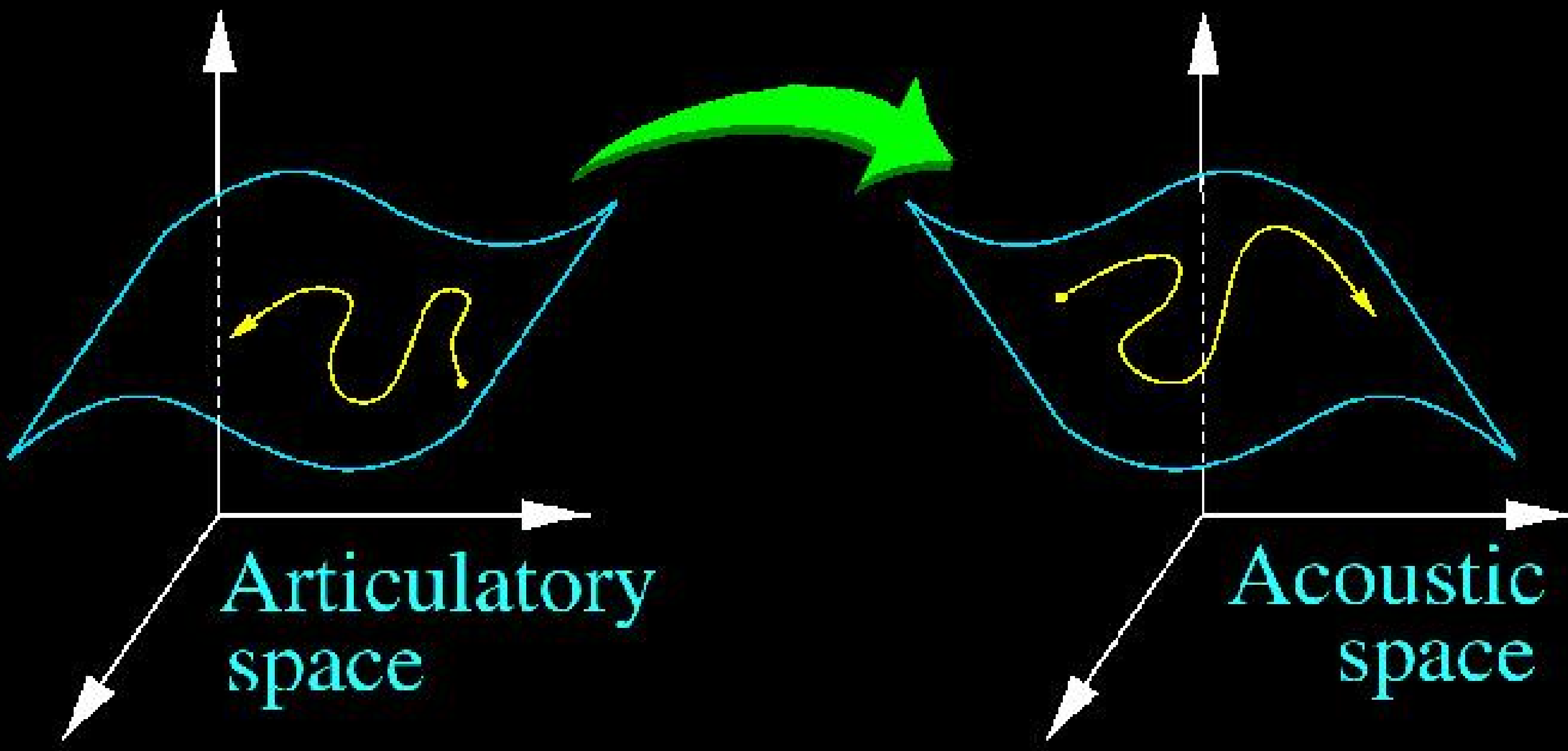
Filling in missing movements

Application to unknown data (tongue body)



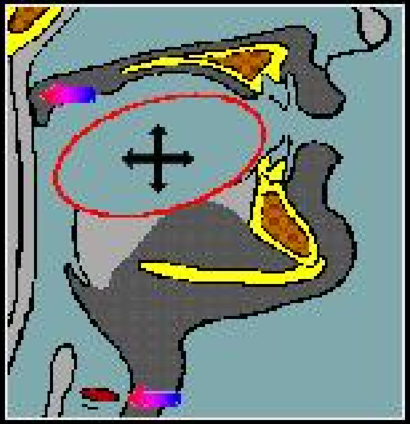
From movements to acoustics

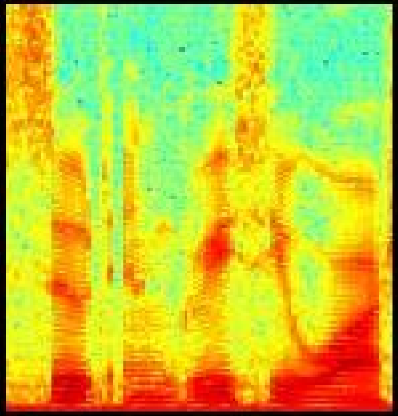
- Is there a forward mapping from articulatory movements to acoustics?



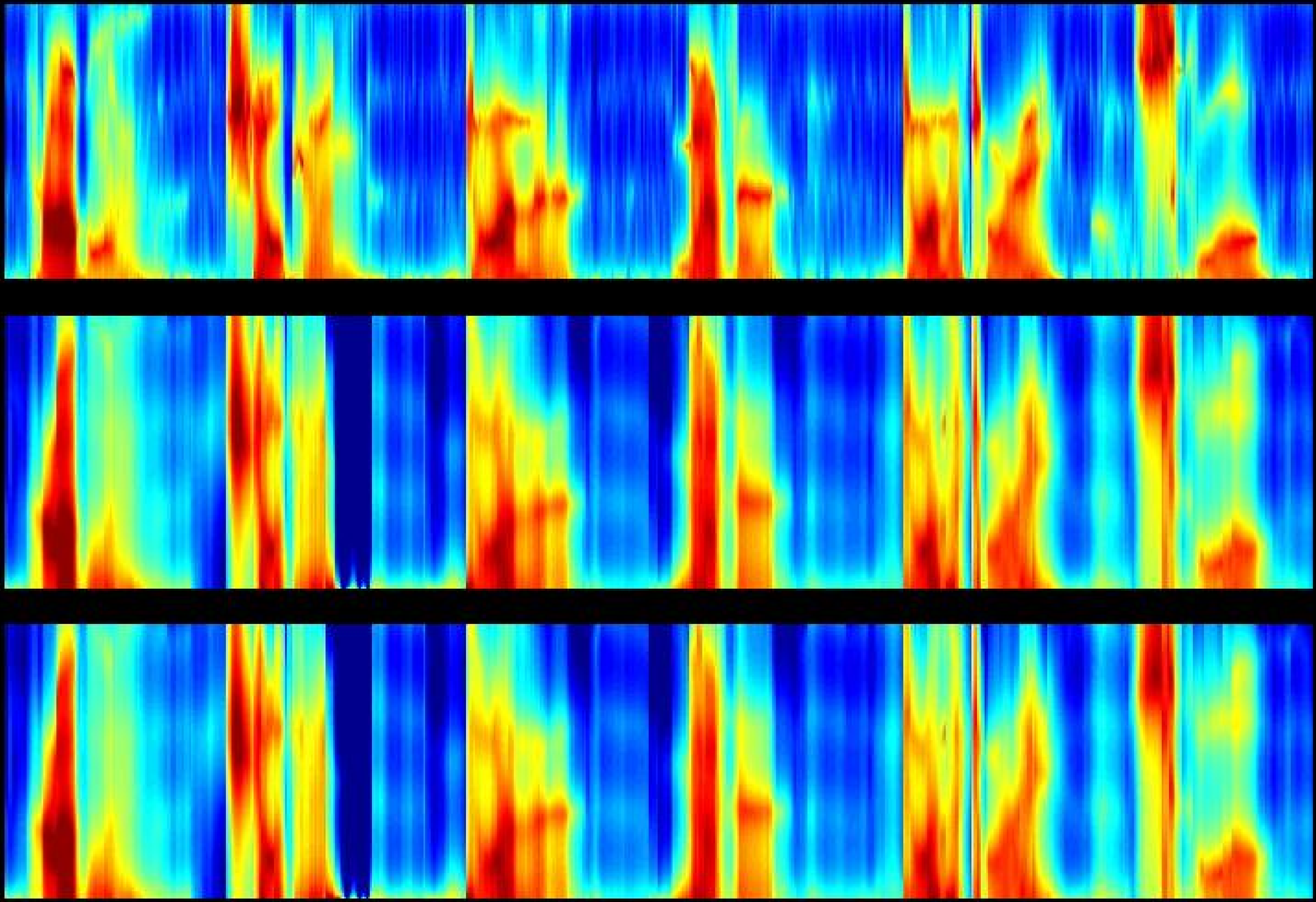
Spectrogram generation

- Train global linear models or mixtures of local linear models to predict spectral shape from midsaggital articulator positions.
- Compare estimated spectrograms with originals.

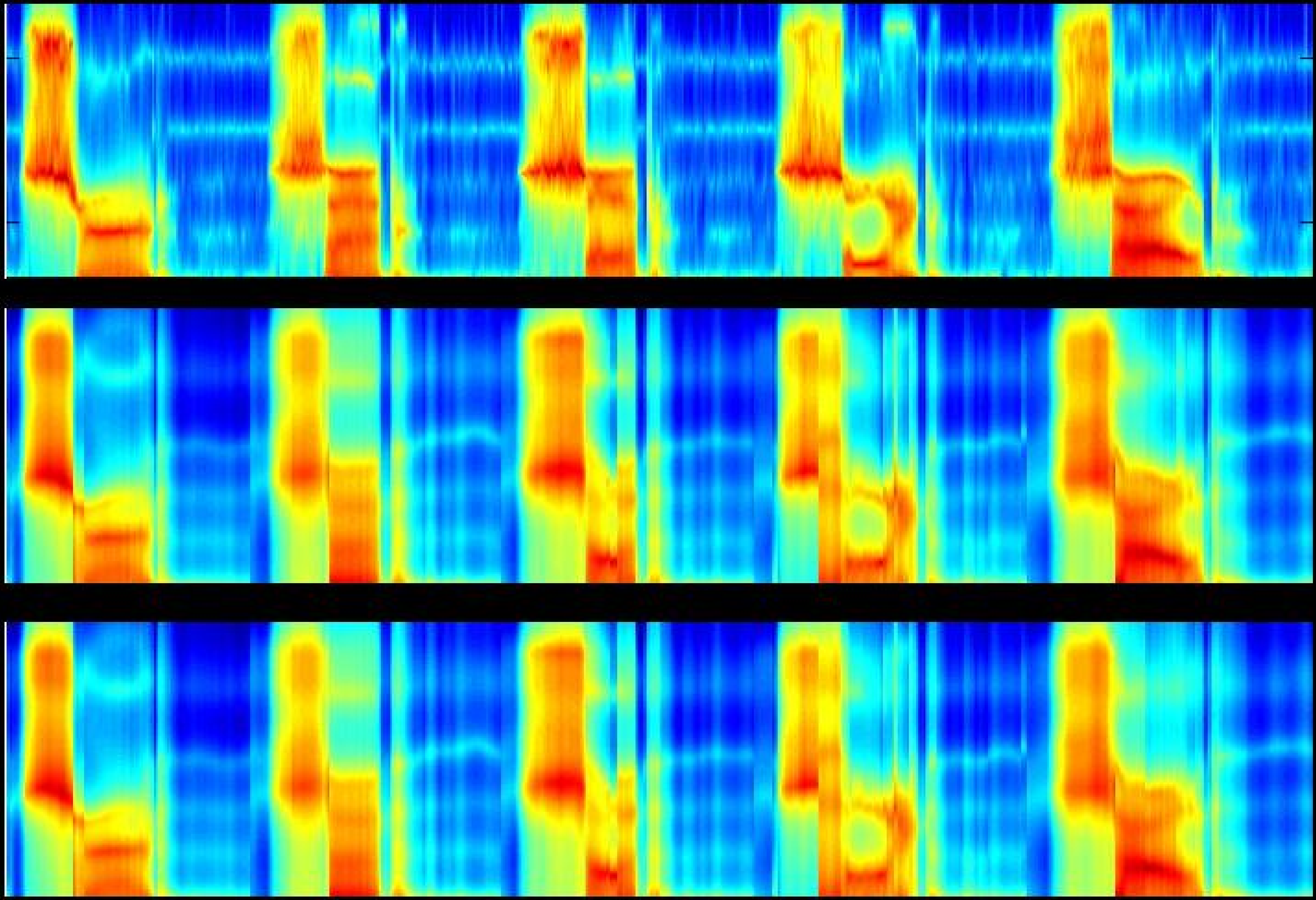




- Start with an instantaneous forward mapping.
- Adding velocities and accelerations to positions gives little improvement.



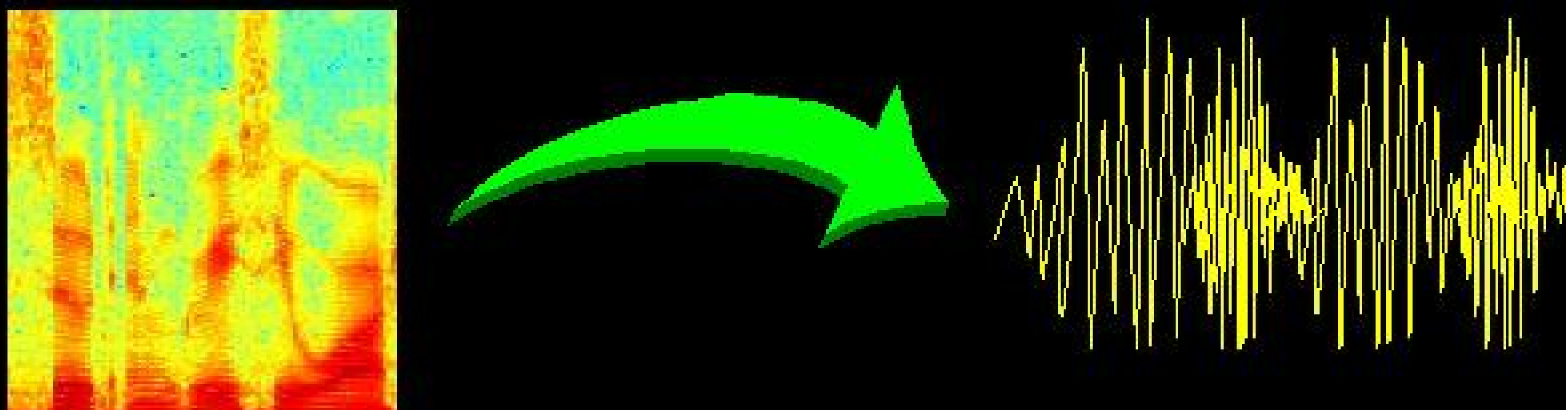
Original (top);Estimated with/without derivatives (middle/bottom)



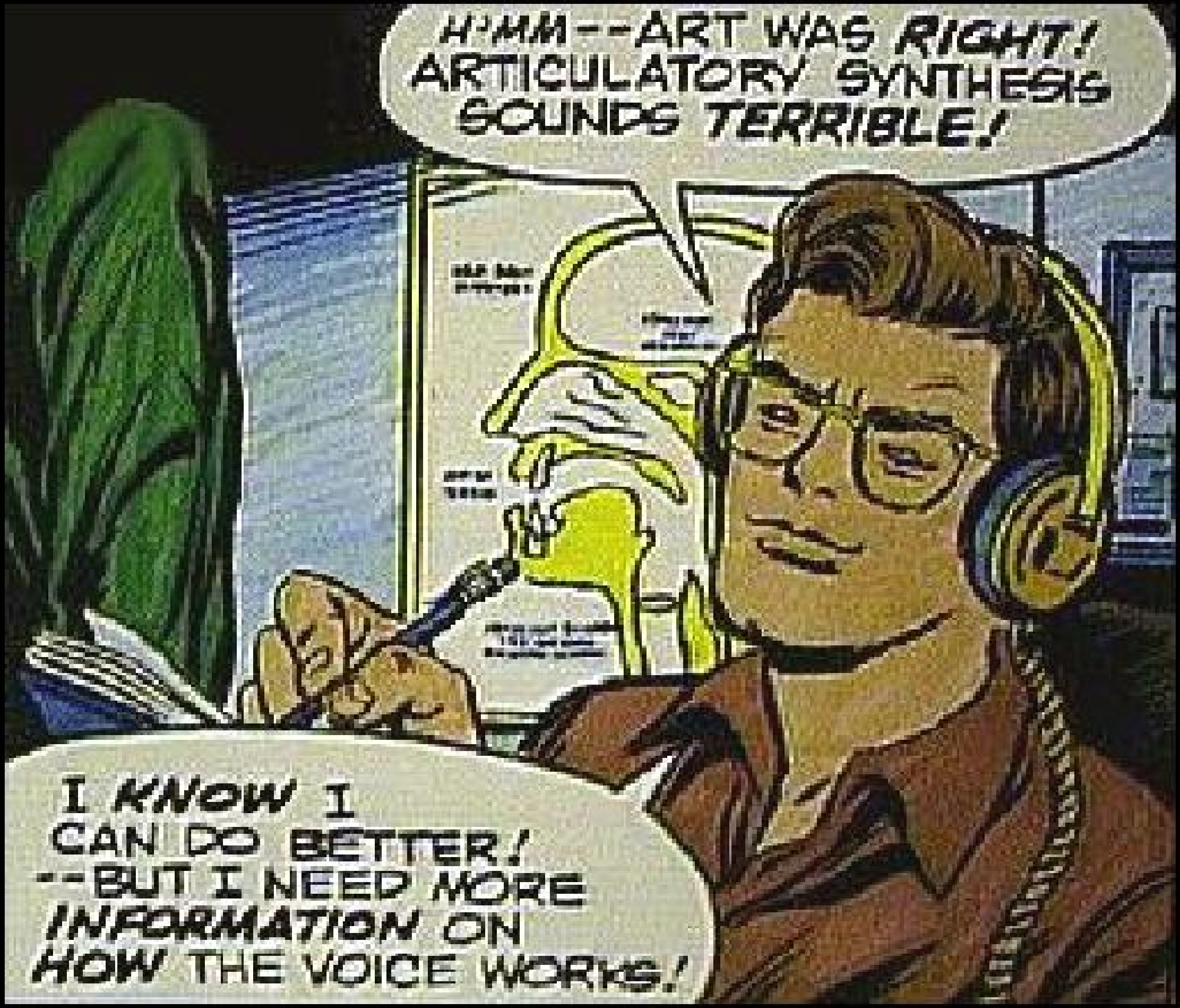
Original (top);Estimated with/without derivatives (middle/bottom)

Resynthesis from spectrograms

- Audio examples can be constructed by inverting original and estimated spectrograms.



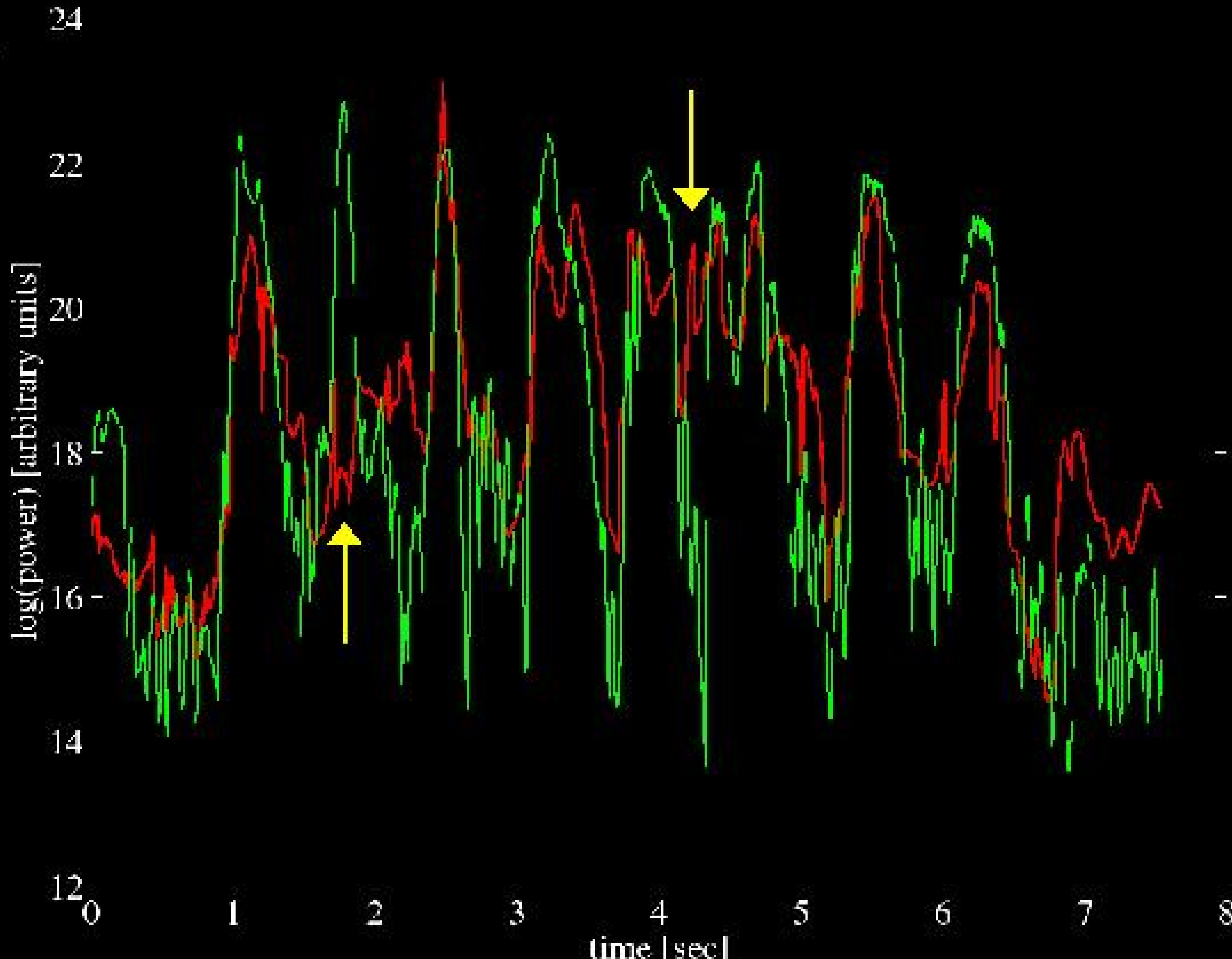
- However: spectrogram inversion is difficult and so resulting audio is noisy even when original (true) spectrogram is used.



[from Ingo Titze, Voices of People and Machines (1993)]

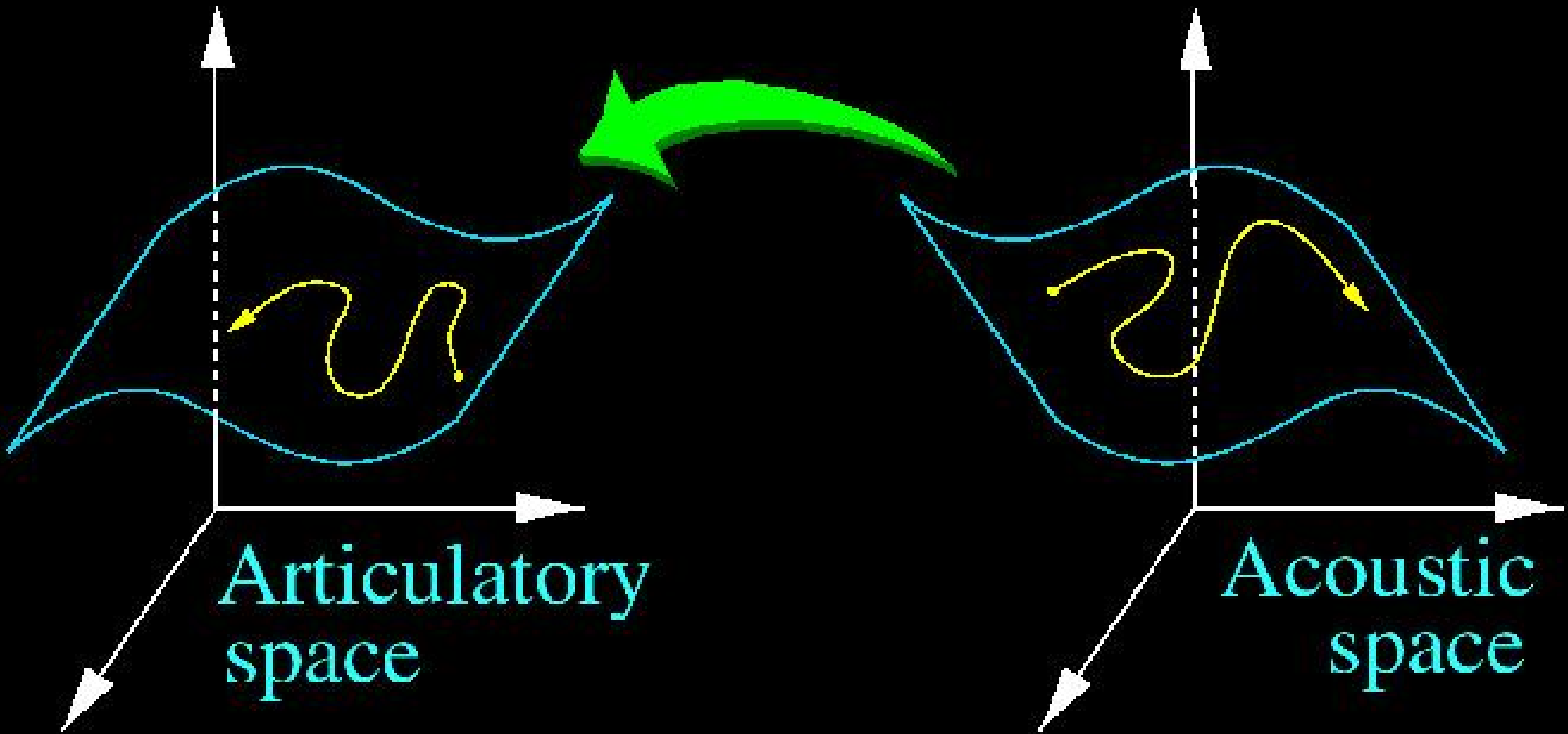
Signal energy estimation

- Generating the power signal is harder (but not impossible)
- Mostly reliable but a few serious failures

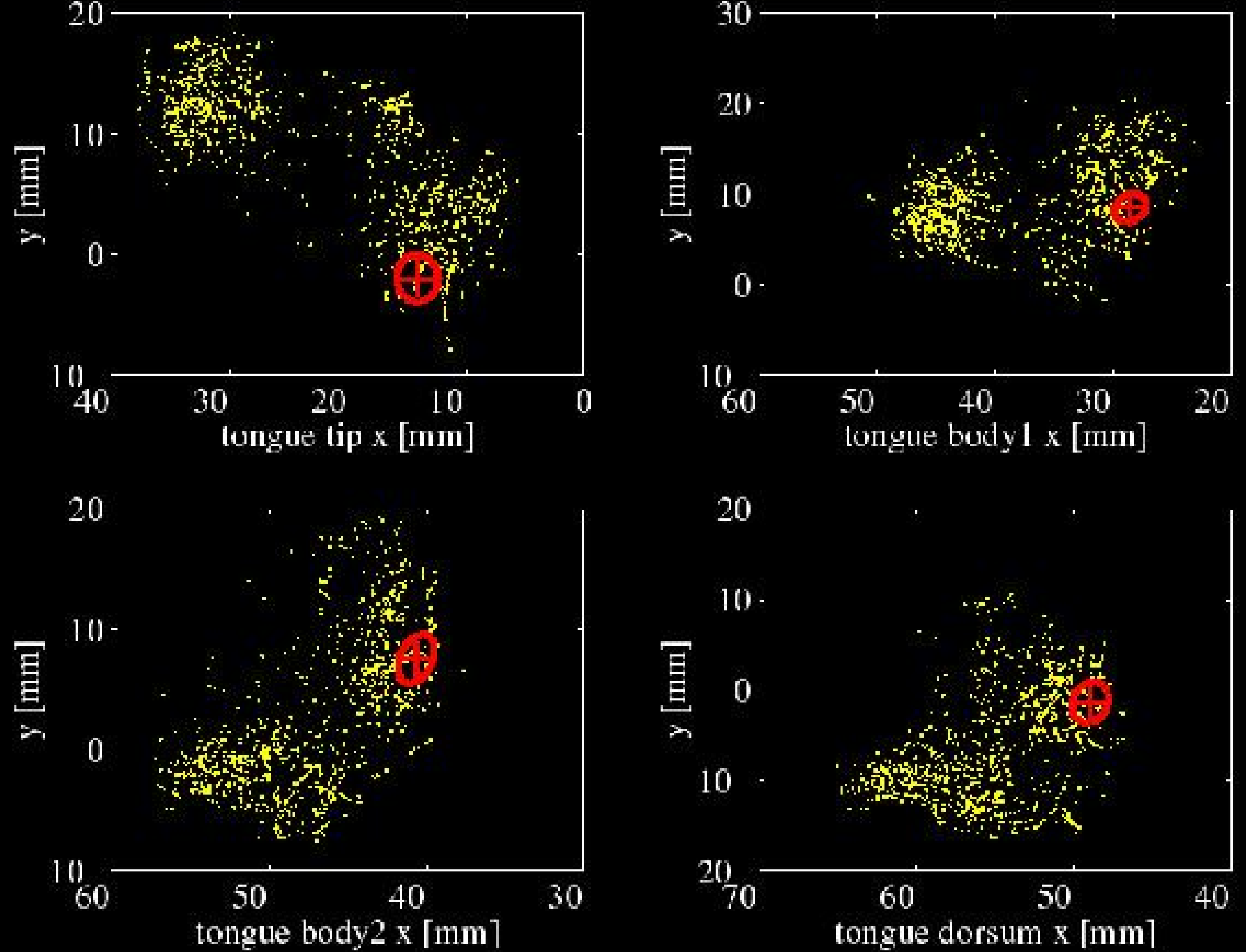


Can you hear the shape of the mouth?

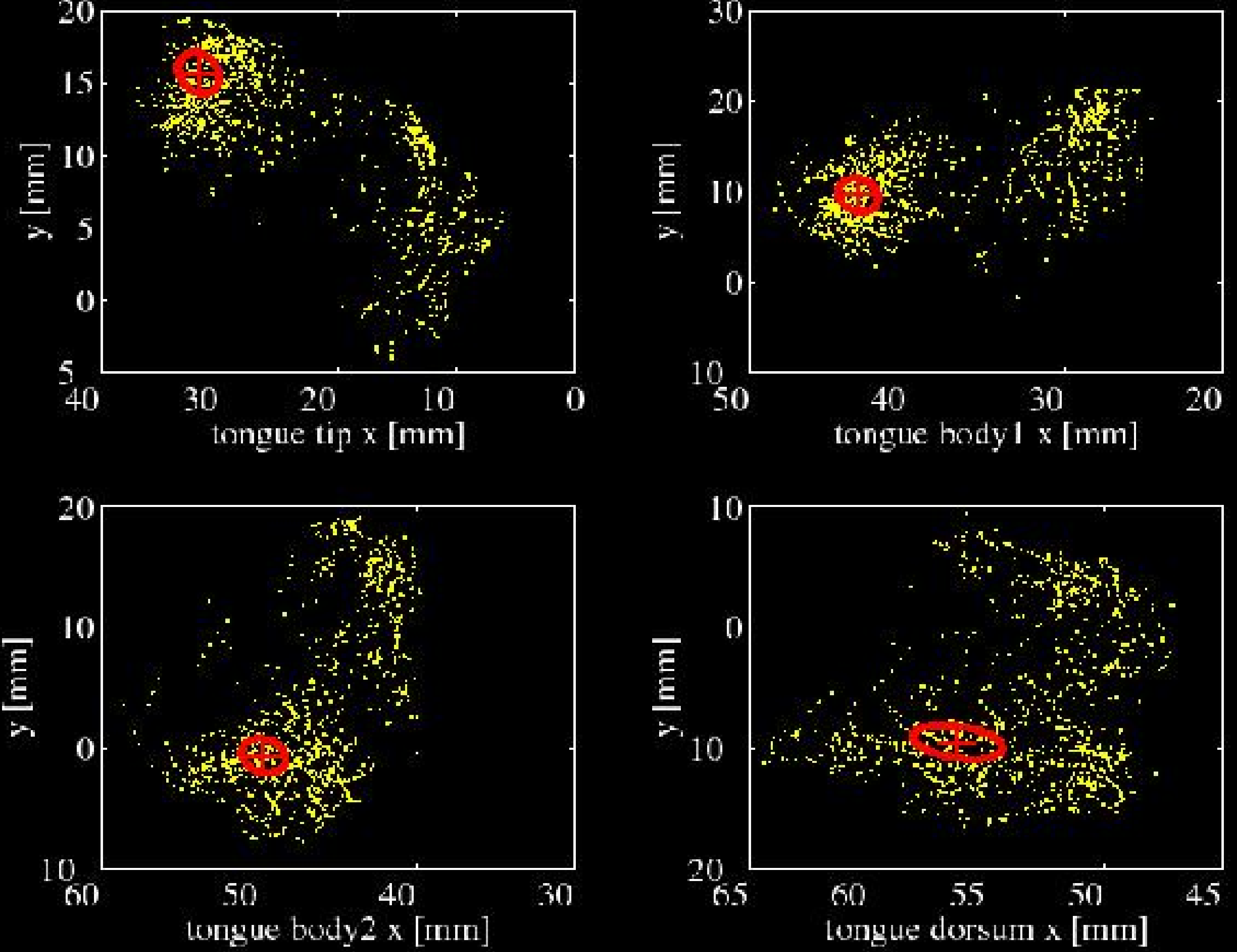
- Is there an inverse mapping from acoustics back to articulatory movements?



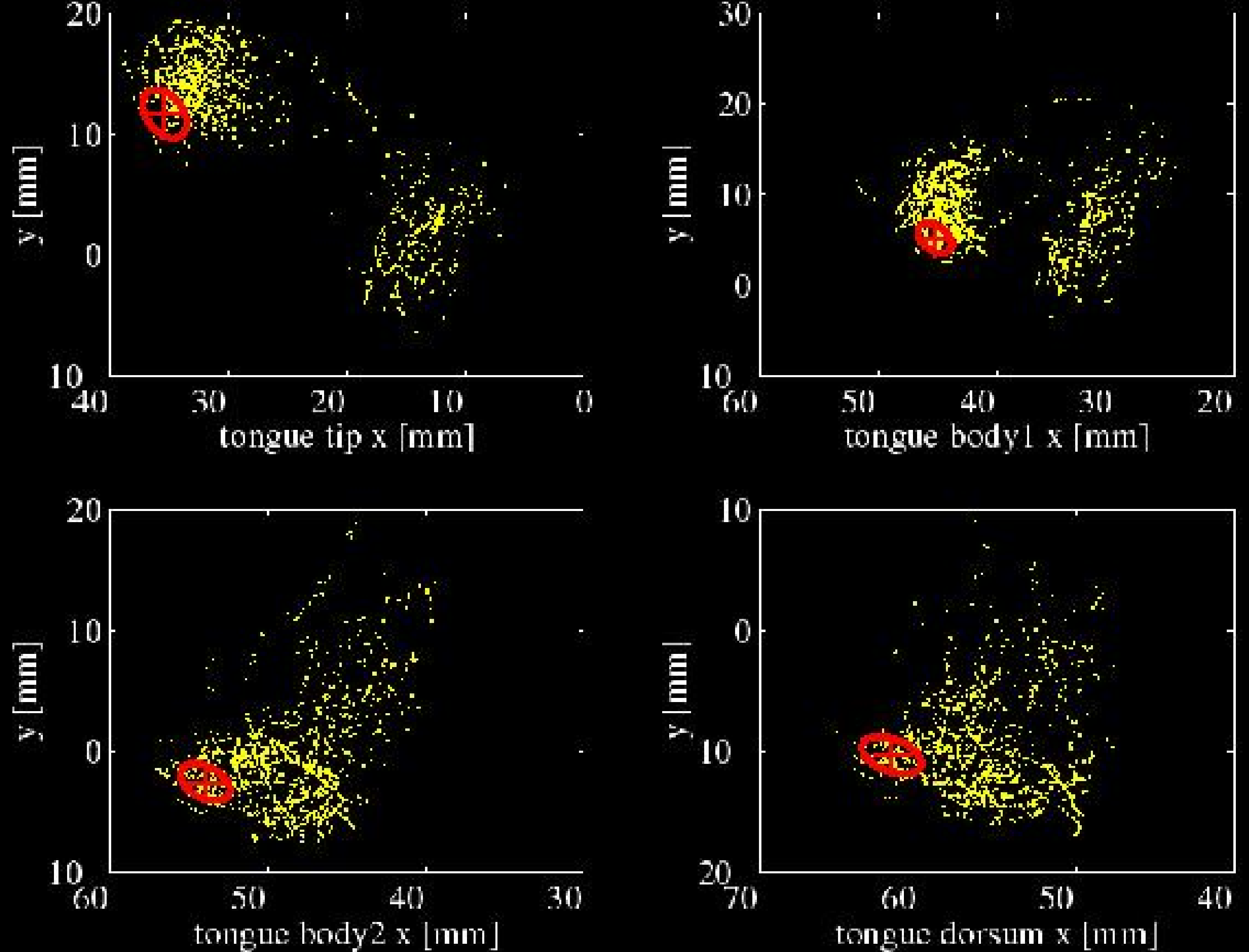
Instantaneous mapping is ill-posed



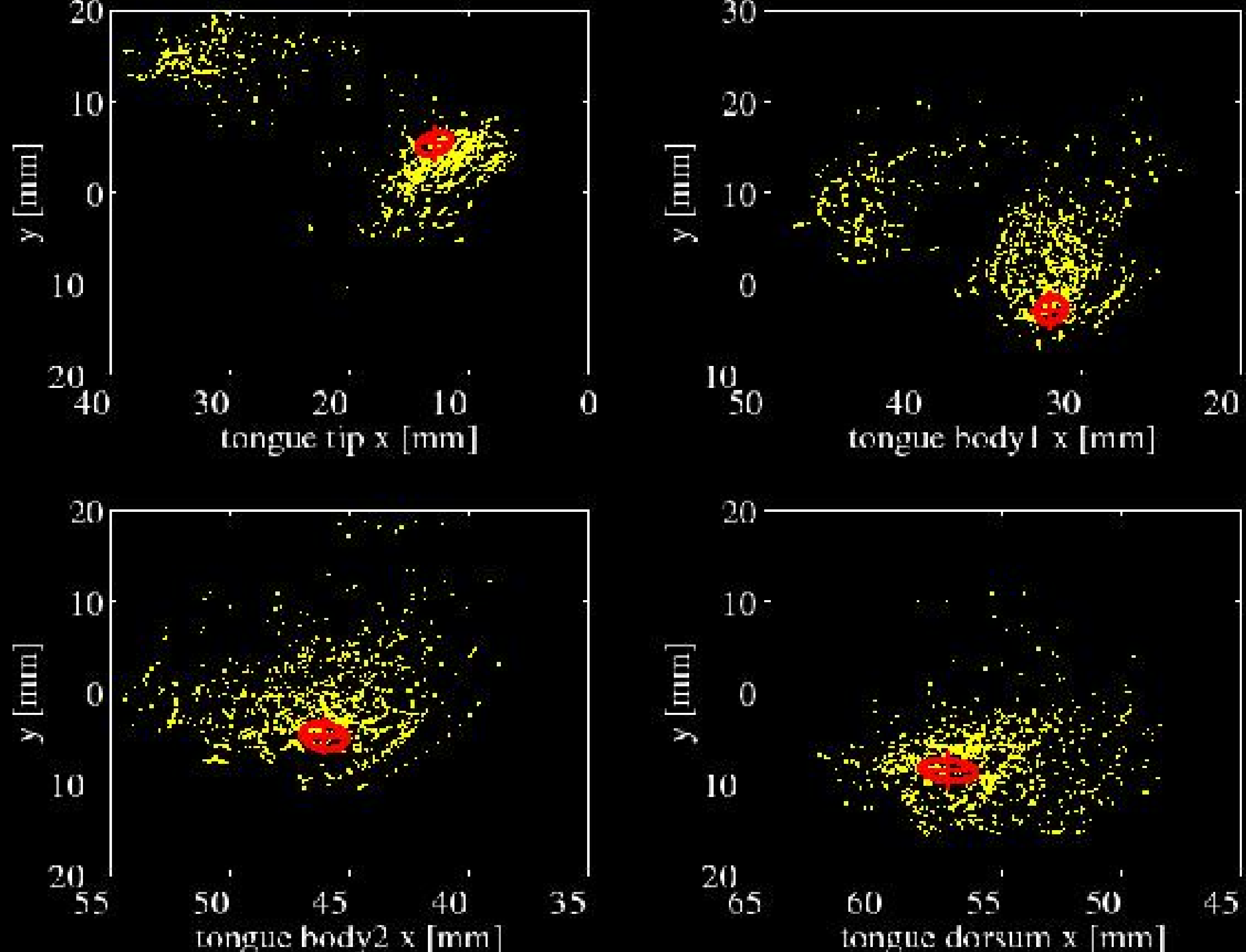
Instantaneous mapping is ill-posed



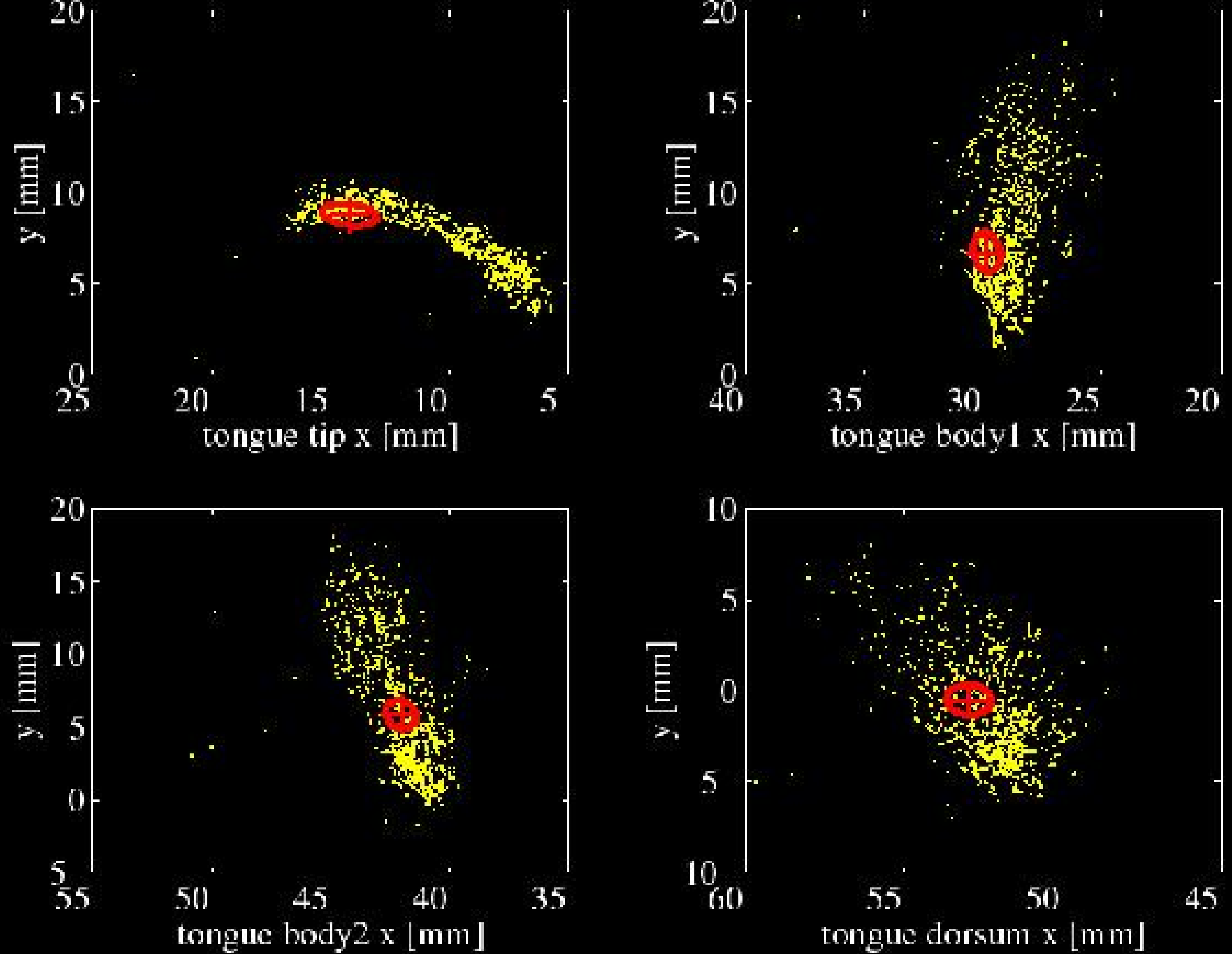
Instantaneous mapping is ill-posed



Instantaneous mapping is ill-posed



Instantaneous mapping is ill-posed



Impossible cases

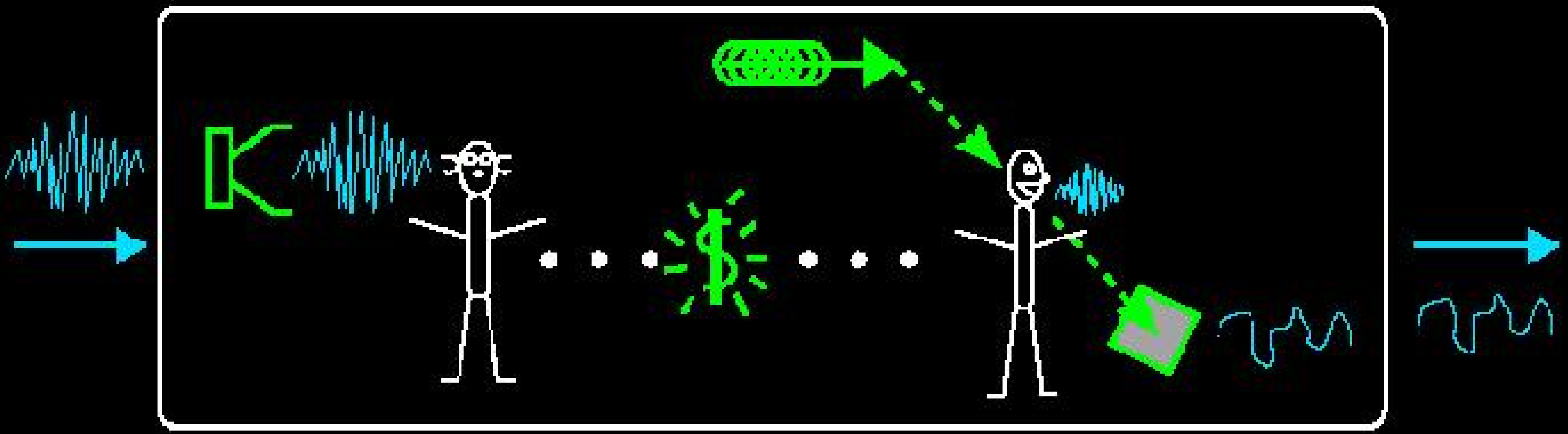
- Cannot recover movements for
  - Steady state sounds (as we have seen)
  - Silences
  - Ventrilloquism effects

Possible case

- But…
  for a single speaker and normal spoken language, the problem seems in principle to be solvable.

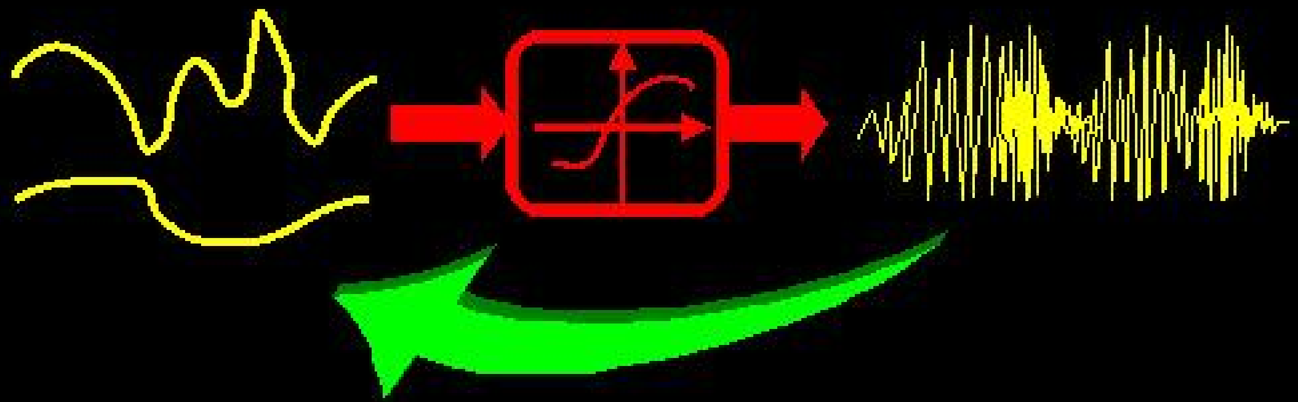
Million dollar solution

1) Original speaker listens to audio;
2) Gets paid $1,000,000;
3) Repeats what they said exactly;
4) Movements are recorded during repetition.



State estimation from entire trajectories

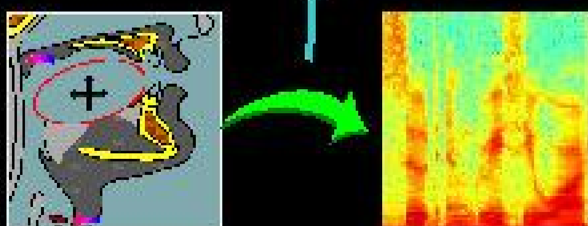
- Generative models with underlying states or control signals can sometimes be inverted to recover underlying states from noisy observations.



- This is called state inference.
  e.g.
  Kalman filter for
  linear dynamical systems,
  Viterbi decoder for HMMs.
- Not instantaneous: uses the entire time sequence of observations.

Constructing a simple LDS

- An instantaneous forward mapping is the key part of a **linear dynamical system** (LDS). We already have this in hand from before!



$$y(t) = Cx(t) + \text{noise}$$
$$x(t+1) = Ax(t) + \text{noise}$$

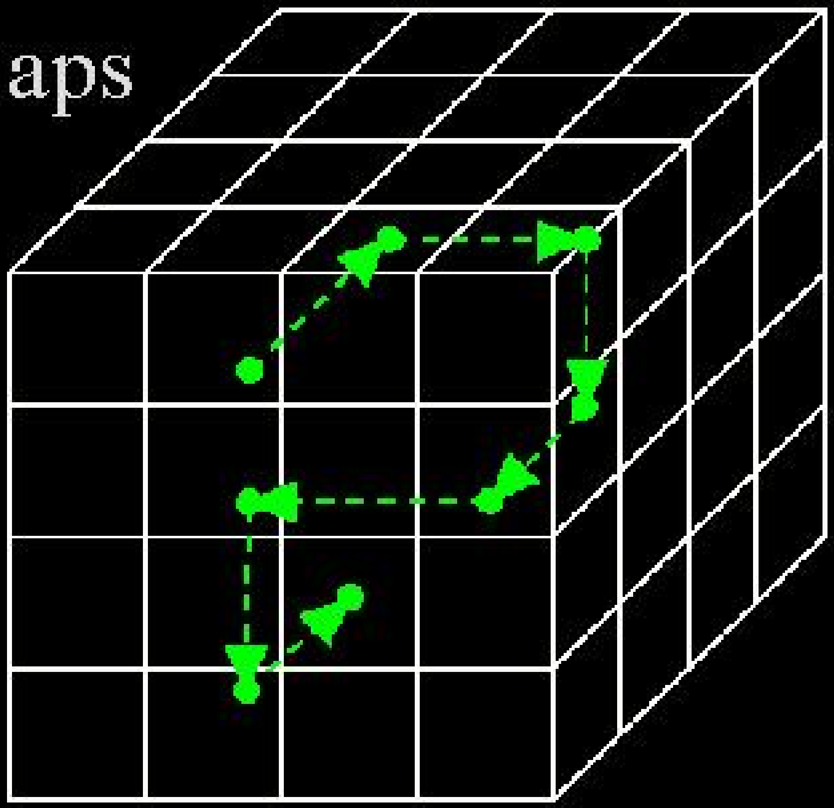
- We just need a simple dynamical model for the states. Setting **A=I** gives a random walk. Could also use momentum.

Direct Kalman smoothing?

- Construct a **LDS** using global forward model then use Kalman Smoothing
- A completely supervised approach: learning is easy.
- But: the global model is not powerful enough
  - global forward synthesis is poor
  - Kalman smoothing on acoustics using **LDS** from global model gives poor recovery of movements
- Inverting the mixture of local models would be better, but is hard to do (...wait though...)

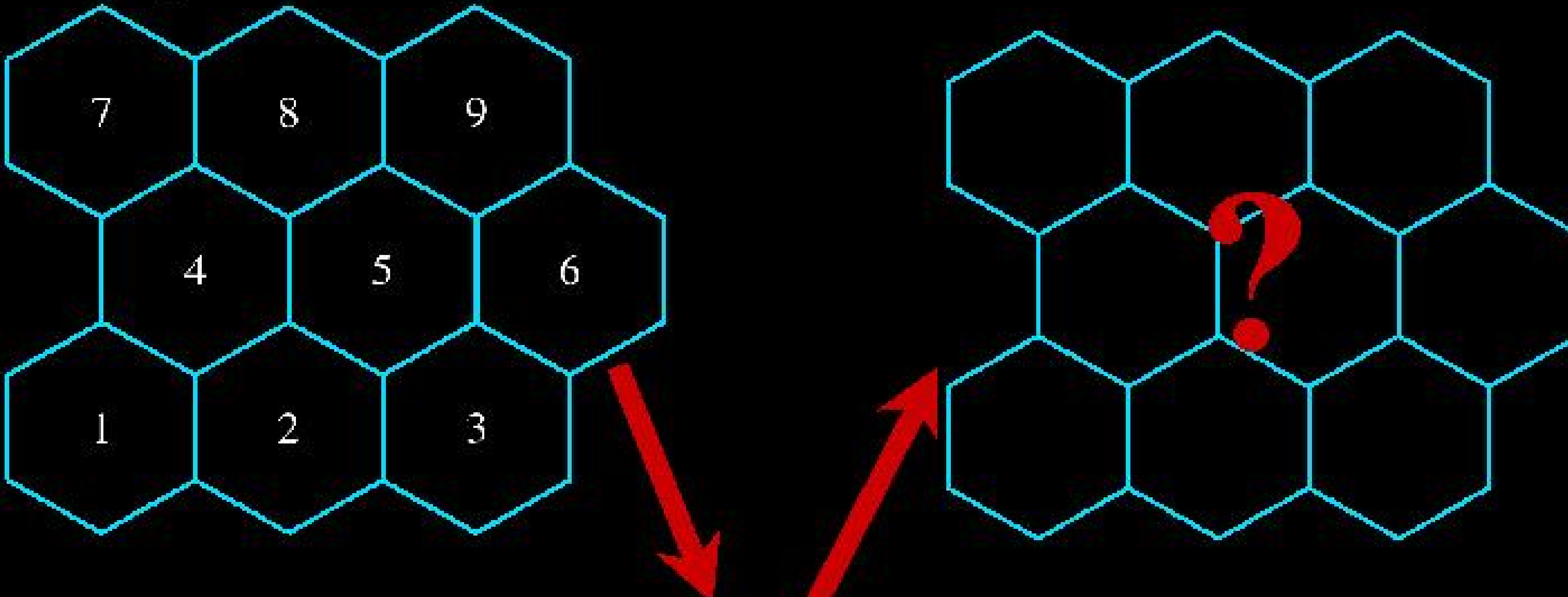
Self-organizing Markov models

- A much more powerful model is the *self-organizing hidden Markov model*
- Learn low dimensional maps to explain sequences of high dimensional data.
- Latent variables only change slowly/smoothly.
- A completely unsupervised approach



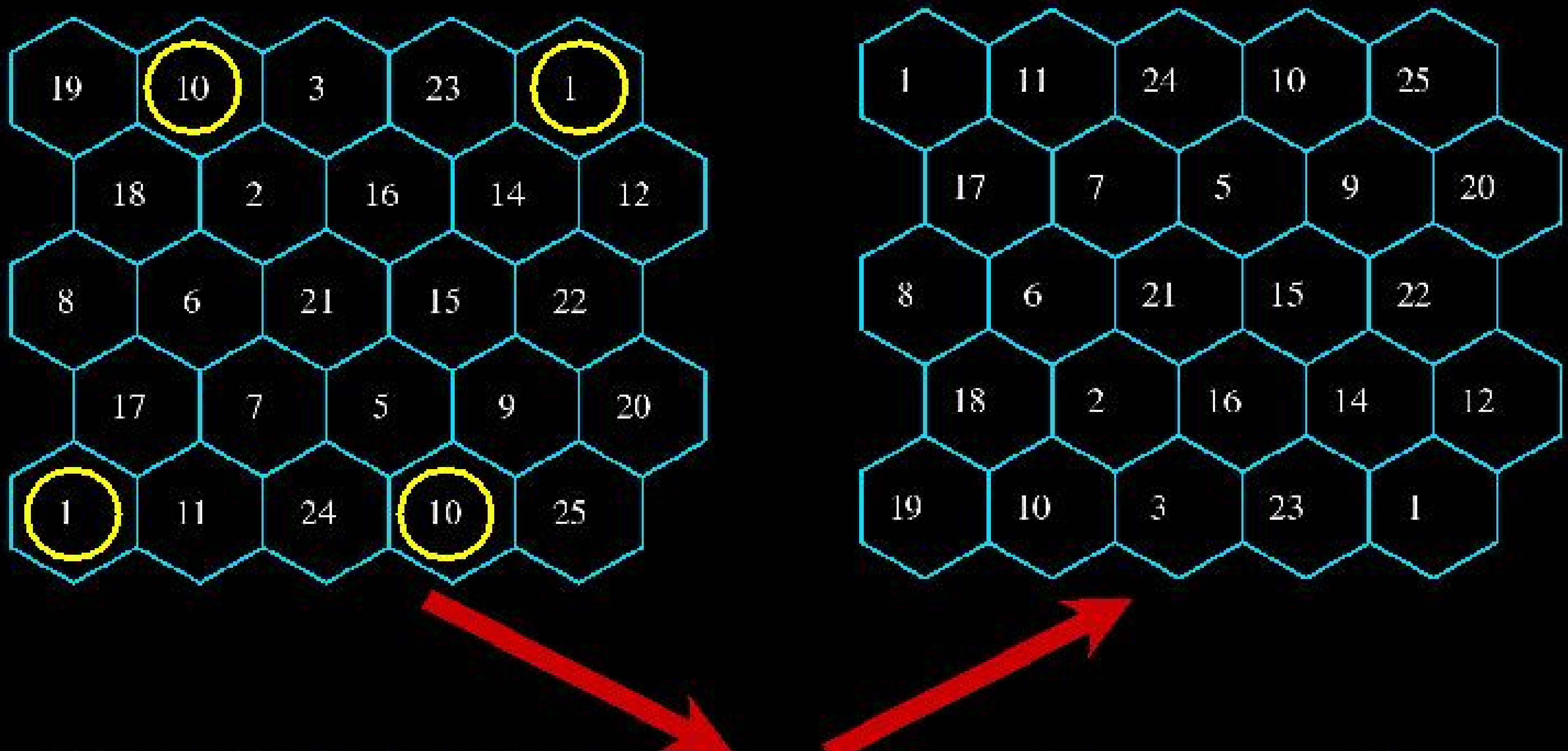
A simple game

- Original map
- Recovered map



"Smooth" Observations: 1,4,1,4,5,2,4,1,4,5,2,1,
2,3,5,4,5,6,9,8,4,1,4,5,4,8,9,6,5,8,4,2,1,2,3,5,6,9,
5,6,3,6,5,9,8,5,6,9,5,9,6,5,4,2,5,2,1,4,2,...

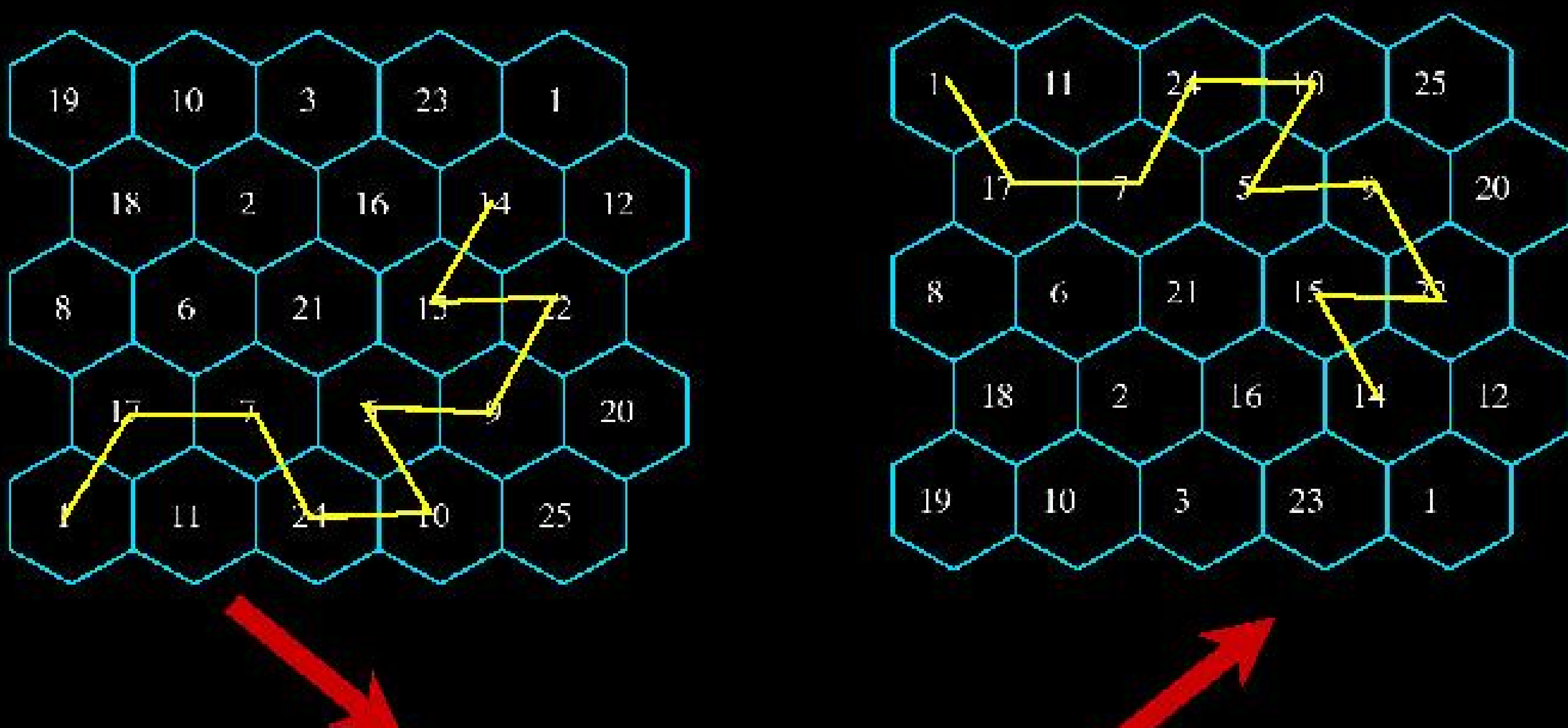
Noise and repetitions



Noisy Observations: 11,7,5,15,16,12,10,2,6,2,10, 7,5,9,9,22,15,21,7,17,8,6,1,24,10,25,10,9,20,22,9,5, 15,14,12,24,23,16,3,16,2,10,3,6,6,18,...

Traces in smooth maps

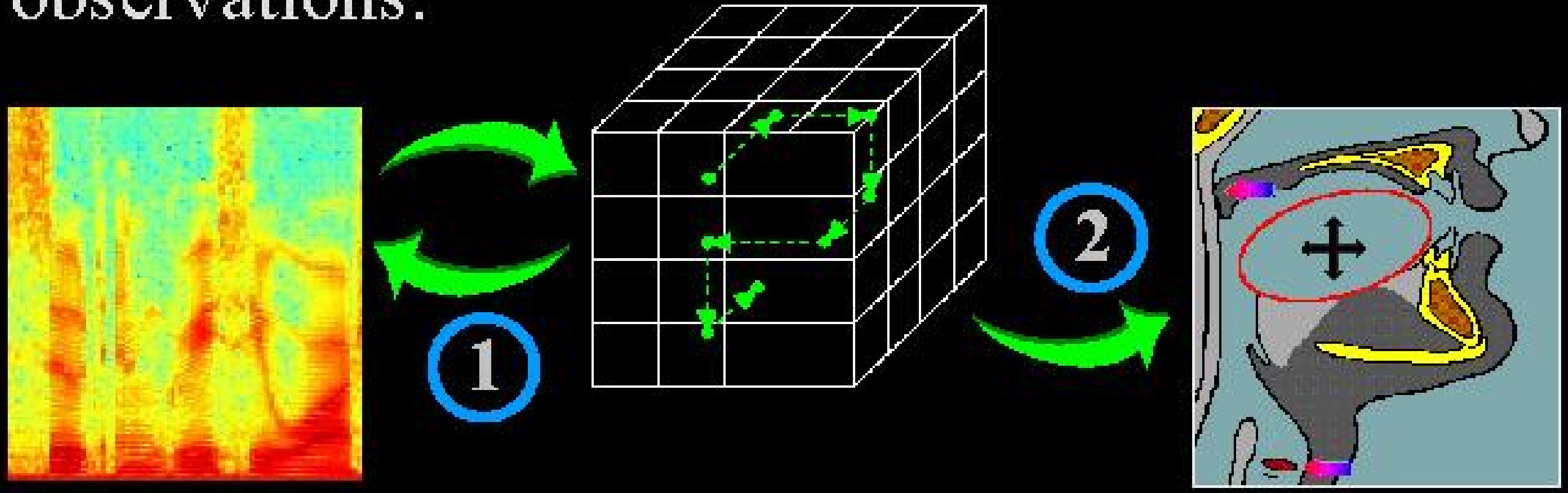
- After the map is learned, new sequences can be **decoded** to give corresponding trajectories.



1,17,7,24,10,5,9,22,15,14,...

Direct application of SOHMM?

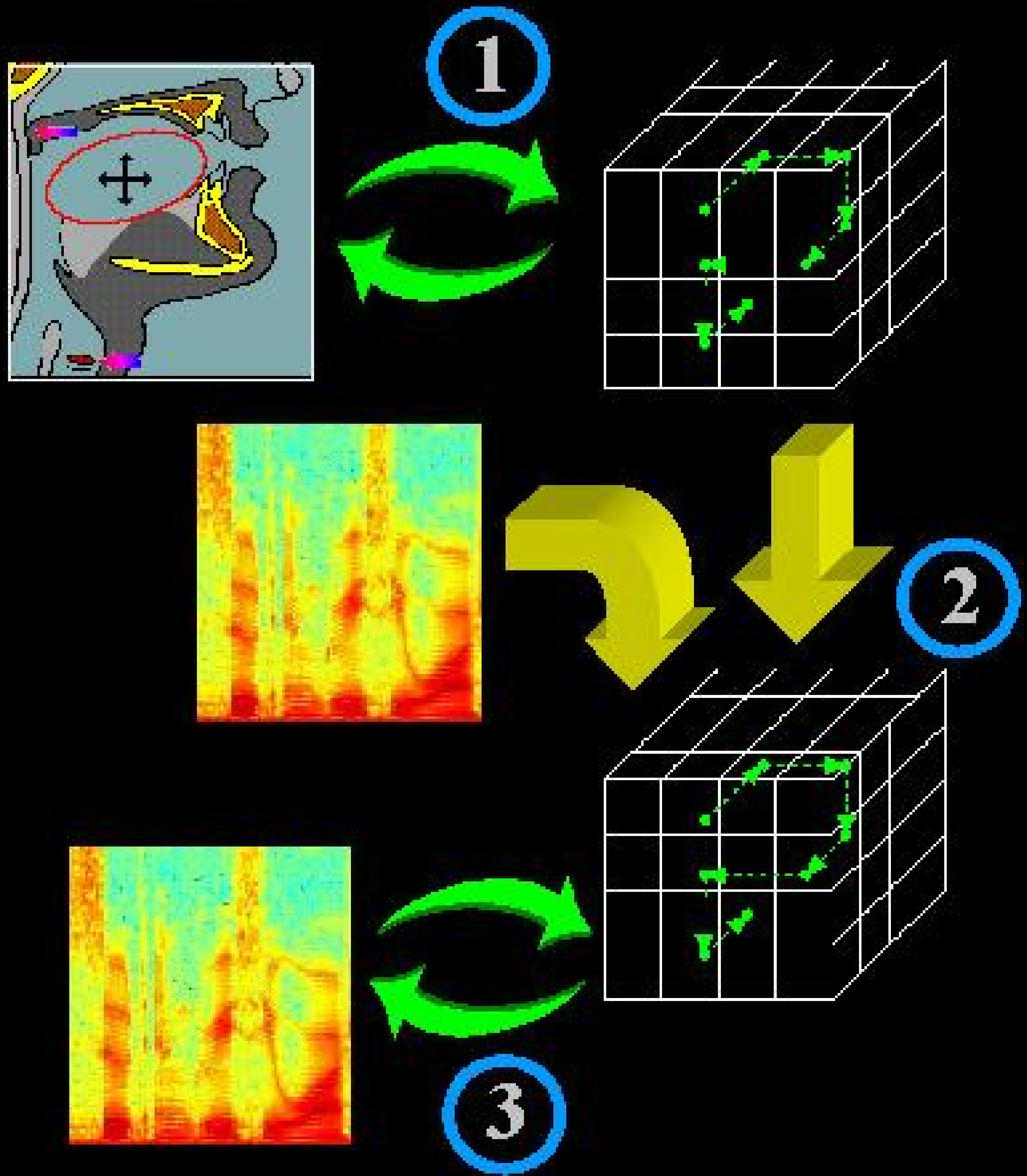
- Train a SOHMM directly on sequences of acoustic observations.



- Inference gives a trajectory in state space. Relate state space trajectories to articulatory movements.
- But: although model is powerful learning is hard.

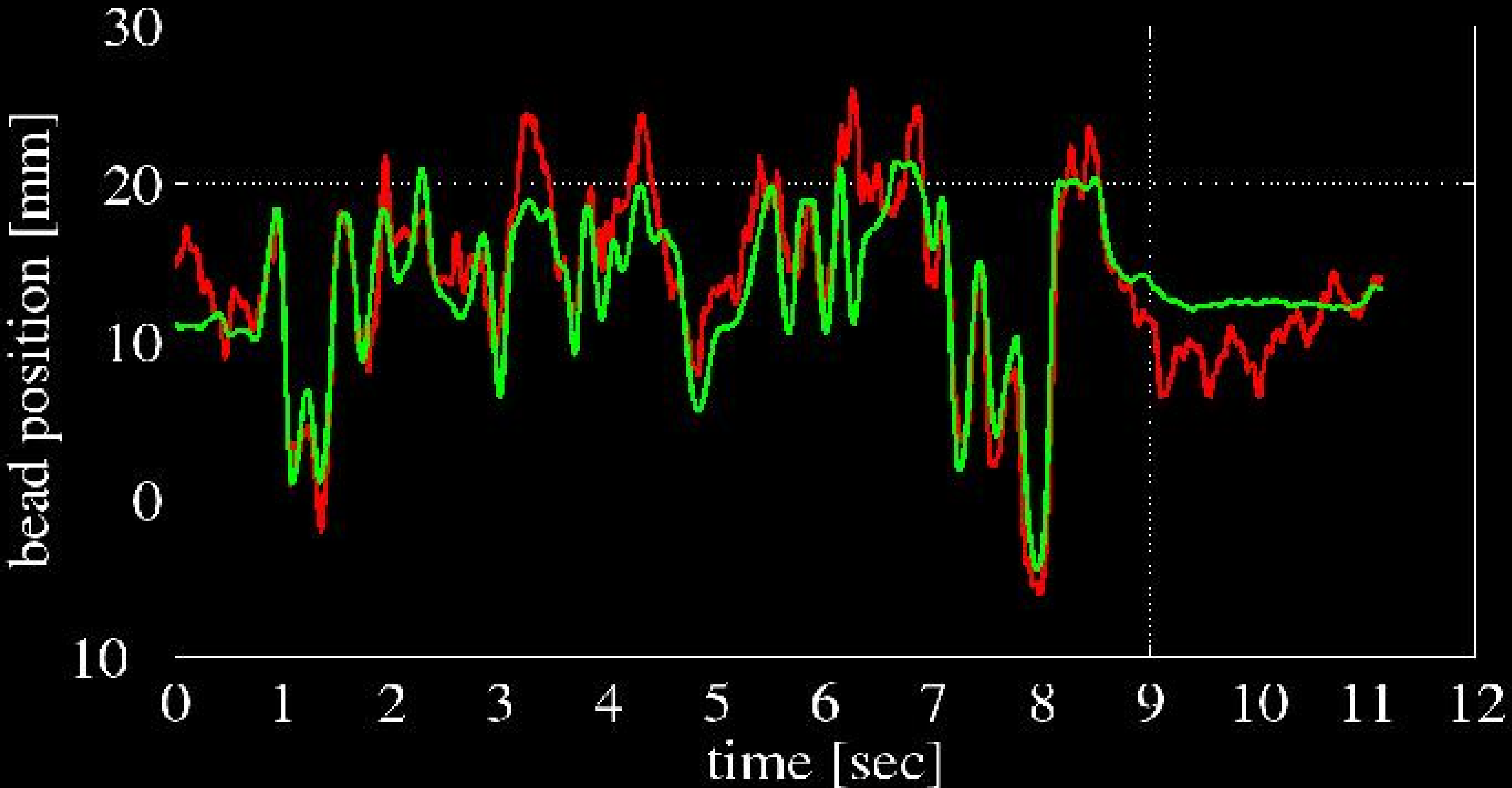
A combined approach

- Idea: train a SOHMM on the sequence of local models (supervised step). This is just geometry.
- Use this model and the measured acoustics in each local model to induce probabilities and convert the SOHMM to a new acoustic SOHMM.
- Retrain (unsupervised step).
- Use coupled inference to do decoding.



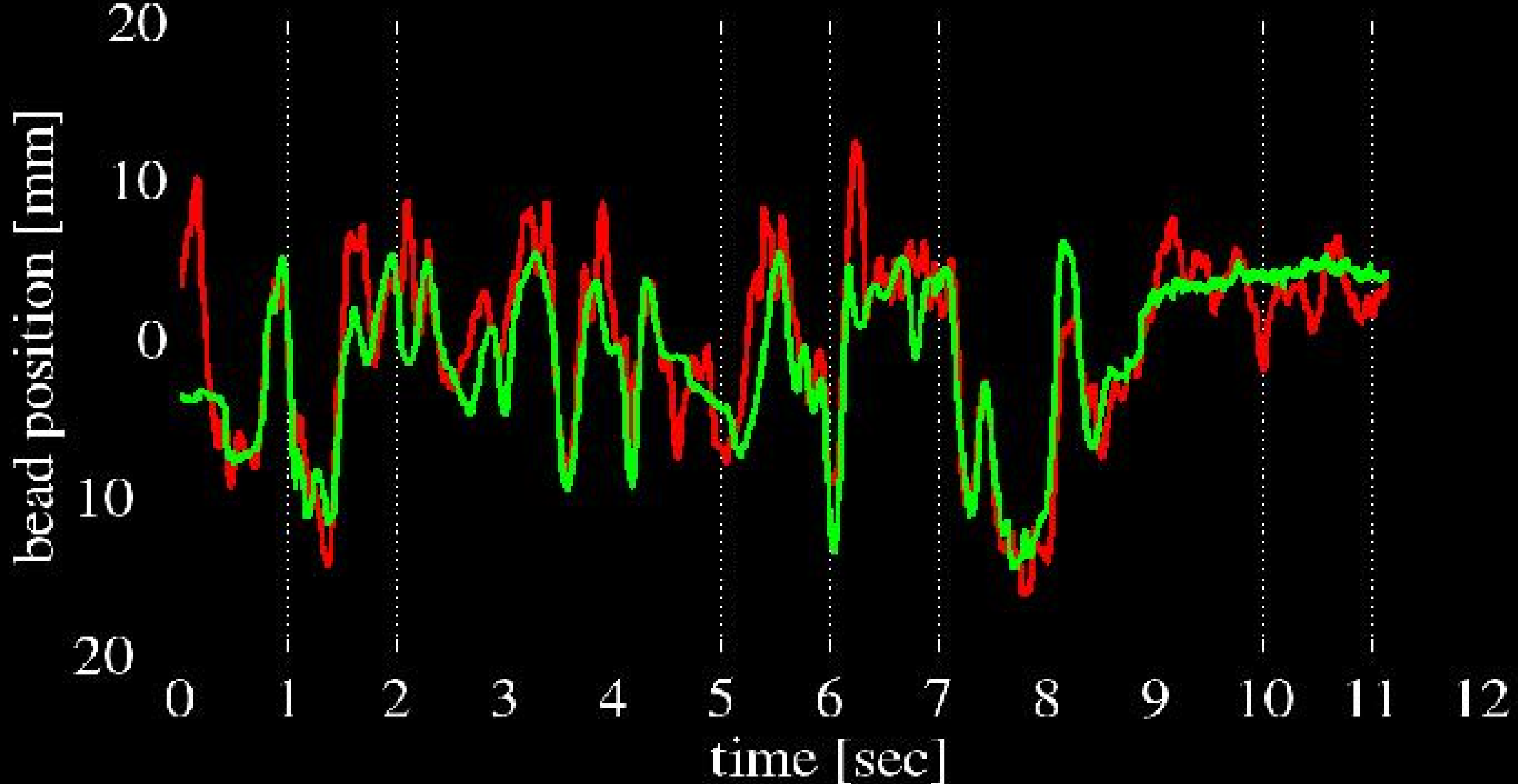
Recovery of movements from acoustics

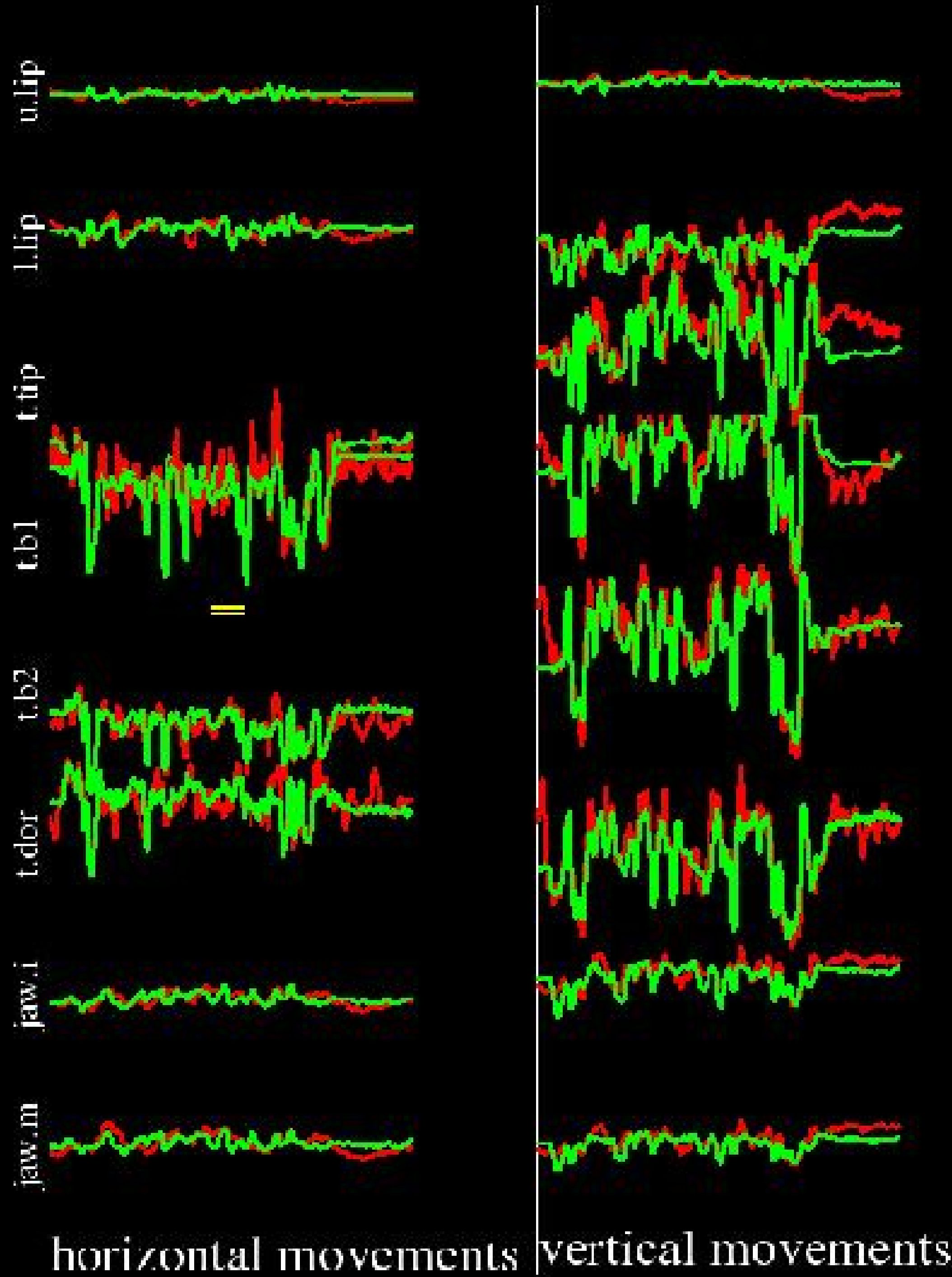
- Estimation of tongue tip vertical motion



Recovery of movements from acoustics

- Estimation of tongue dorsum vertical motion

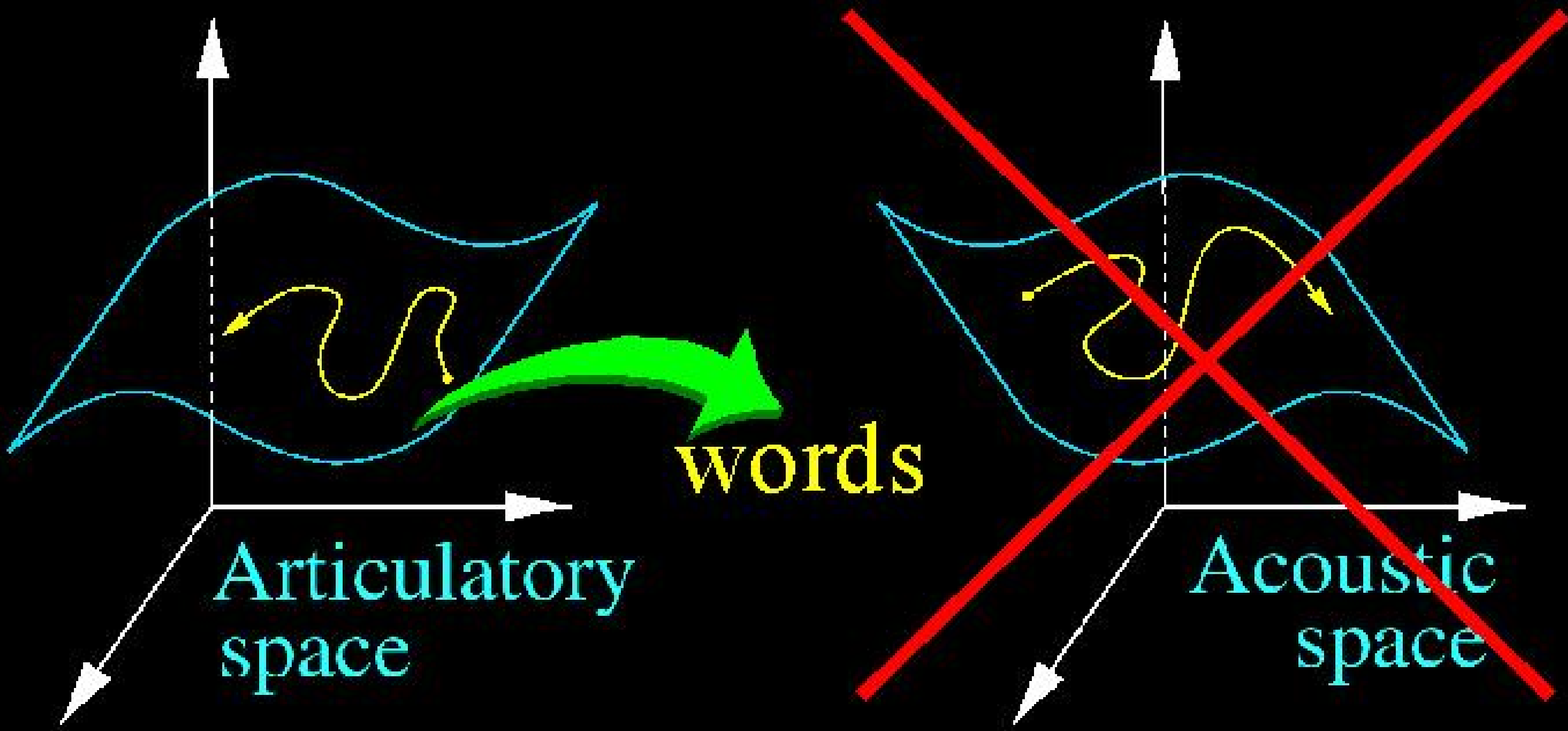




All bead traces estimated from acoustics

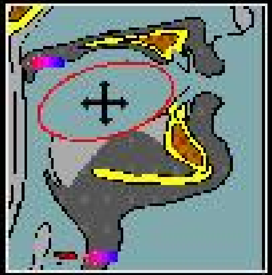
From movements to words

- Can we do simple speech recognition using the true articulatory movements?



"Cheating" experiments

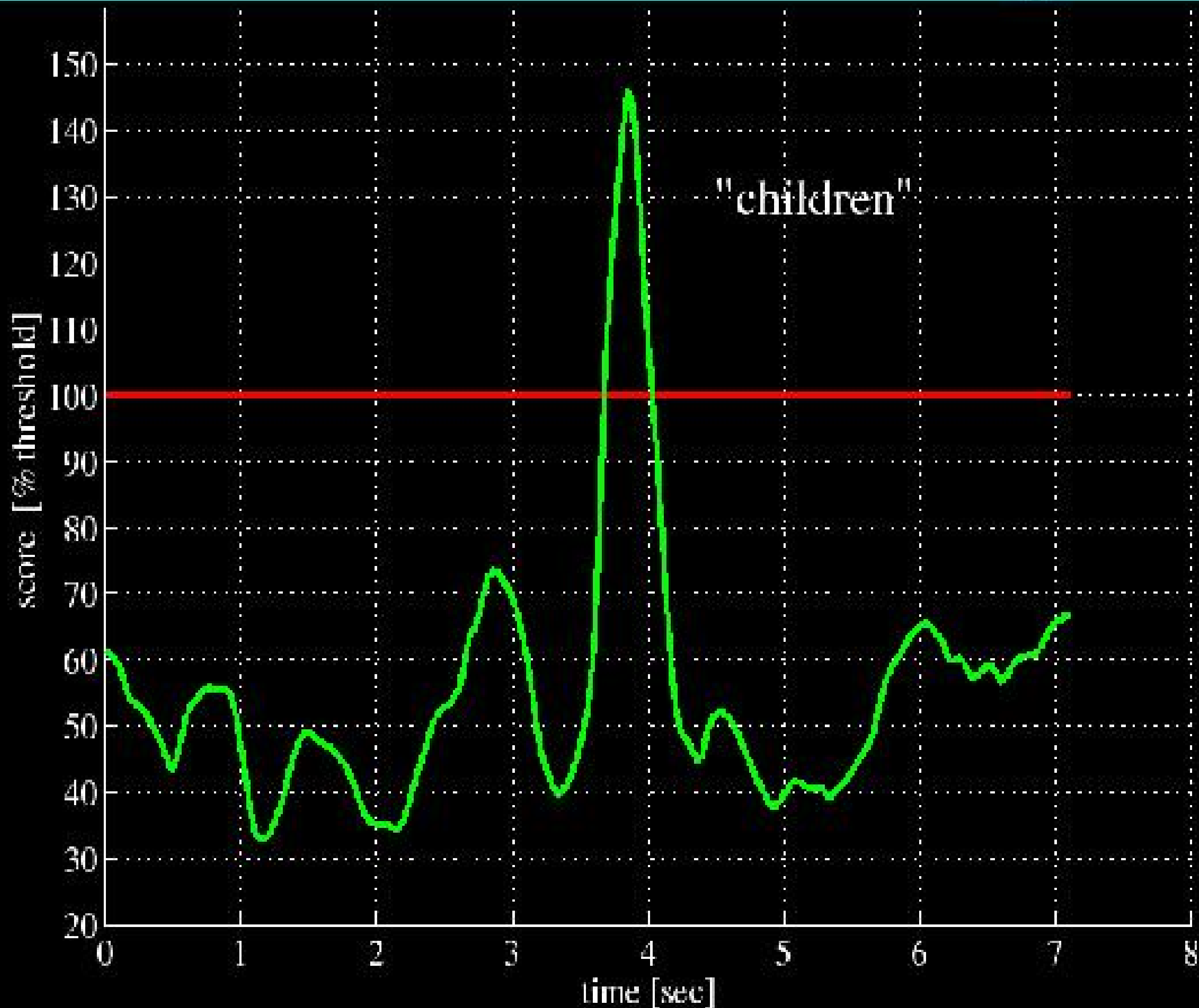
- Use one instance of a word as a template (!)
- Look for matches in entire database (`550ms` template in about `1200sec` of data)
- Simple dynamic time warping algorithm **using true articulatory movements**.



"science is fun"
s-ay-ax-n-s ih-zs f-ax-n

- Result? **Perfect performance on an easy task**. A simple threshold on matching score finds all instances of the word with no false positives.

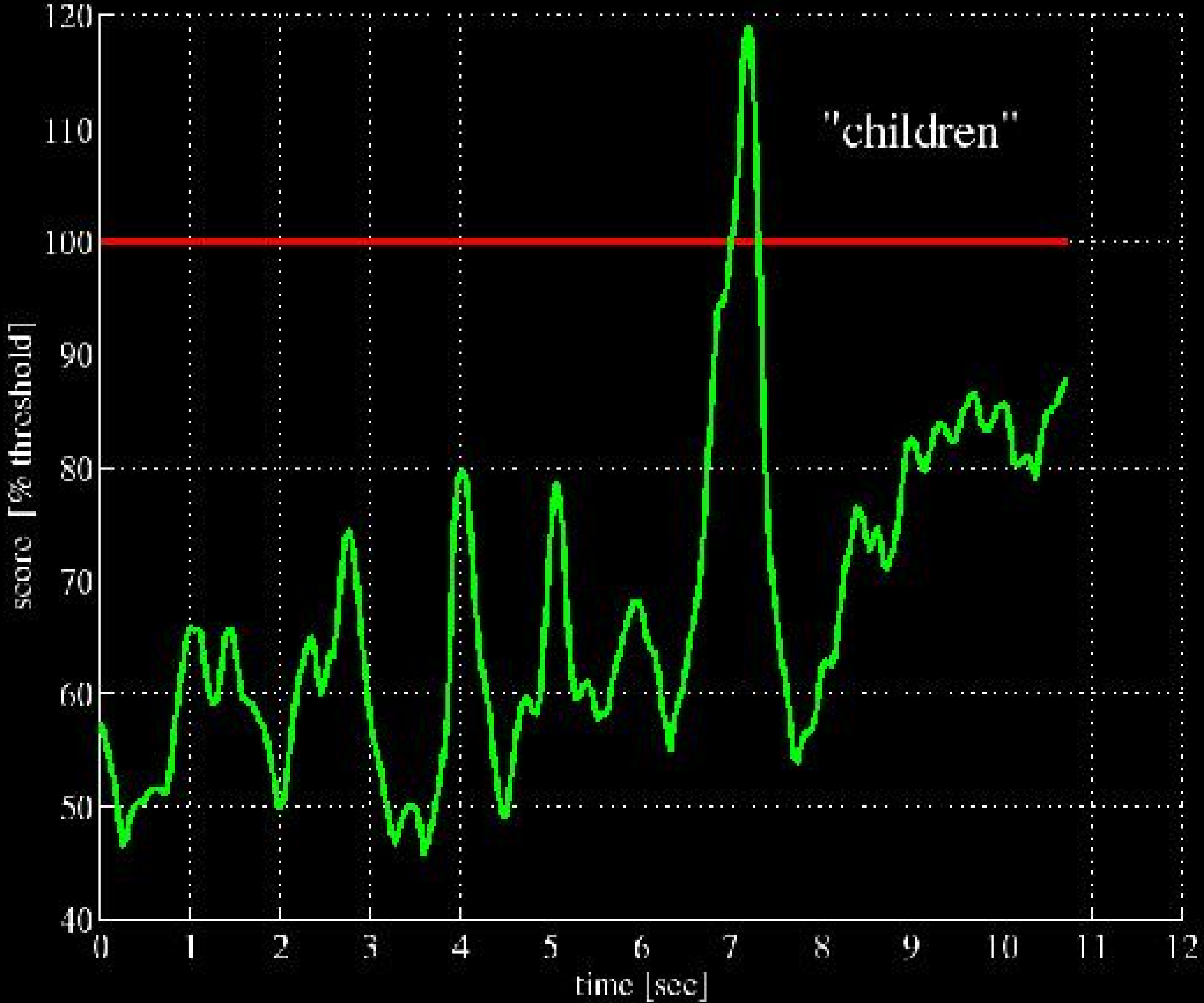
Isolated word spotting



- Using true articulator movements for speaker dependent isolated word recognition

(shown here: jw45/tp025)

Continuous speech

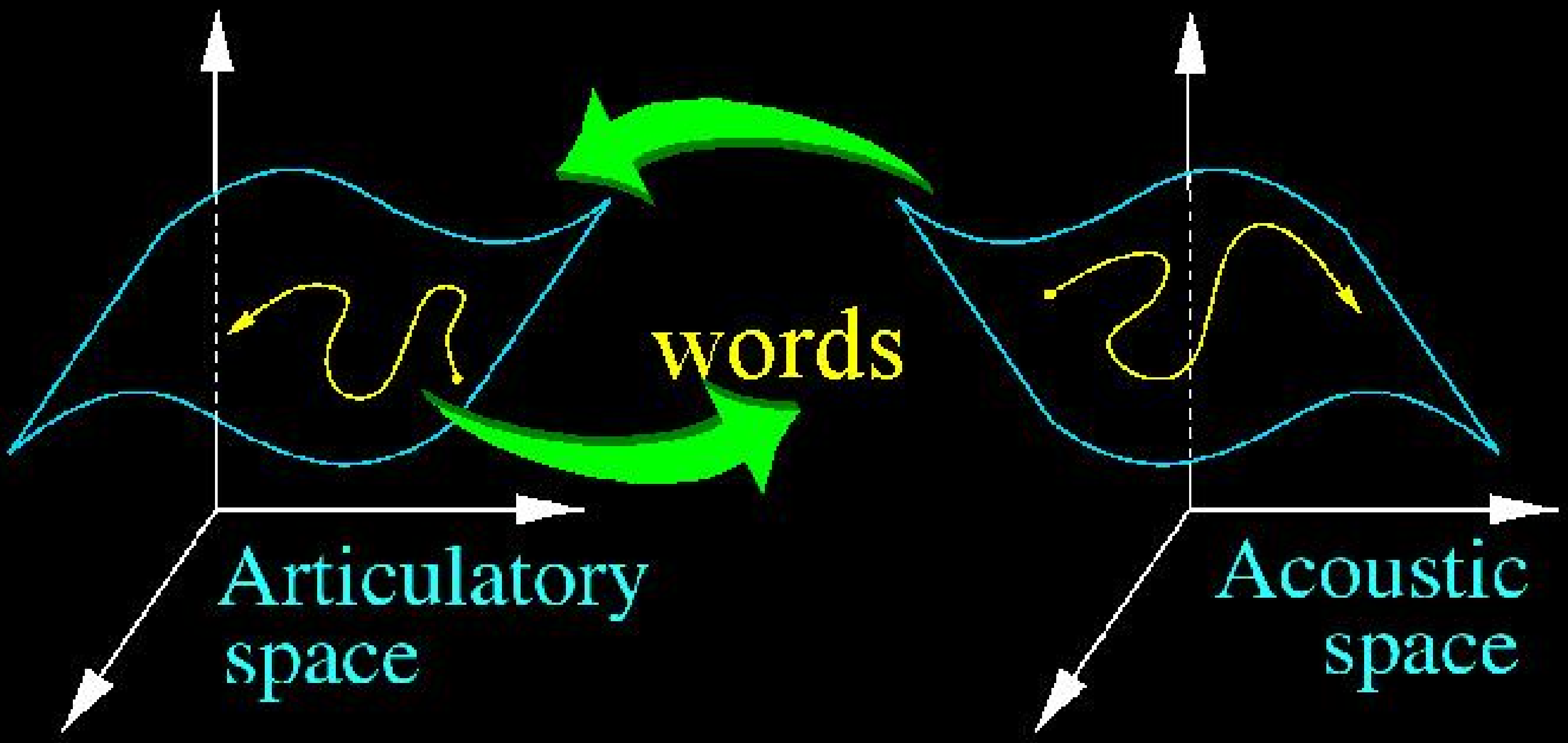


- Using true articulator movements for speaker dependent continuous speech recognition

(shown here: jw45/tp064)

Articulatory speech recognition

- Recover movements and then do recognition



Results: one small step...

- Use one instance of a word as a template (!)
- Look for matches in all isolated word tasks (total of about 400 words)
- Simple dynamic time warping algorithm using recovered articulatory movements.

"science is fun"
s-ay-ax-n-s ih-zs f-ax-n

- Result?
  Perfect performance on a really easy task.

Articulatory word spotting

- Using recovered articulator movements for speaker dependent isolated word recognition

(shown here: jw45/tp002)

Other research groups

- Los Alamos National Labs (Hogden, Nix, Zlokarnik)
- Rutgers CAIP (Flanagan, Sinder, Chennoukh)
- Cambridge (Blackburn)
- MIT (Papcun)
- Bell Lab (Sondhi, Schroeter, Levinson)
- Waterloo (Deng, Ramsey, Sun)
- Caltech (Barr, Fain)

Representation is everything

- Hard computations are best solved by knowing the right way to look at the problem
- Probabilistic generative models explain variability and separate signal from noise.
- They apply to more than speech.
- Consider handwritten digits: in the right representation, dynamic time warp can be used for recognition!

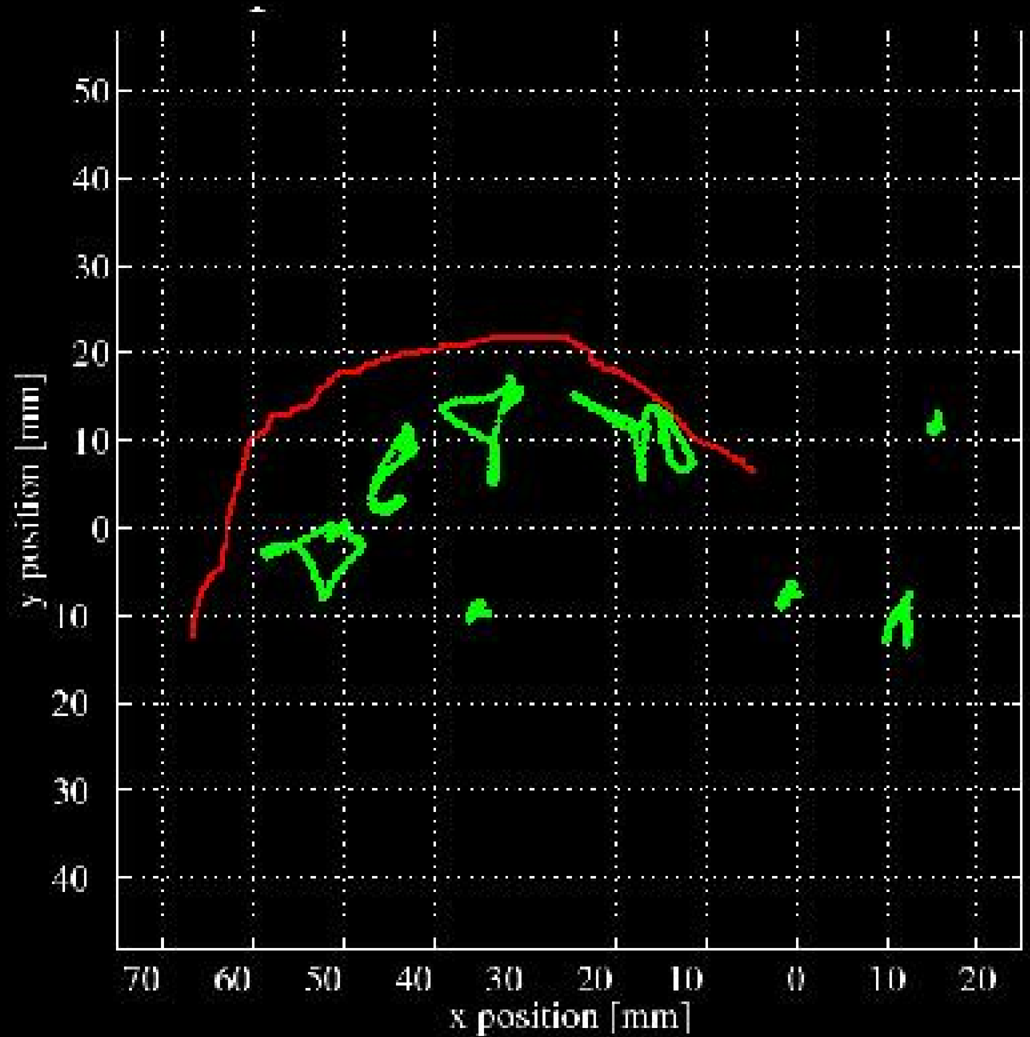
Acknowledgements

- John Hopfield and group (Caltech/Princeton)
- Abeer Alwan & Pat Keating (UCLA)
- Pietro Perona & Yaser Abu-Mostafa (Caltech)
- Bell Labs colleagues
- Simon Blackburn, Dan Fain, John Hogden
- John Westbury and his team (Wisconsin)

- special thanks to Laura Rodriguez

Phase space

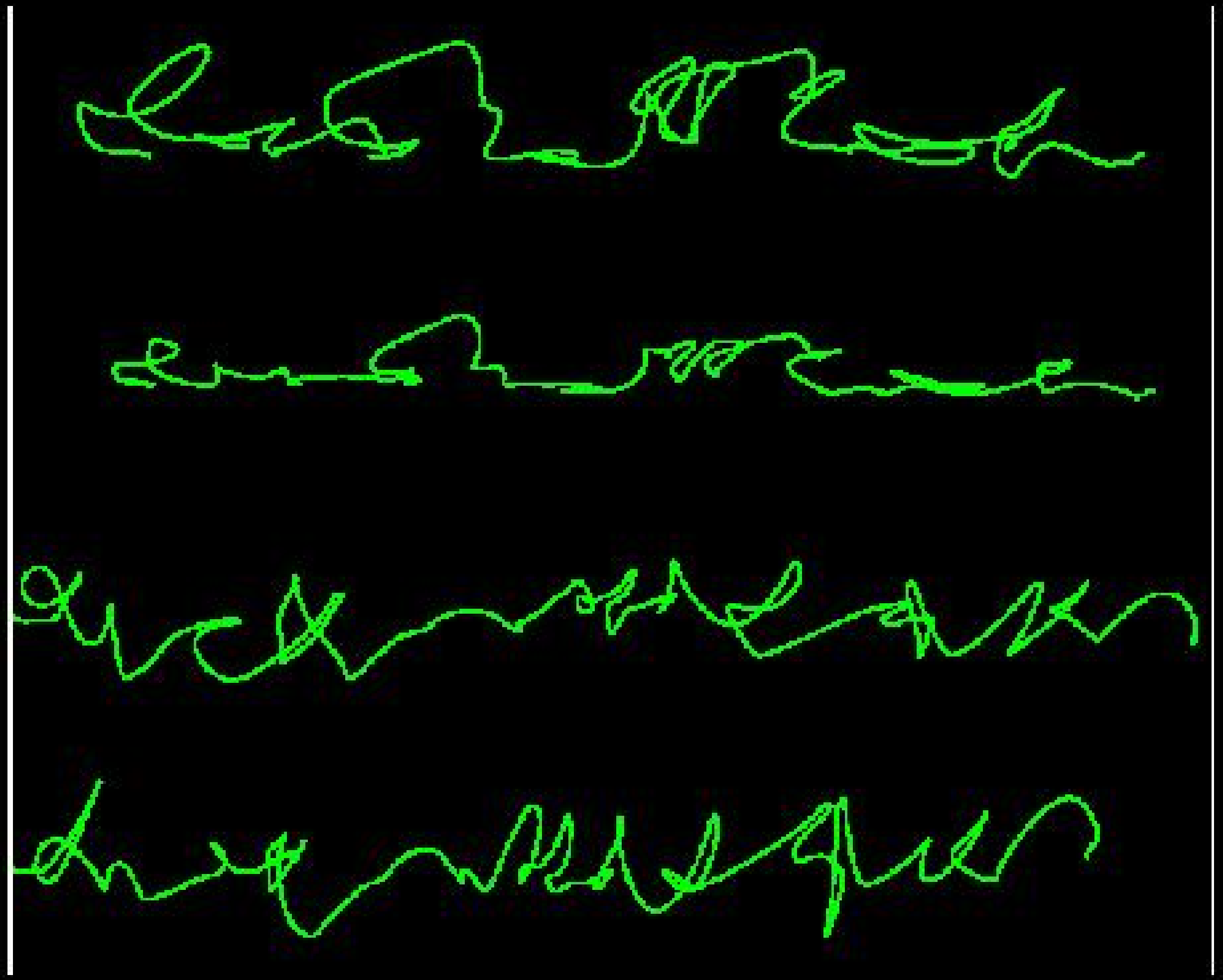
- If a pen were attached to each bead, the resulting traces would constitute a "phase space" representation in which time was implicit.



Articulatory “signatures”

- Now slide the paper underneath the pens at a constant velocity in a constant direction.

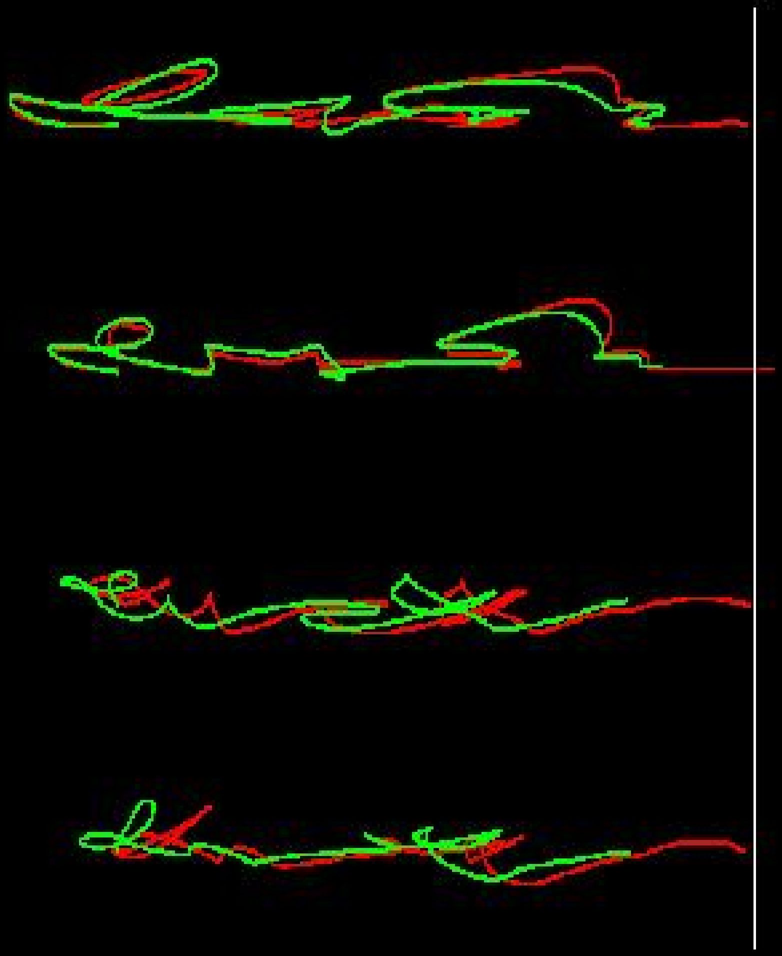
(four tongue beads shown here)



Signatures for recognition?

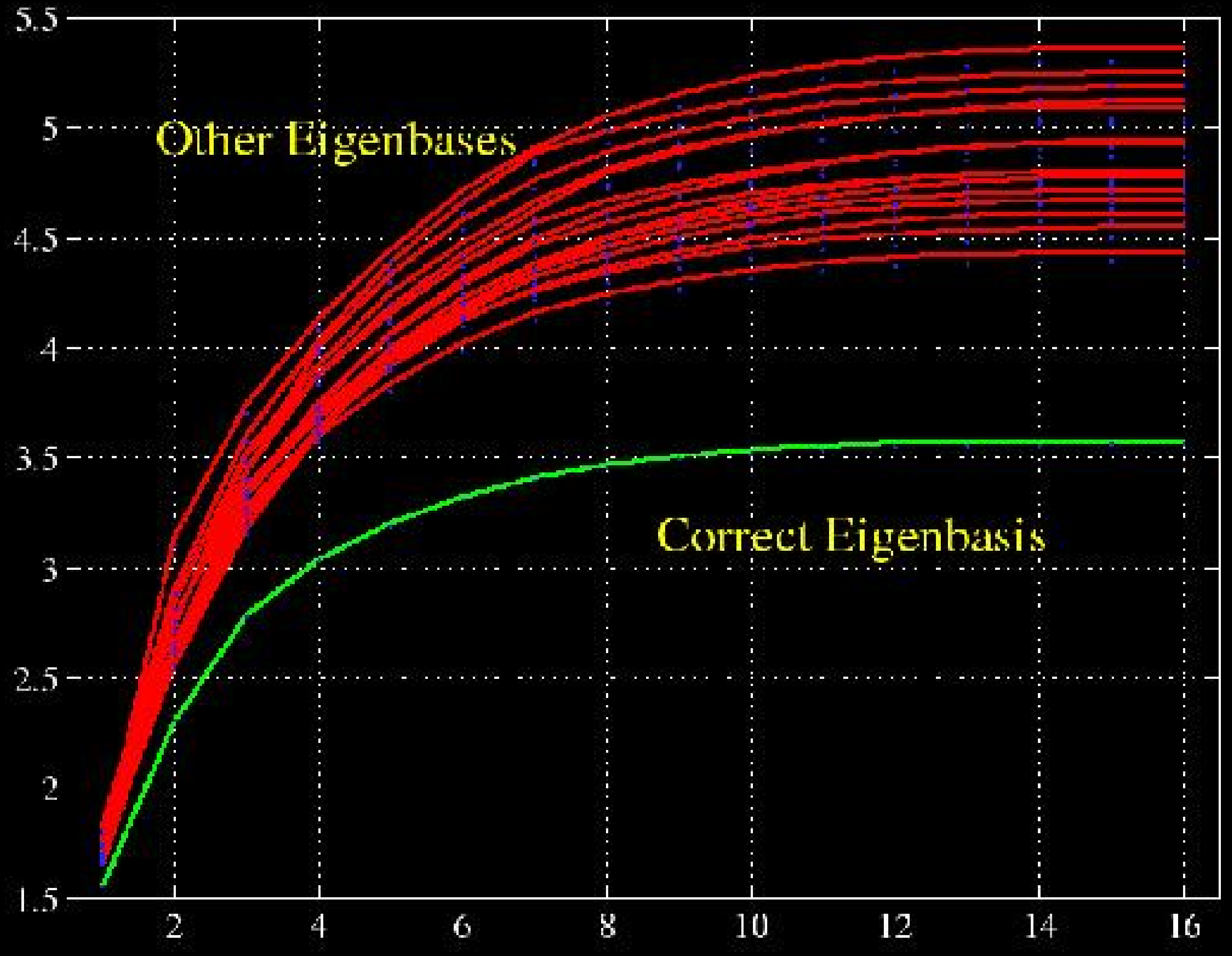
- Two repetitions of the same phrase by the same speaker at different times.

- Dynamic time warping assumes that all variation is in time. Signatures allow variation in a mix of time-space variables.



Speaker identification

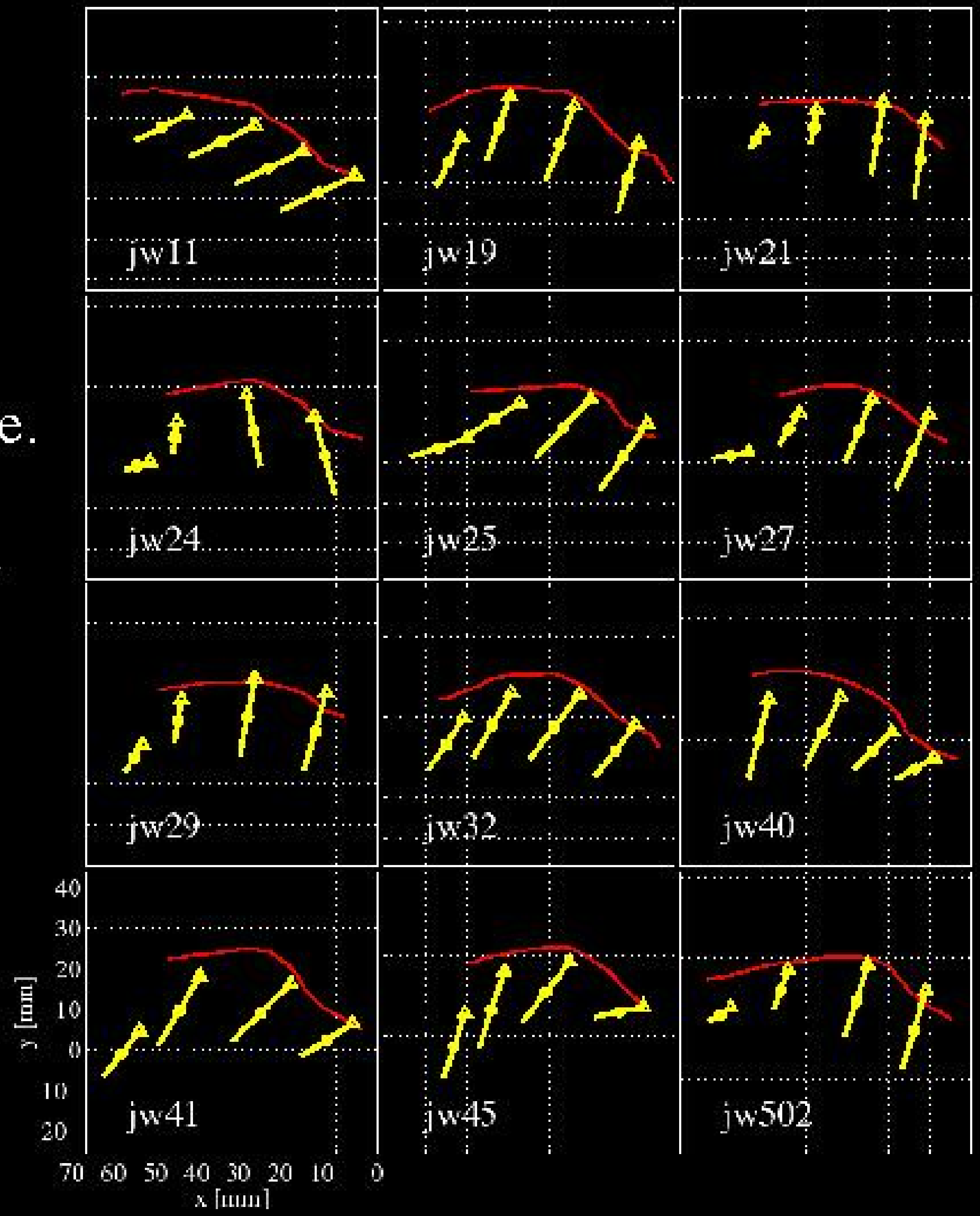
- Speaker identification using articulator positions (no means)
- Time independent
- Lines for 10sec of data, dots for 1sec



What do the eigenvectors look like?

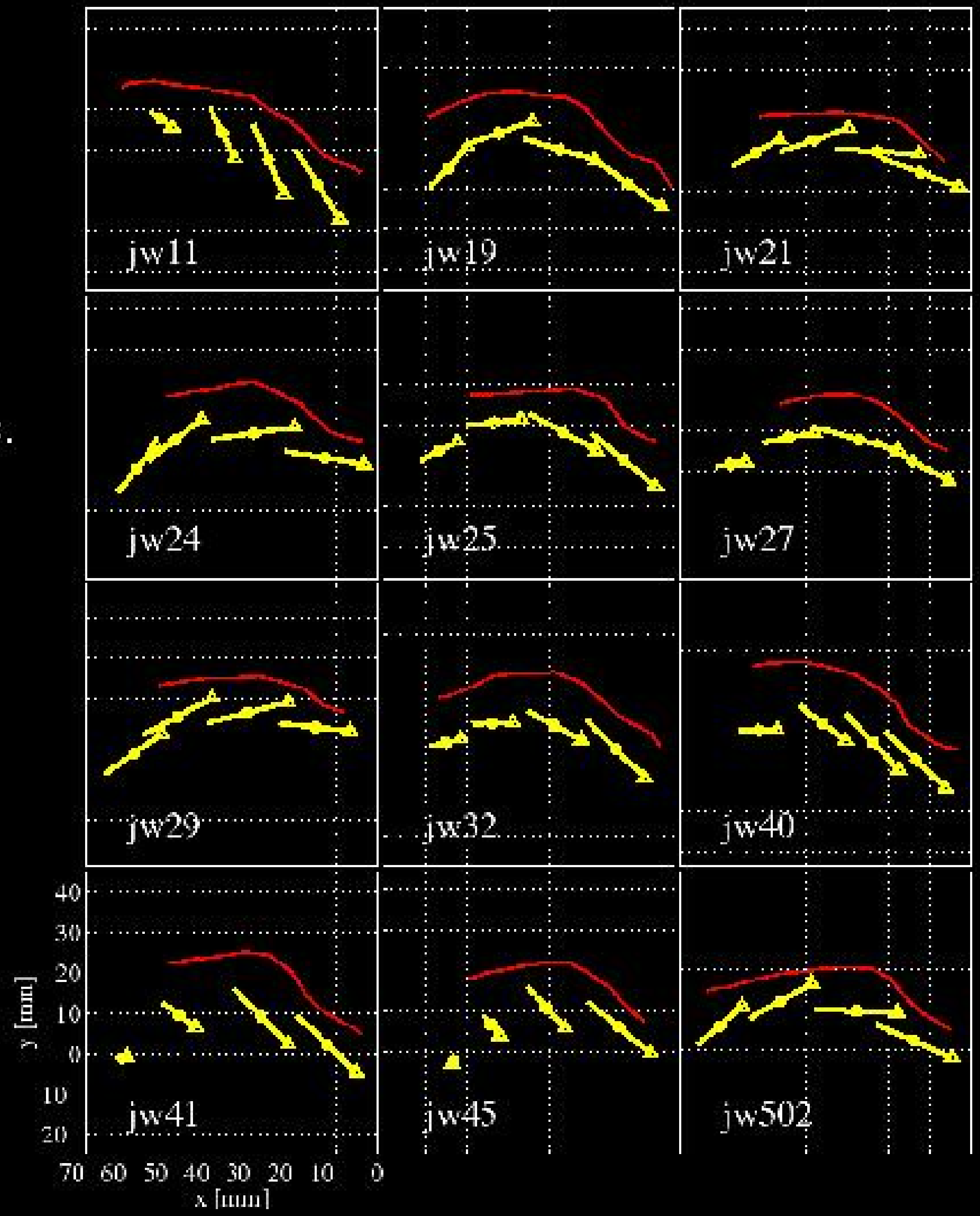
Leading eigenmode.

Towards and away from the palate.



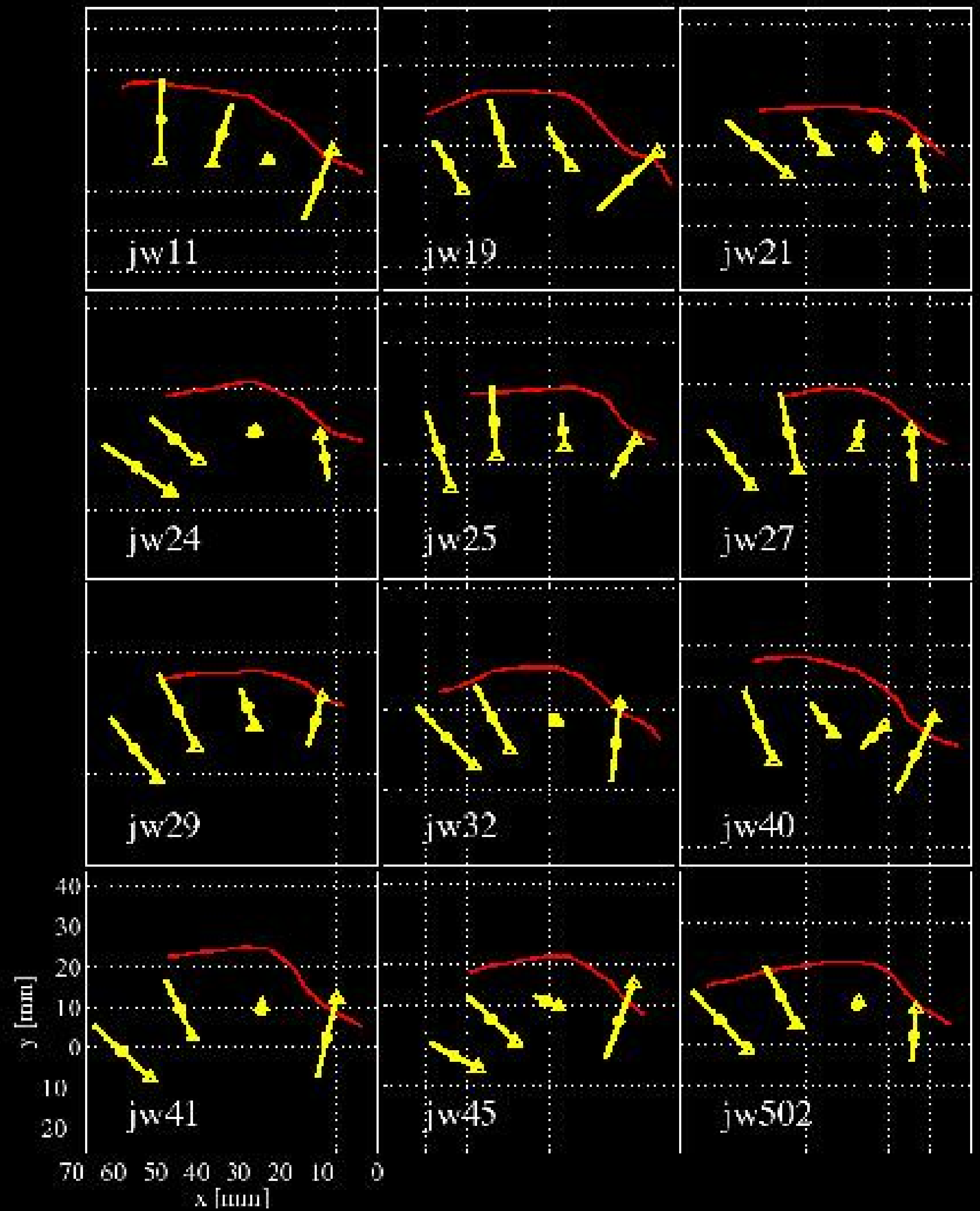
Second eigenmode.

In and out
of the mouth
(parallel to palate).



Third eigenmode.

Rocking/pivoting of the tongue.



Some history

- "Universal Service" (1910,1915,1934)

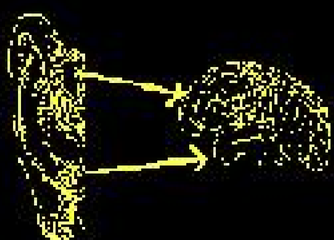
- Radio Rex toy (1922)
- Sonograph, Audrey (1950,1952)
- Theory: HMMs and Viterbi (late 1960's)
- Pierce's caustic letter (1969)
- LPC,DTW, ARPA (early 1970's)
- HMMs in speech (mid 1970's)
- Theory: *EM* algorithm (1977)
- IDA symposium in Princeton (1980)

use what you know

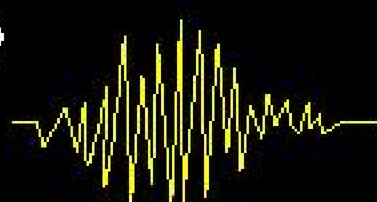
science is fun

~~swim green why fish~~

possible messages constrained → language modeling



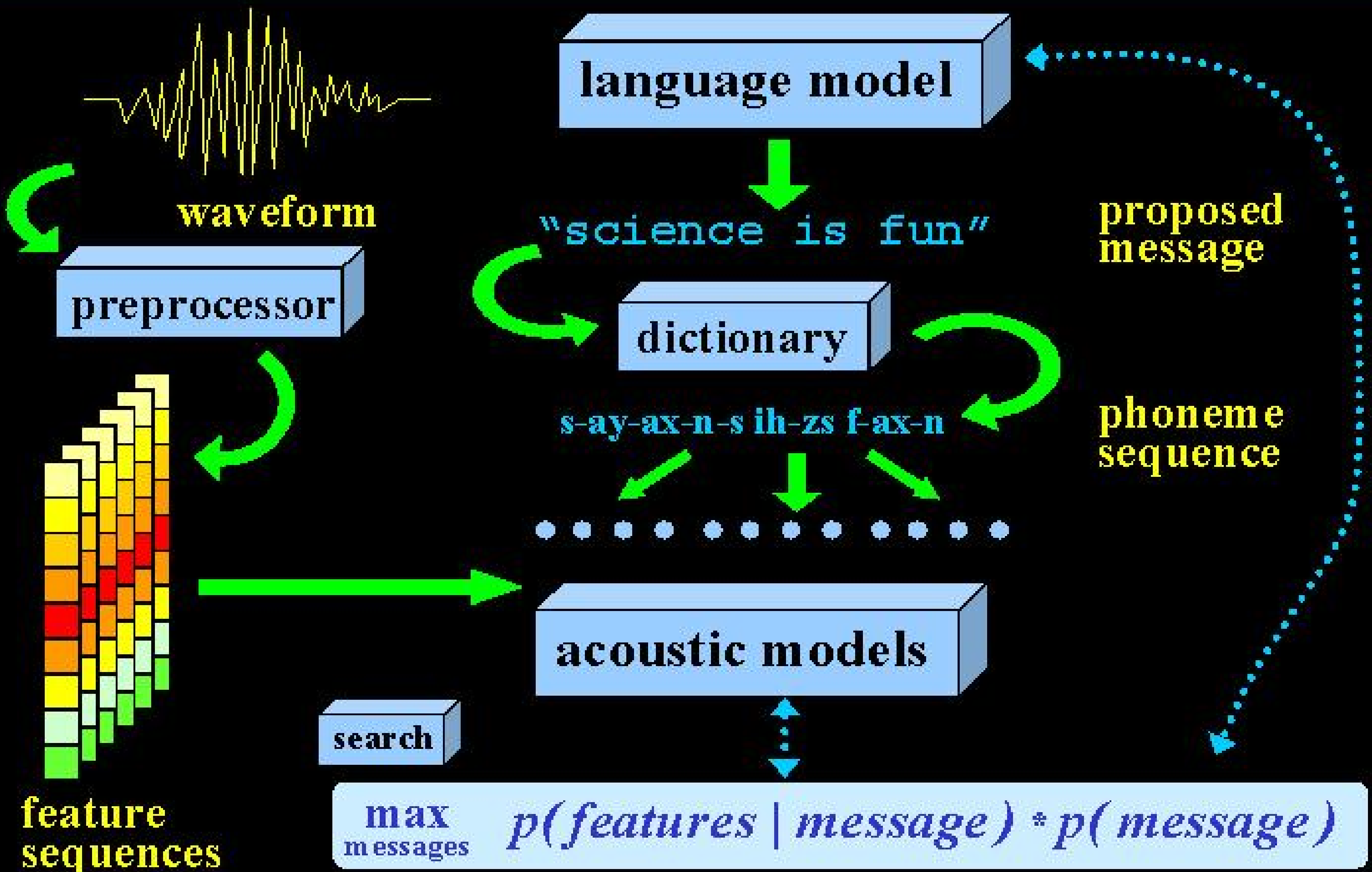
decoder constrained → perception studies for noise & preprocessing



encoder constrained → production studies for variability robustness

my work

Current approach: statistical model



Problem solved ?

- Speaker ID: noisy, mixed gender, instant add

  false rej = 0.1% hundreds of users (open)
  false acc = 0.01% (thousands if set phrase)

- Synthesis: http://www.att.com/aspg/odemo.html
- May 1997 recognition evaluation from NIST:

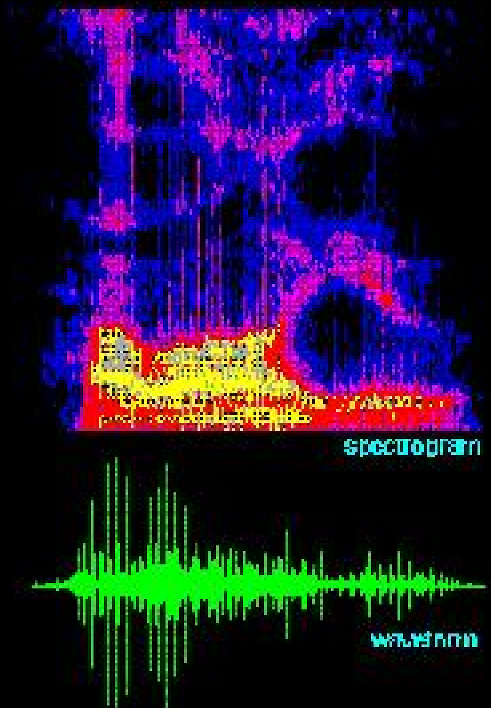
System	SWB	CH	AVG(wer)
BBN.......	35.5	53.7	44.9
BU........	41.5	58.2	50.1
CMU-ISL...	35.1	54.4	45.1
CU-HTK....	39.2	57.6	48.7
DRAGON....	39.9	57.4	48.9
SRI.......	42.5	57.5	50.2

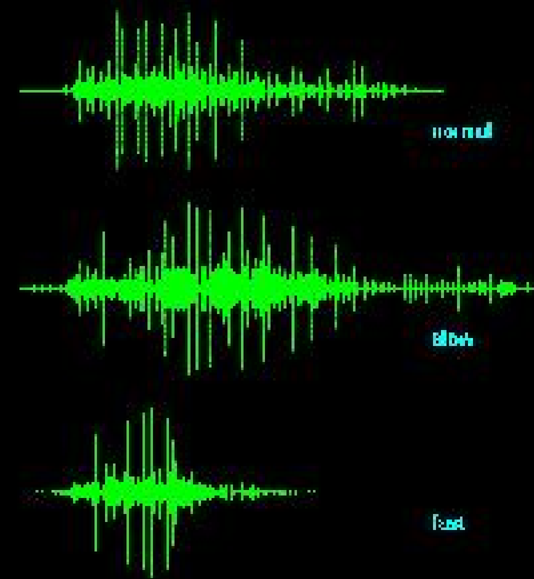
(typical: 300MB Sparc ULTRA takes 800x realtime)

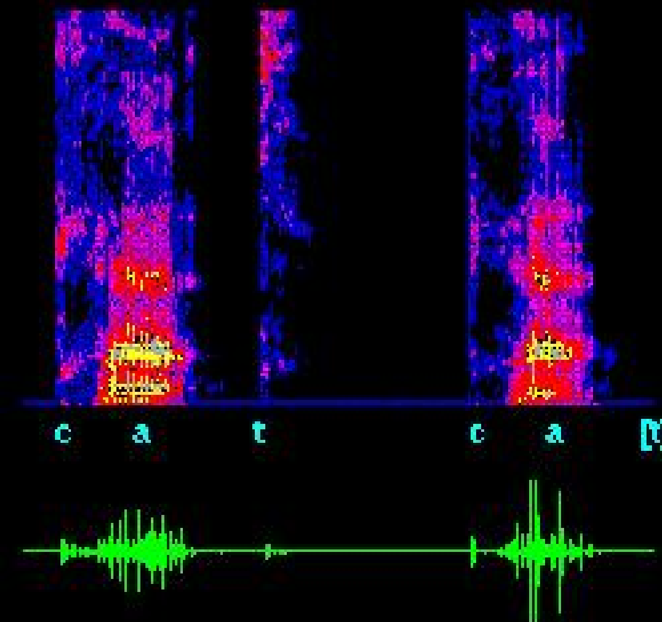
what hurts current systems?

- spontaneous, informal, conversational speech breaks the language models ✗
- noisy, reverb/coloured, multi-source speech breaks the acoustical preprocessor ✗
- variability between & within speakers (e.g. deletion, co-articulation, rate, prosody) breaks the pattern recognition algorithms ✗

examples of variability







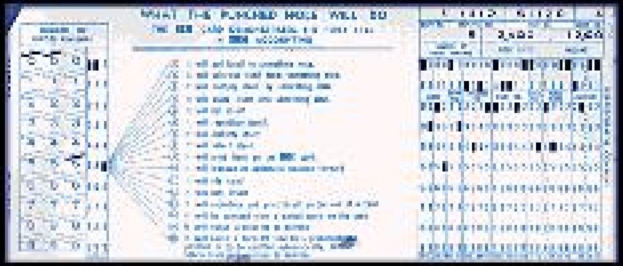
"how are you"

- rate
- co-articulation
- deletions

Forces driving the technology

- Computers are everywhere
  Good human-machine interface?
  How about the spoken word.

- Electronics are getting tiny
  "You can't type with toothpicks."

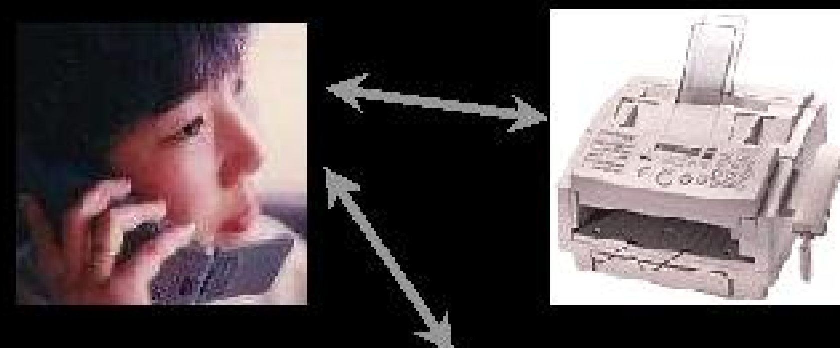


- Computers move & access
  a lot of information,
  do very complex tasks,
  and all in real time. Interactive systems.

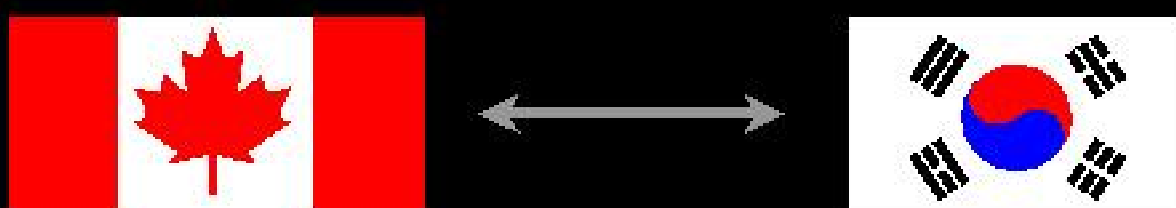
three cool applications

- Blind reading & hands-free typing

- Phone to fax/email

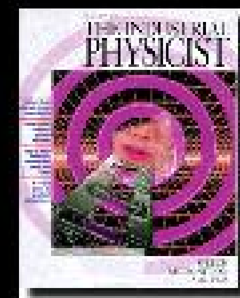


- Real time interactive translation



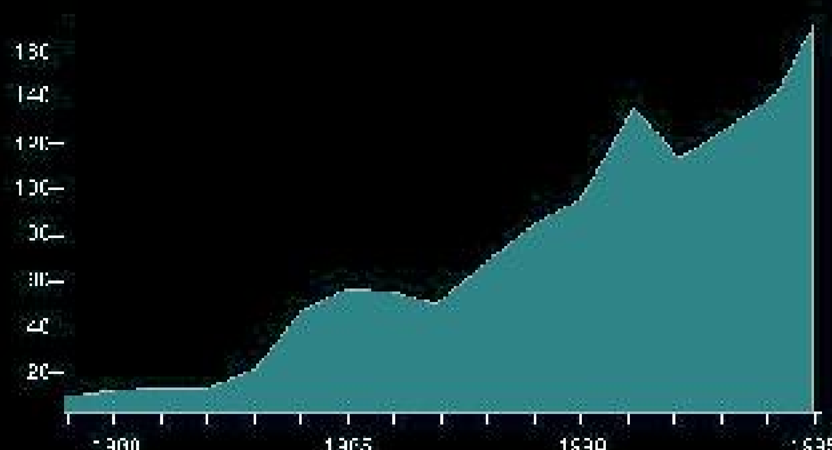
Excitement

- Magazine covers



- Patents/year

Patents as a Function of Time

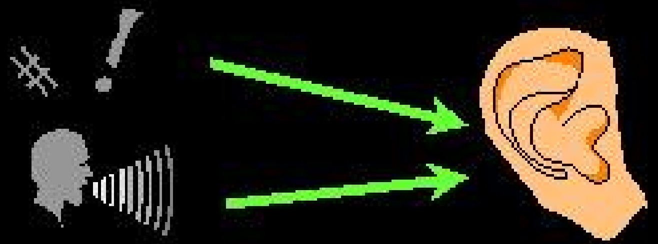


- National Medal of Science

future research areas

- better acoustical pattern recognition
- Video!

- Source separation, noise reduction



- Synthesis: prosody, emotion, voice conversion
- more sophisticated language models

ex: "umm...ok, how about going it alone?"

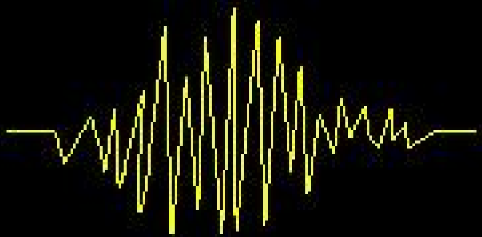
A research paradox?

- half-century of work by an army of academic & industrial researchers
- state of the art >50% word error rate on a noiseless task
- no real breakthroughs in more than 20 years
- too hard for machines?

- enormous amount of wonderful "free" data (100's hours)
- huge compute power (9Gb RAM)
- input/output/errors all well defined
- everything but the algorithm (which exists!)

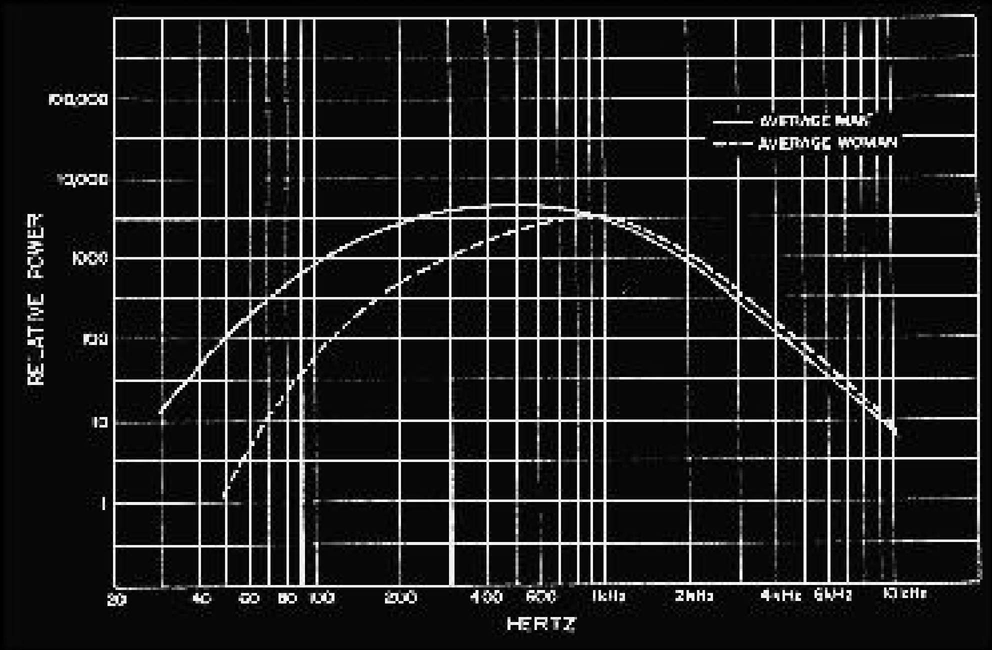
Statistics of natural speech

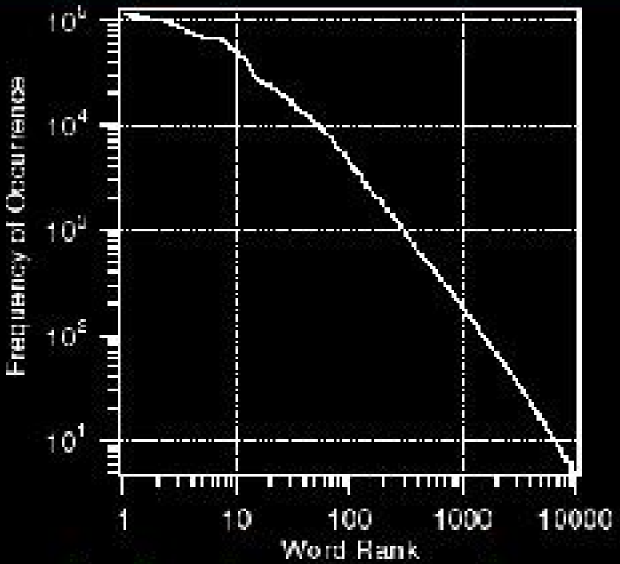
speech is a 1d pressure wave signal with:



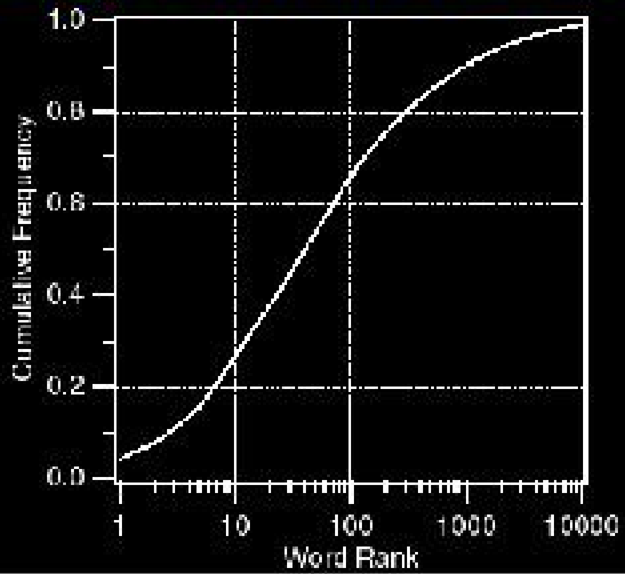
a logarithmic amplitude distribution

a smooth, power law (1/f) spectrum

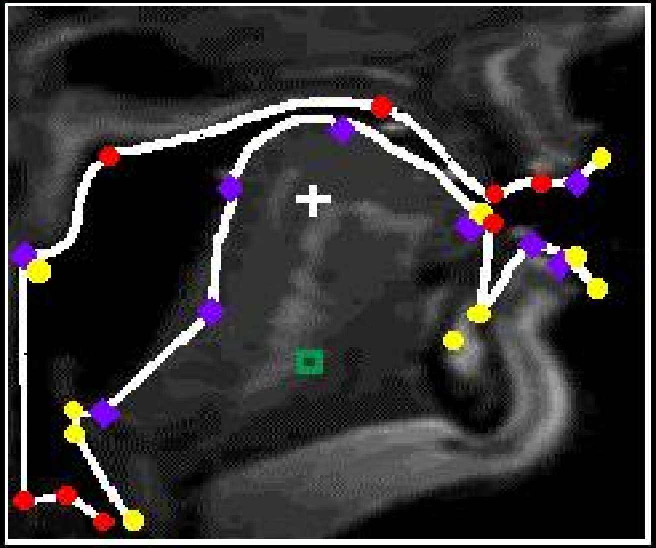


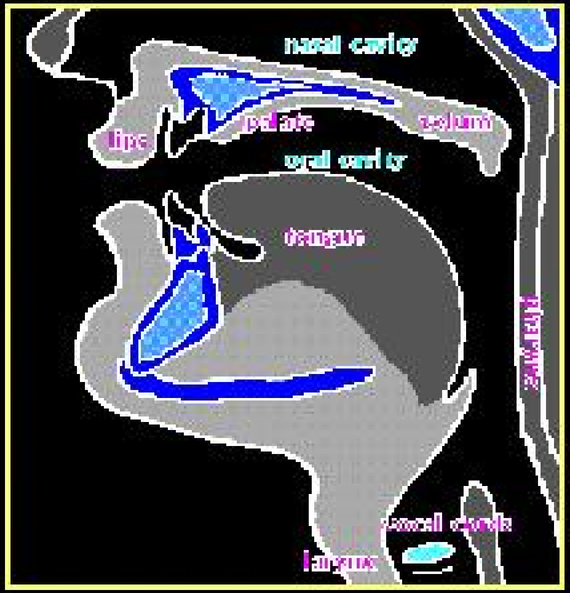


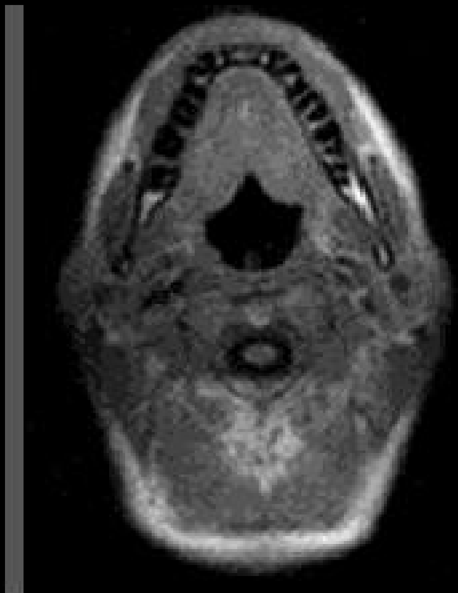
Zipf's law word stats

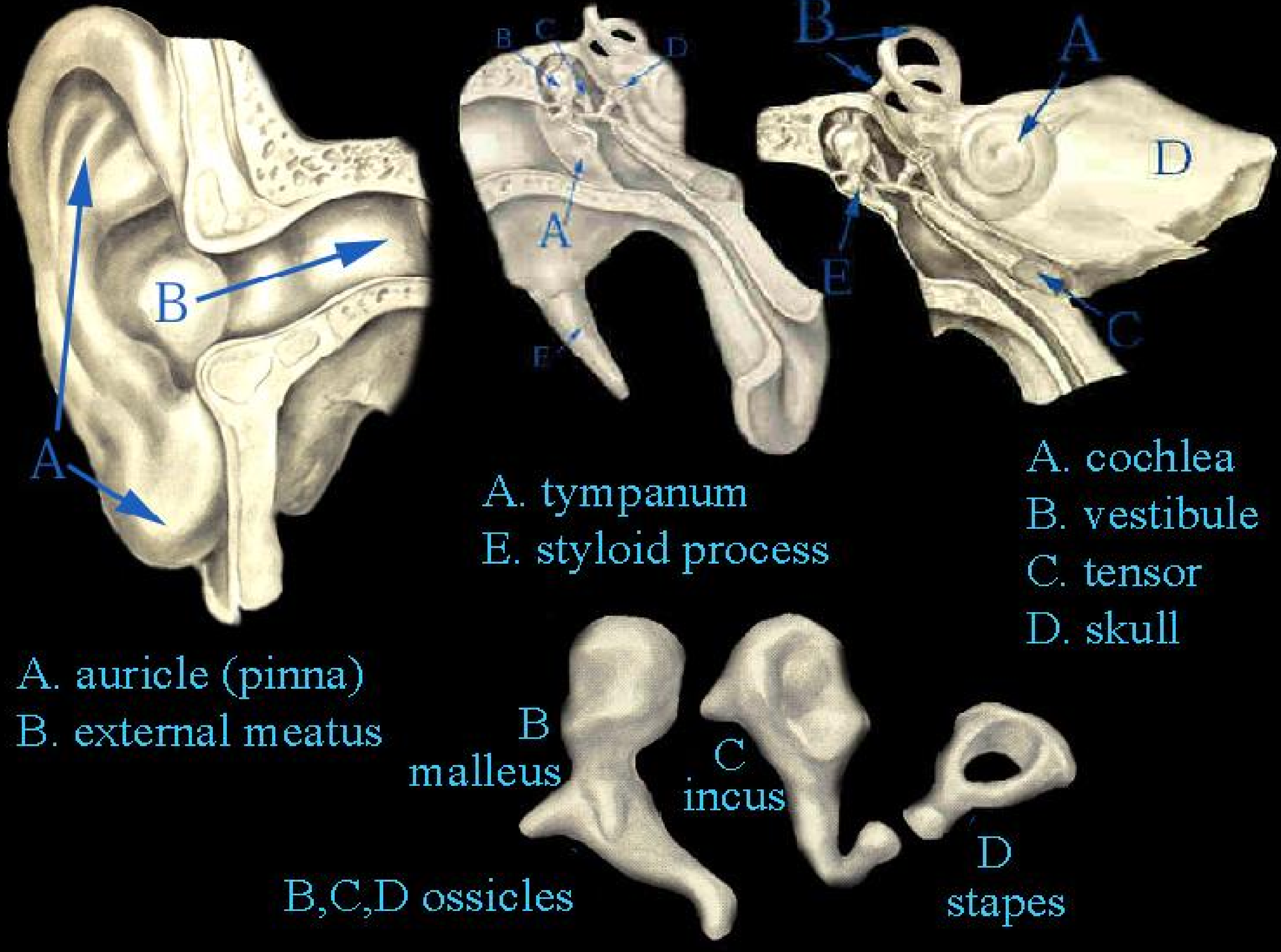


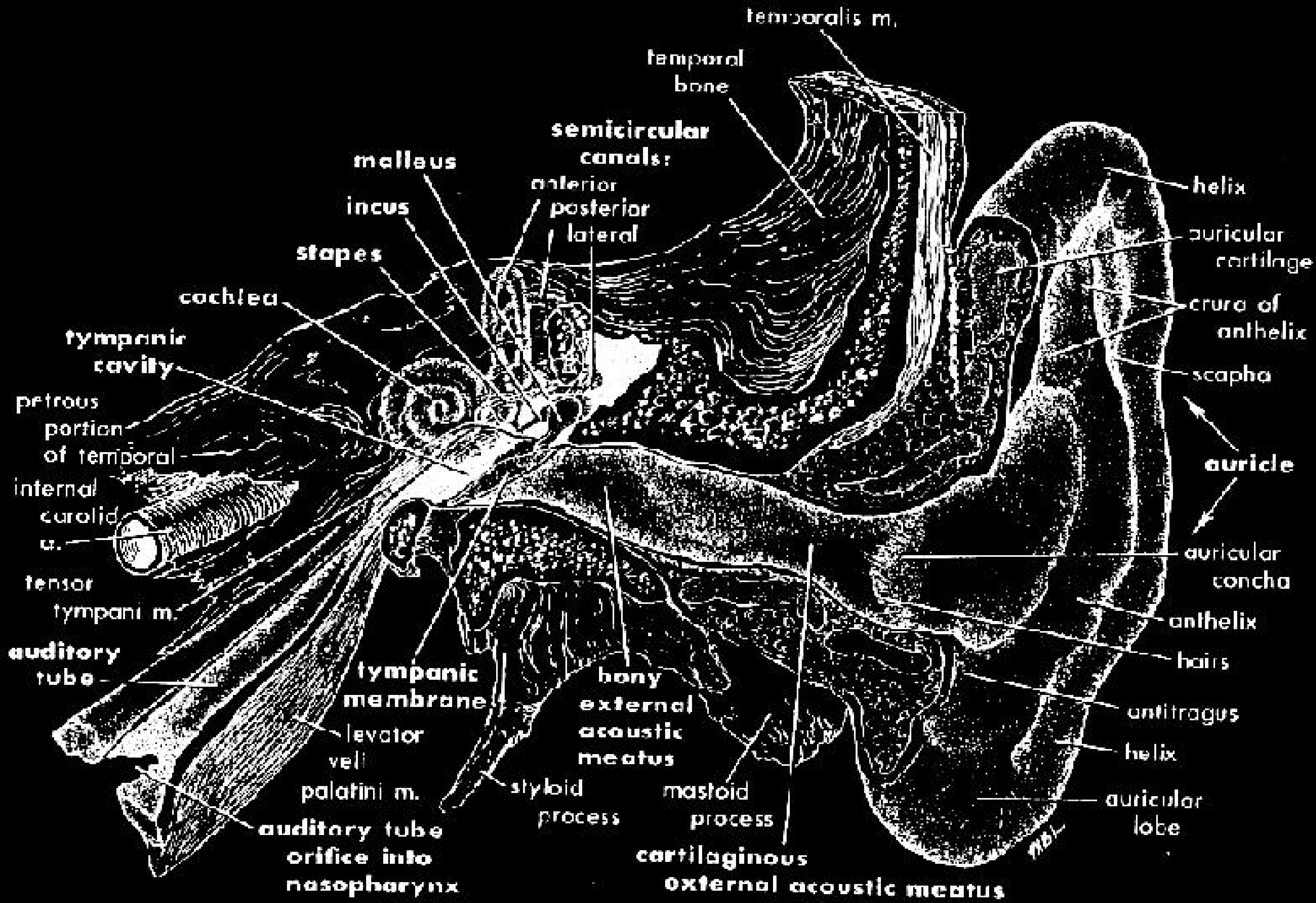
Speech production organs








A
B
B
C
D
A
E
B
A
D
E
C
A. auricle (pinna)
B. external meatus
A. tympanum
E. styloid process
A. cochlea
B. vestibule
C. tensor
D. skull
B
malleus
C
incus
D
stapes
B,C,D ossicles



temporalis m.
temporal bone
semicircular canals:
anterior
posterior
lateral
malleus
incus
stapes
cochlea
tympanic cavity
petrous portion of temporal
internal carotid a.
tensor tympani m.
auditory tube
tympanic membrane
levator veli palatini m.
auditory tube orifice into nasopharynx
styloid process
bony external acoustic meatus
mastoid process
cartilaginous external acoustic meatus
helix
auricular cartilage
crura of anthelix
scapha
auricle
auricular concha
anthelix
hairs
antitragus
helix
auricular lobe

Speech perception organs