



THE UNIVERSITY *of* EDINBURGH
informatics

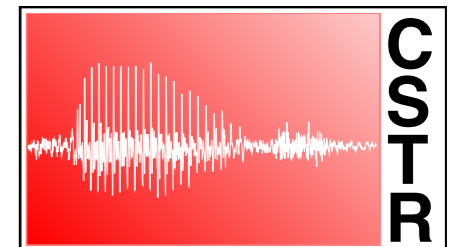
On Measuring and Estimating Speech Articulation

Dr Korin Richmond

Workshop on Speech Production in Automatic Speech Recognition

August 30th 2013

Lyon, France

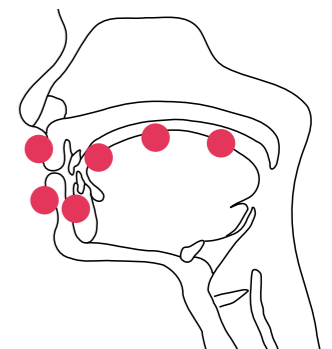
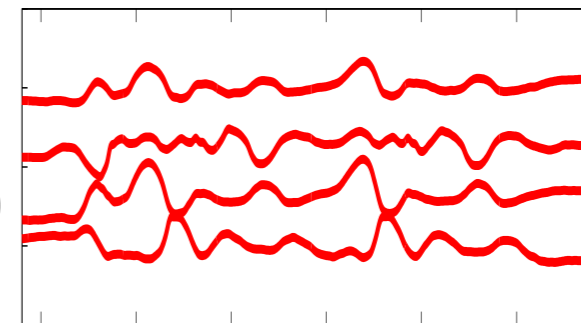


Talk outline

1. Why articulation?
2. Available articulatory data
3. Inversion with Artificial Neural Nets (ANN)
 - MLP v's Mixture Density Network
 - Deep ANN models
4. Is this any good though?
 - is it an adequate articulatory representation?
 - what is the best performance possible?
5. Summary

Why might articulation be useful?

- An articulatory representation of speech has attractive properties
 - relatively slow, smooth
 - physical constraints - (e.g. no “jumps”)
- Constraints potentially useful for
 - low bit-rate speech coding
 - speech synthesis
 - speech training
 - avatar animation/lip synching
 - and of course ASR...



How to capture speech articulator movements?

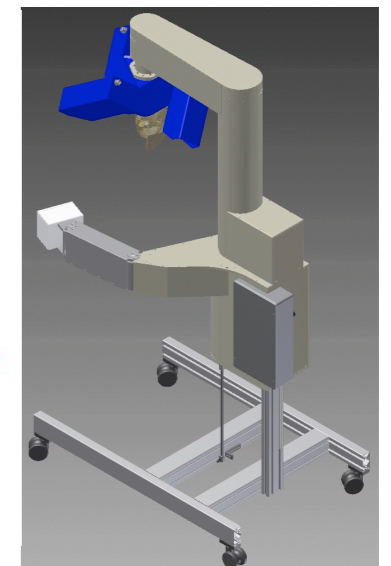
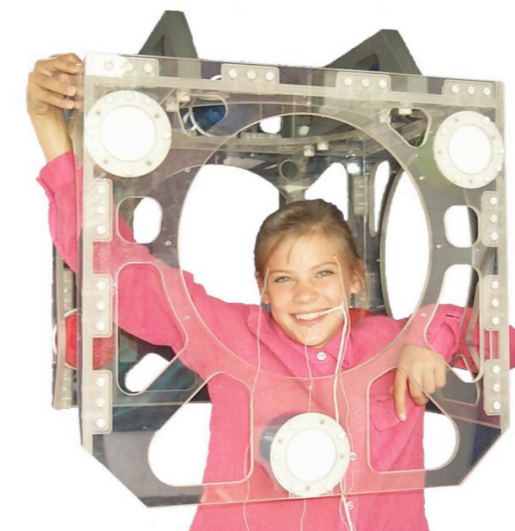
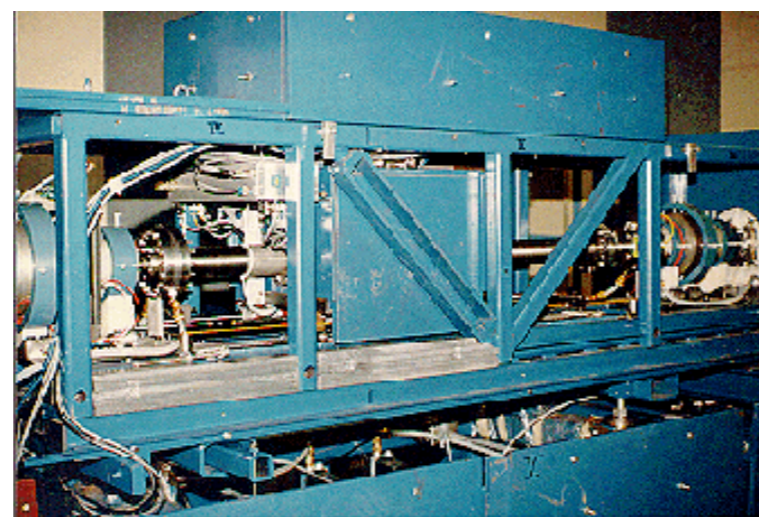


- **Imaging:** Video, Ultrasound, X-ray Cineradiography, MRI

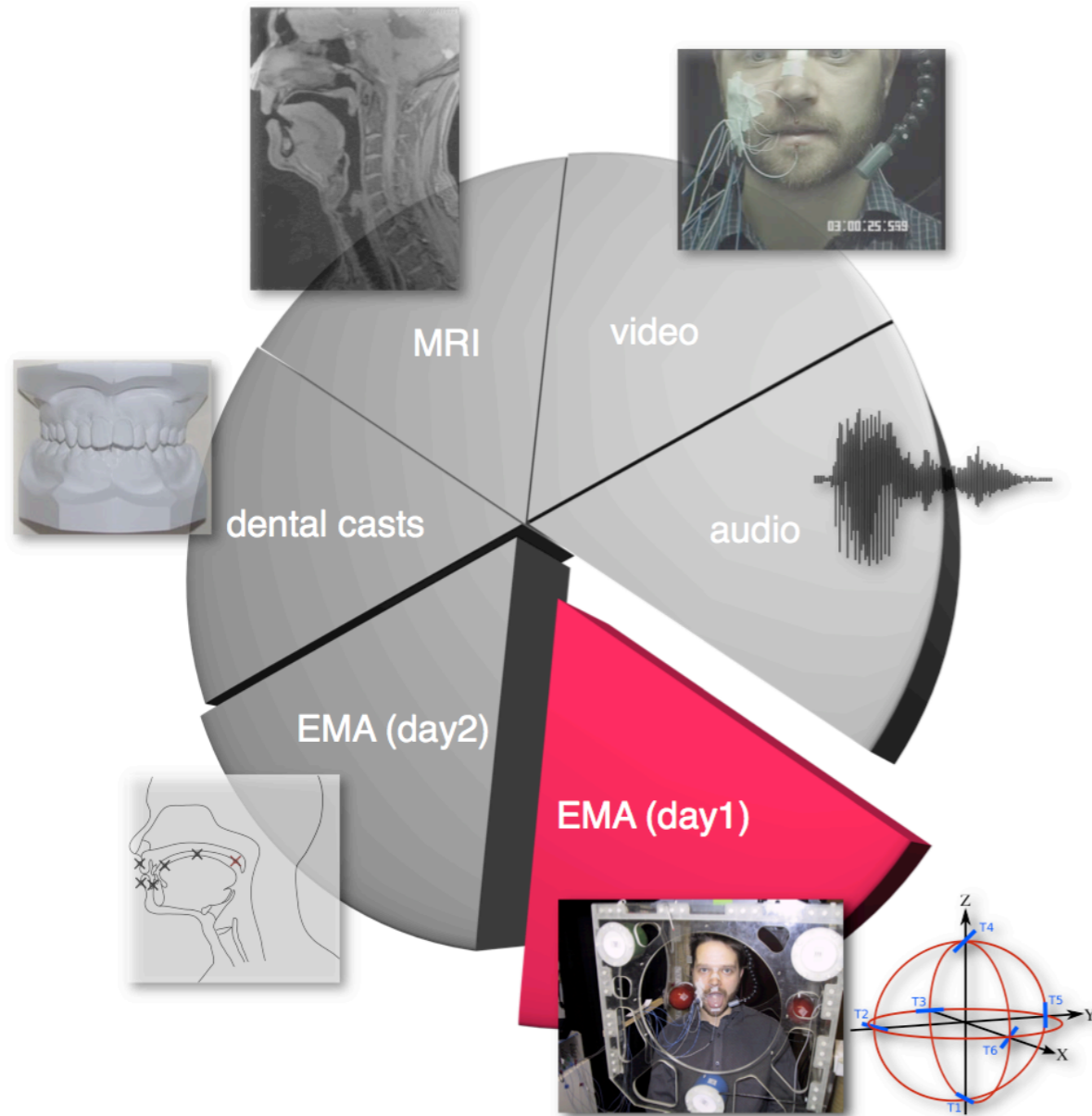
- **Contact:** Electropalatography, laryngography



- **Point tracking:** Mocap, X-ray microbeam, Electromagnetic articulography (EMA)



Advertisement: mngu0 articulatory corpus



Collaborators =>

EMA: Phil Hoole (LMU, Germany)
Simon King (Edinburgh)

MRI: Ingmar Steiner (Saarland, Germany)
Ian Marshall (Edinburgh)
Calum Gray (Edinburgh)

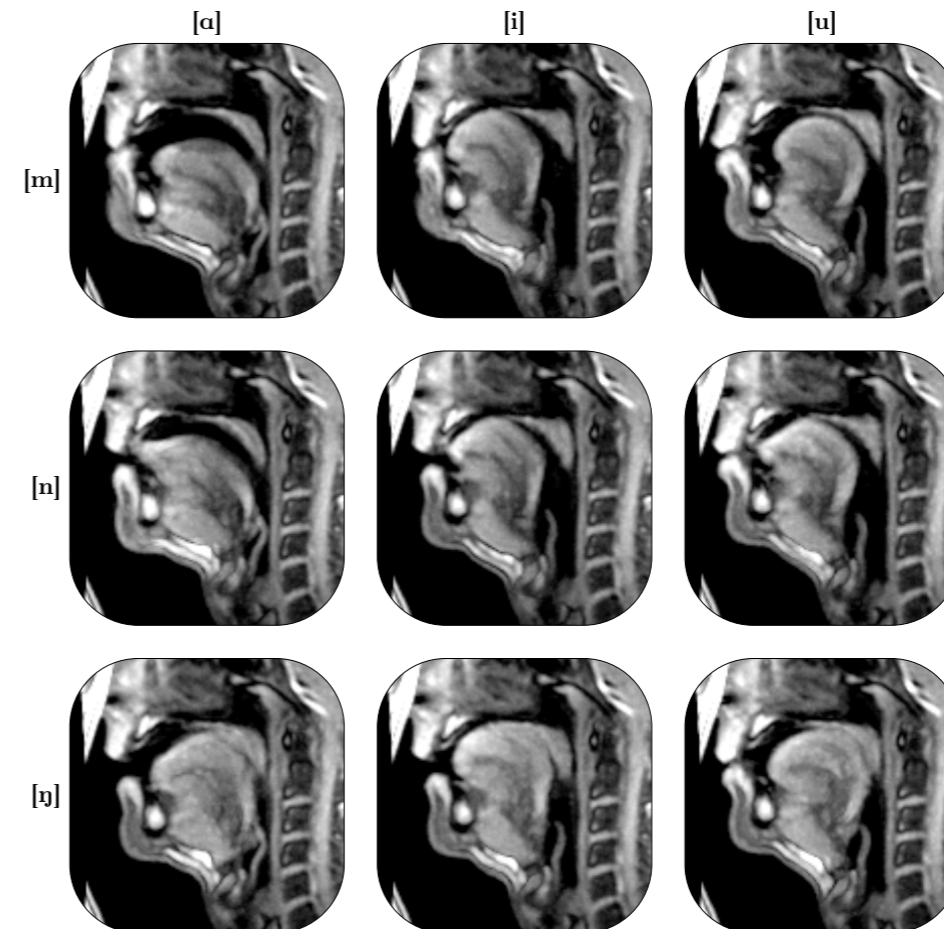
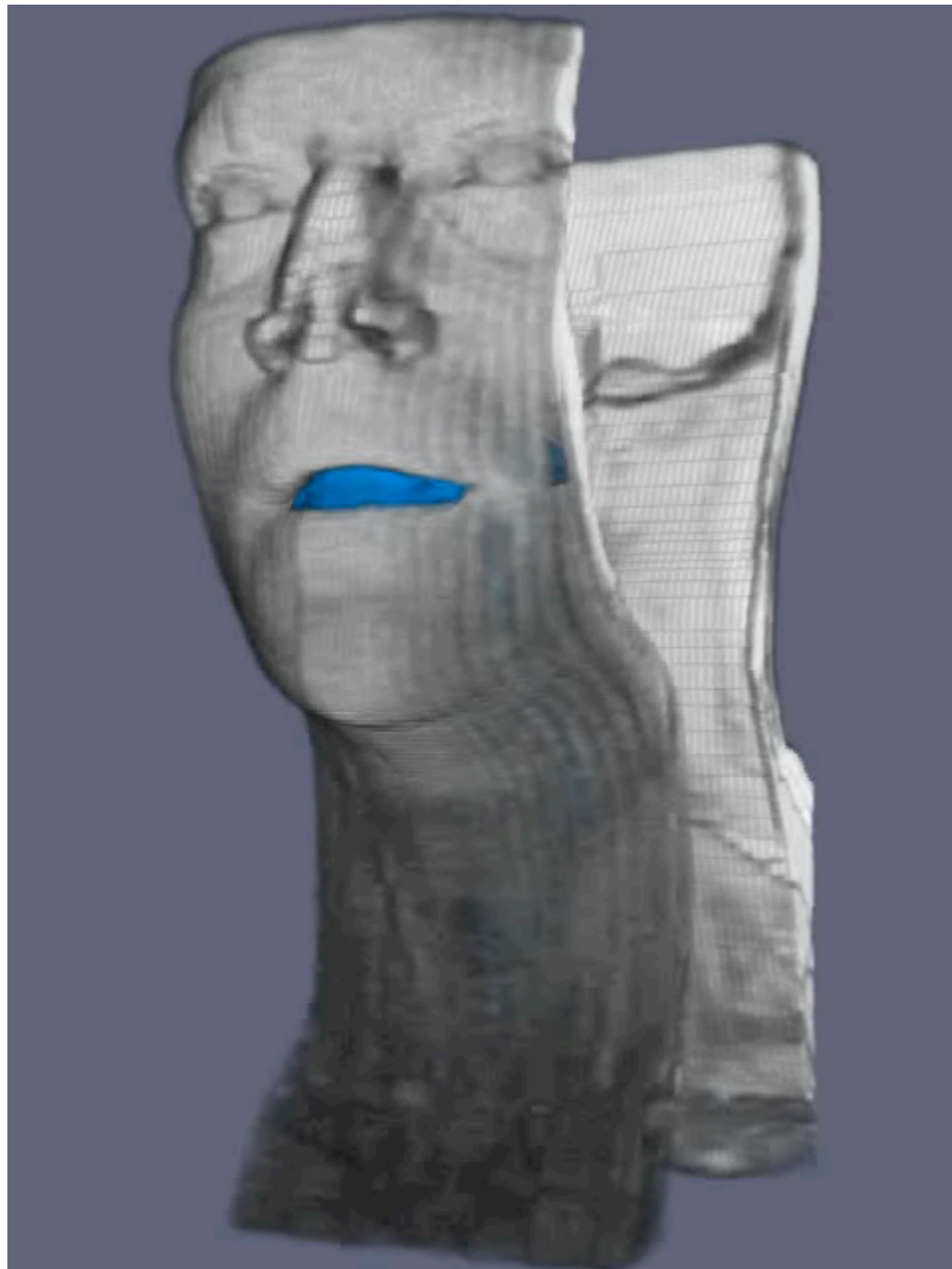
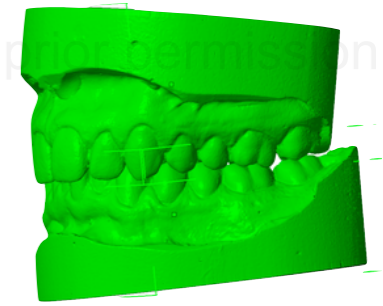
Dental: Laurie Littlejohn (CoreDental, Glasgow)

mngu0 - EMA (day1) data set

- Carstens AG500 EMA
- Articulators: Upper and lower lips, jaw, and three tongue points
- >1,300 phonetically-rich utterances
- Good audio



mngu0 - vocal tract anatomy



- 3D volume (26 slices, 4mm, 256x256px)
13 vowels, 16 consonants
- Midsagittal “dynamic” scans, with 16
cons & 3 vowel contexts (eg. “apa”)
- Acoustic reference recordings

mngu0 web forum

Website => hub for mngu0 activity

Registration required

Default non-commercial licence

User uploads strongly encouraged



The screenshot shows a web browser window with the URL <http://www.mngu0.org/>. The page features a header with a logo consisting of a sagittal cross-section of a human head and the word "MINGUO" in large, bold, black letters. Below the logo is a navigation menu with links for Home, News, Publications, and Register. A search bar is located in the top right corner. The main content area is titled "Welcome to mngu0" and contains the following text:

Welcome to the home of the *mngu0* data set. This is a corpus of articulatory data of different forms (EMA, MRI, video, 3D scans of upper/lower jaw, audio etc.) acquired from one speaker. The purpose of this website is:

1. To distribute the data (in a variety of forms and parameterisations)
2. To provide a repository and forum for research work which makes use of this data

The first part of this corpus to be released is a large set of utterances recorded by a male British English speaker in a Carstens AG500 Electromagnetic Articulograph. More details can be found in the initial announcement paper for mngu0, which appeared at the Interspeech 2011 conference.

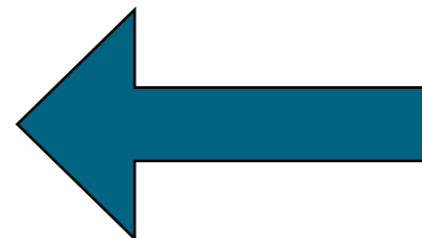
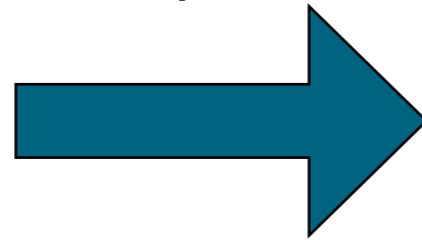
In order to gain access to the mngu0 data and other related resources, it is necessary to register as a user of this website, using the "register" tab above.

At the bottom of the page, there is a footer with the text "All content © 2010-2012 by Korin Richmond" and a row of links: "Powered by Plone", "Valid XHTML", "Valid CSS", and "WCAG".

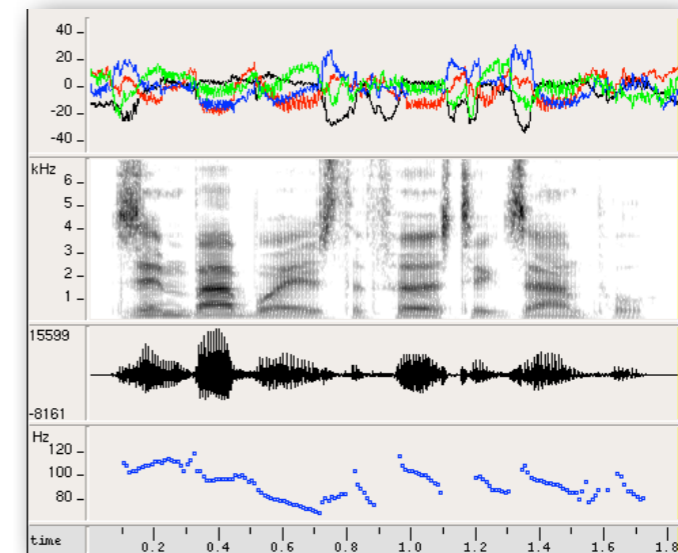
The inversion mapping problem



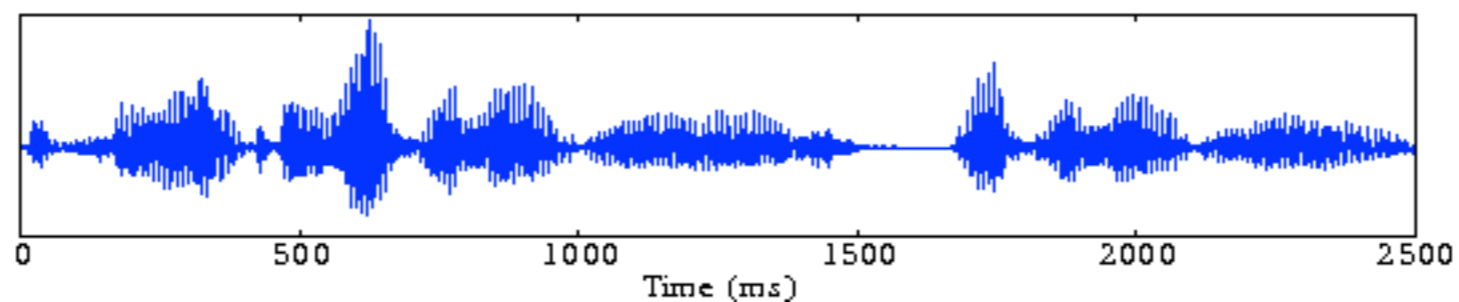
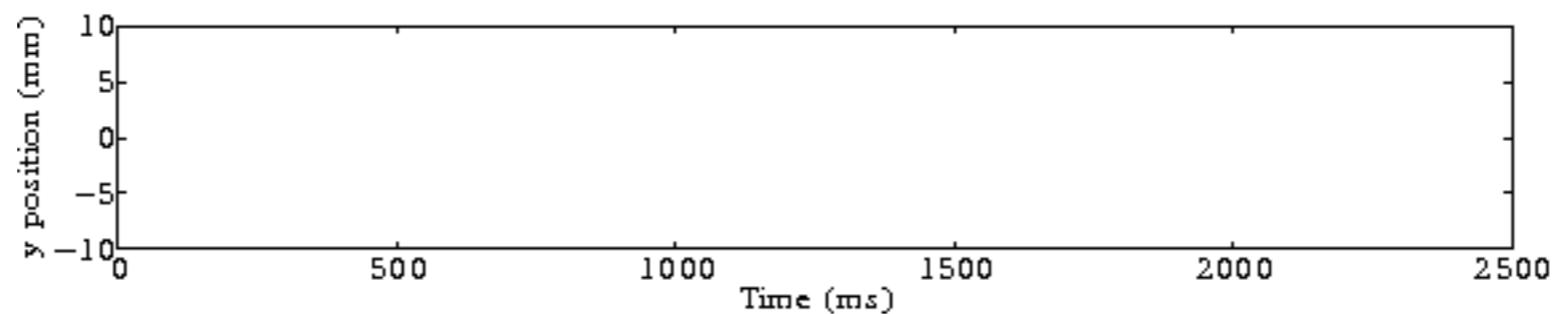
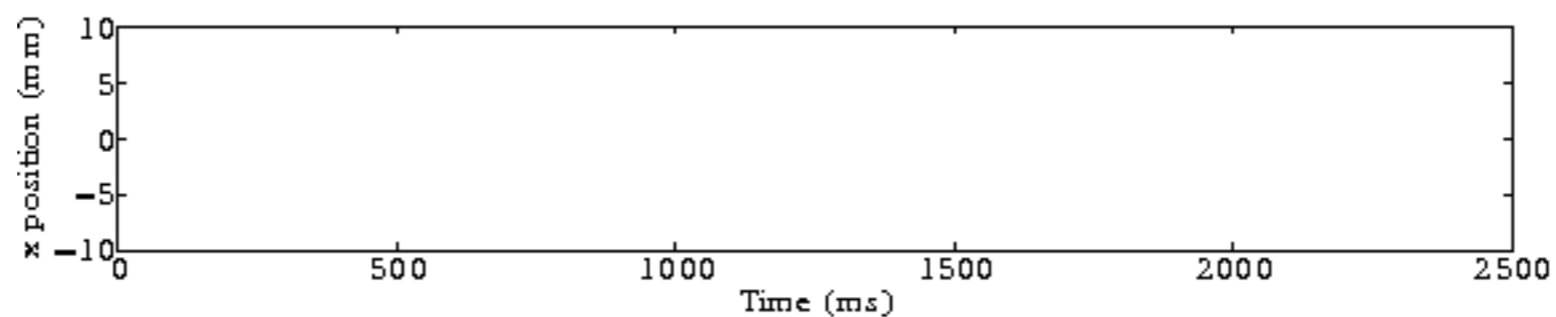
Speech production



Inversion mapping



ANN for the inversion mapping



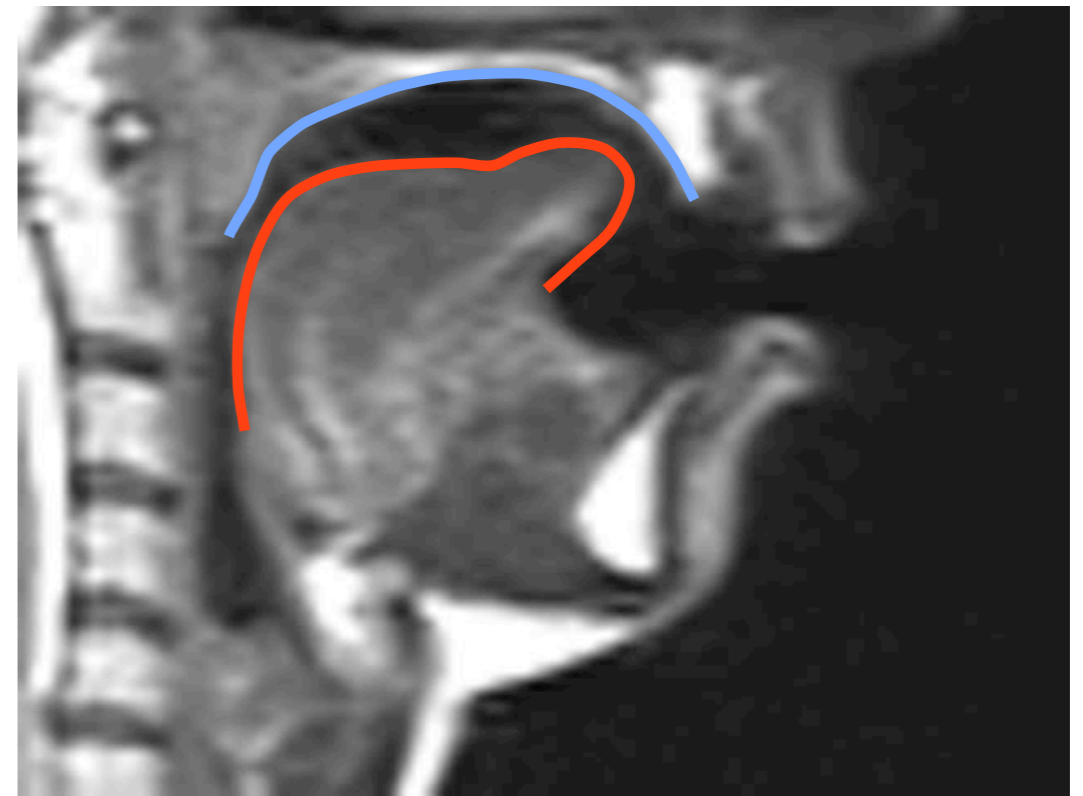
(thanks to Benigno Uría for animation)

Inversion mapping - characteristics

- Interesting modelling problem:
 - non-linear
 - one-to-many mappings (=ill-posed problem)



“oar”



“oar”

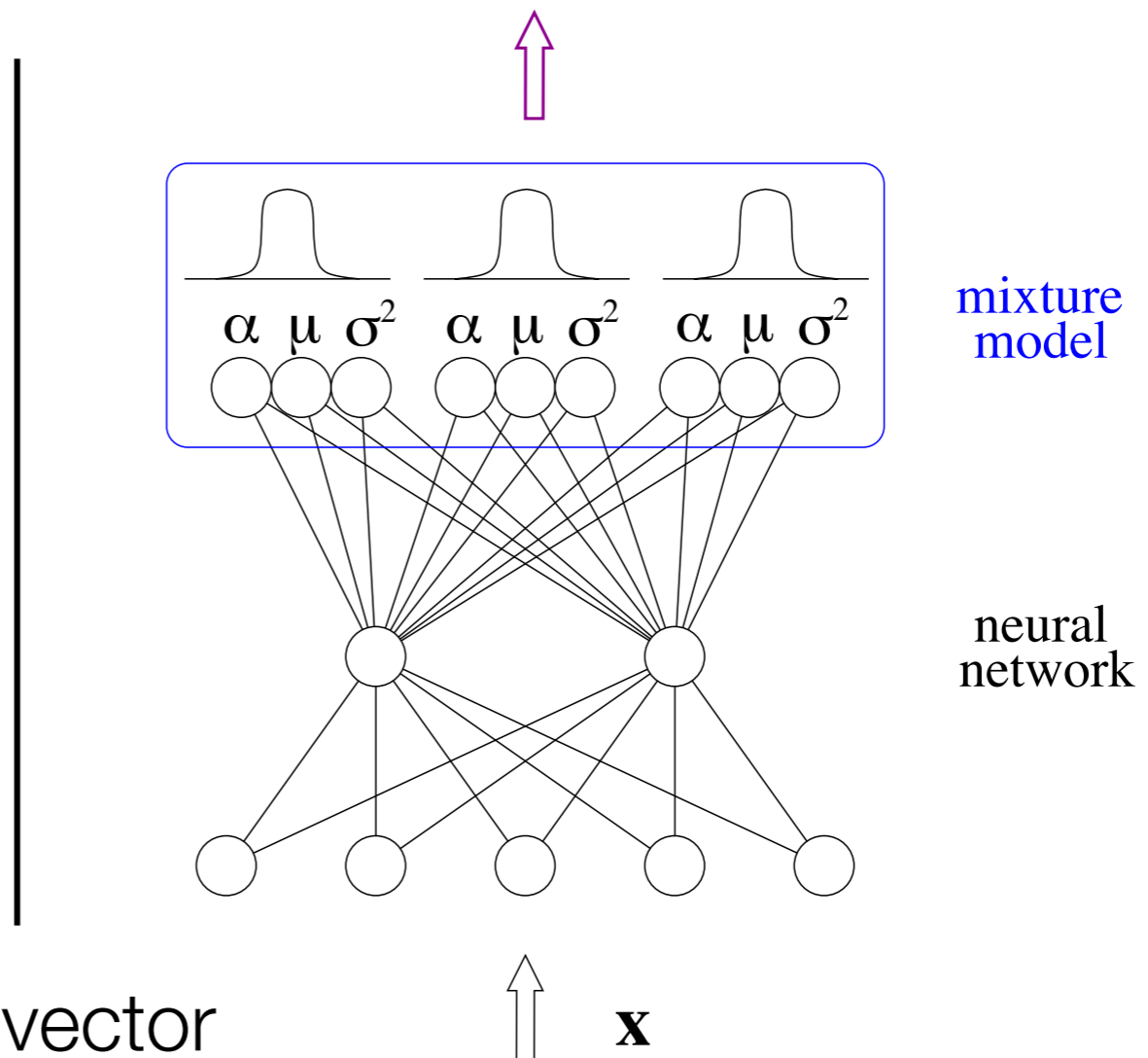
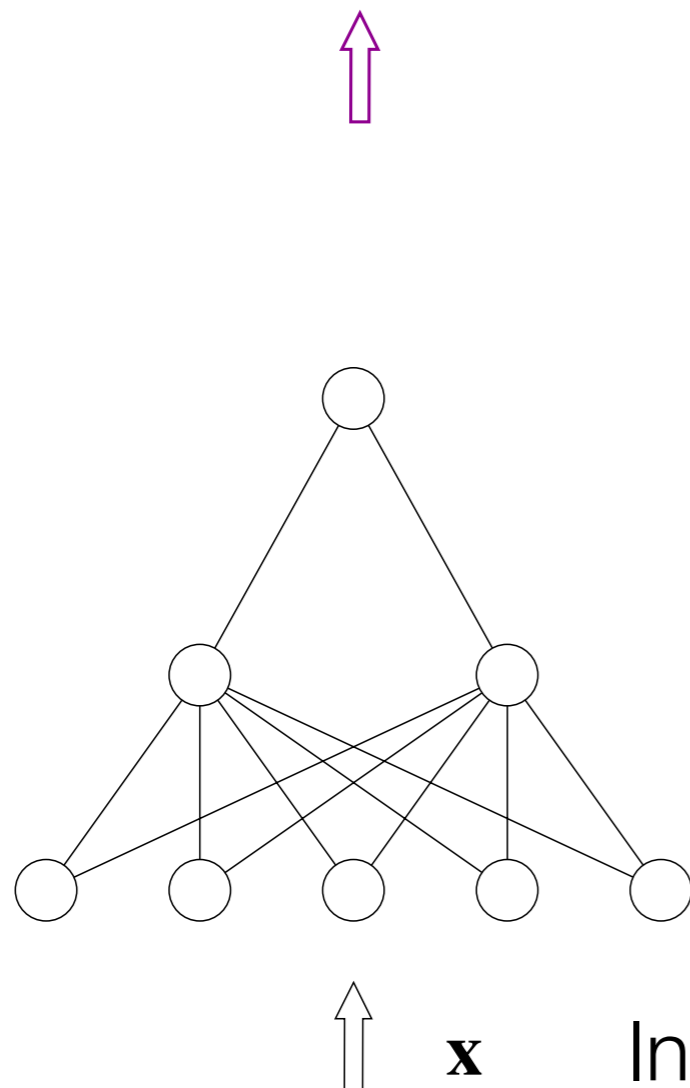
Mixture density network (MDN) suits ill-posed problems

MLP gives
conditional mean

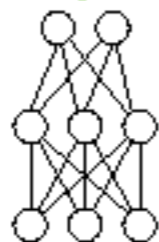
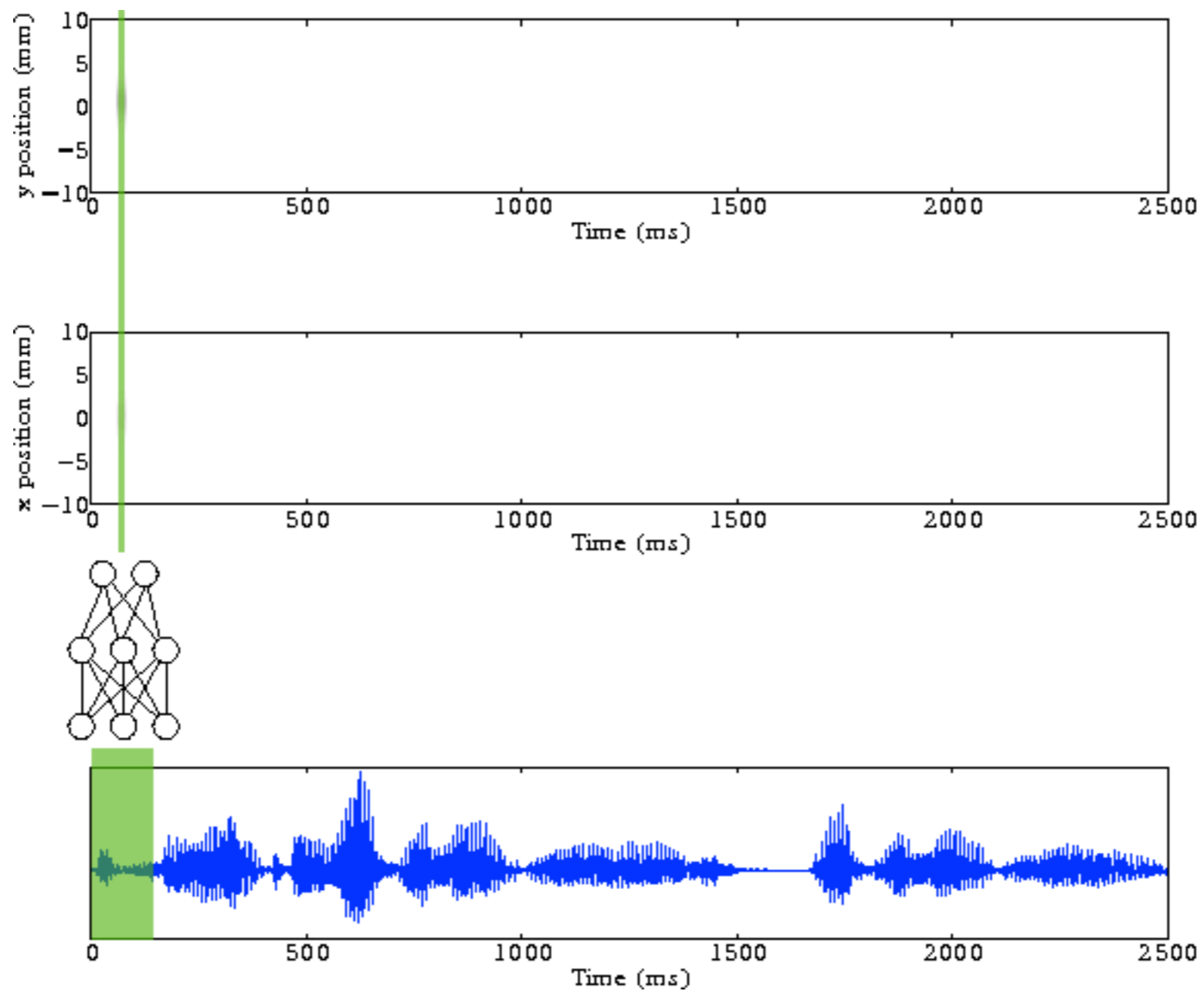
MDN gives
conditional probability density

$E(t|x)$

$p(t|x)$

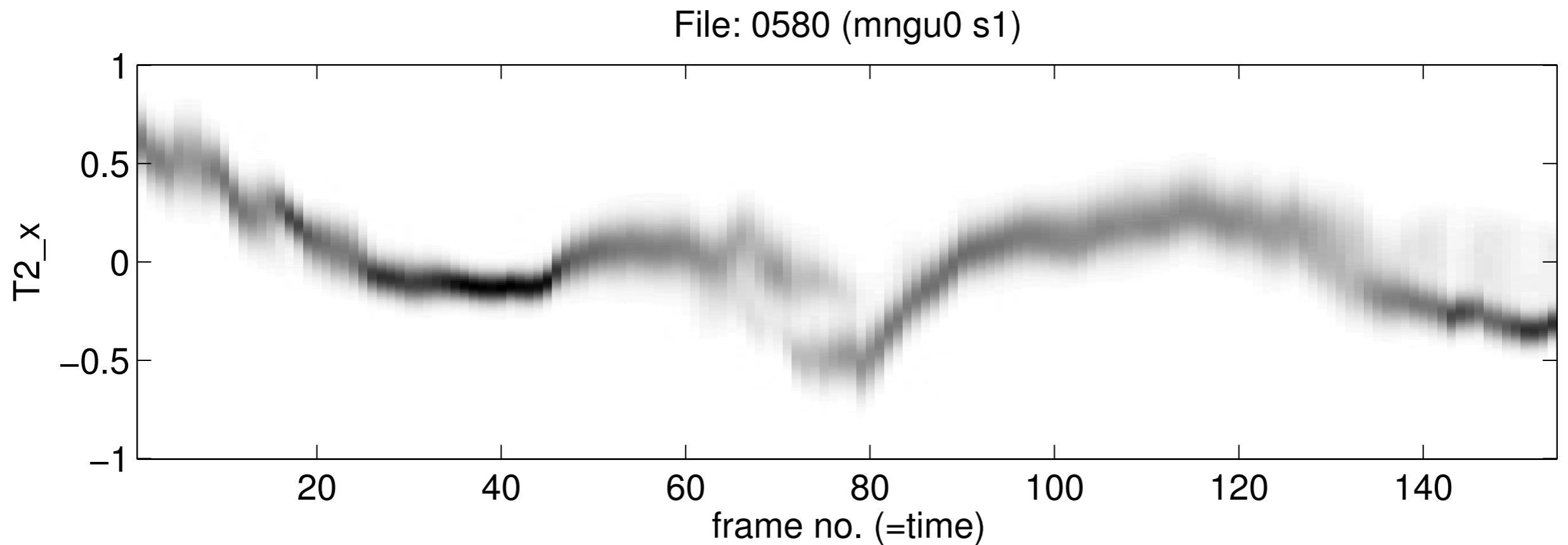


TMDN Inversion summary



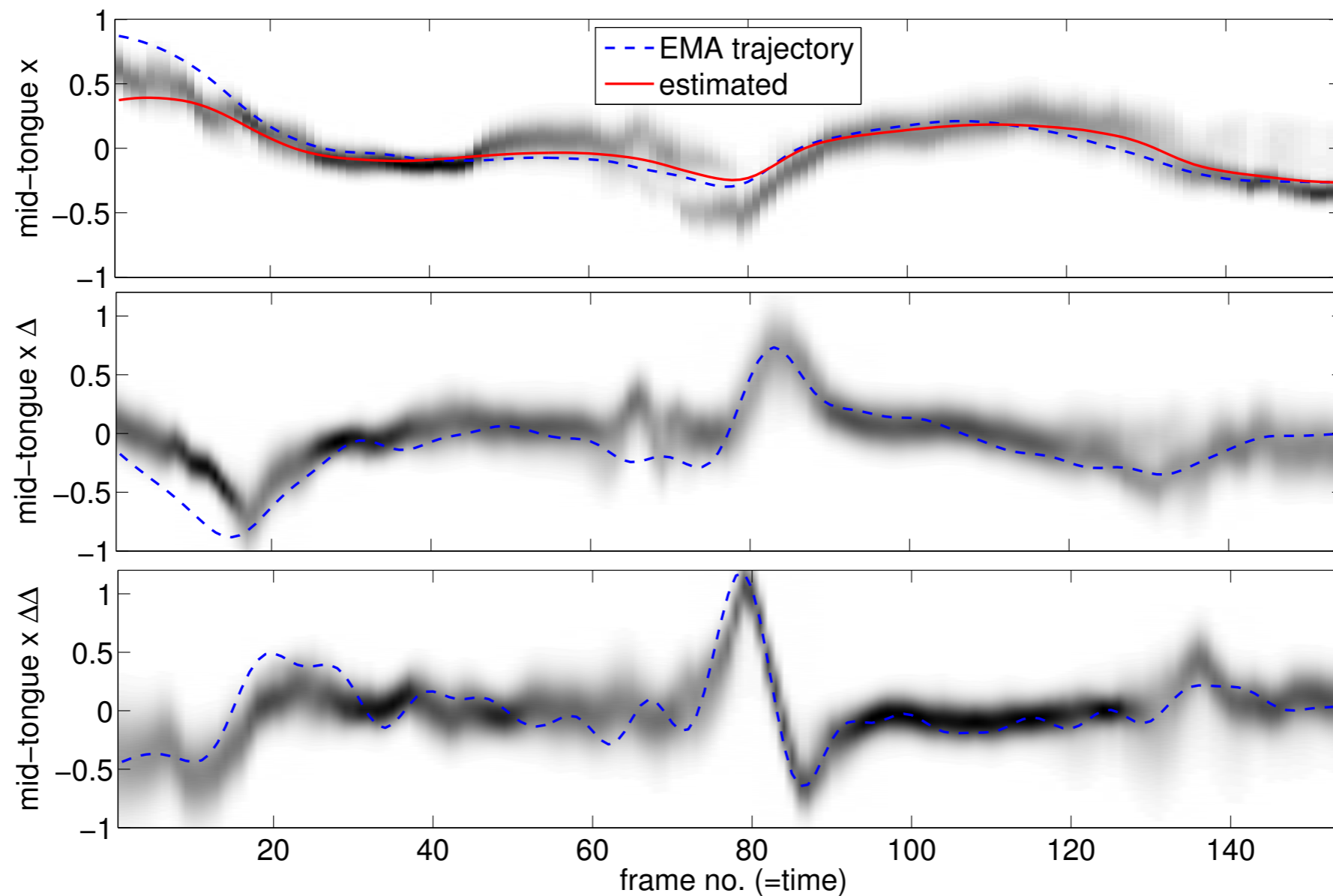
(thanks to Benigno Uría for animation)

MDN output



K. Richmond, S. King, and P. Taylor. Modelling the uncertainty in recovering articulation from acoustics. *Computer Speech and Language*, 17:153–172, 2003.

MDN output - estimating trajectories



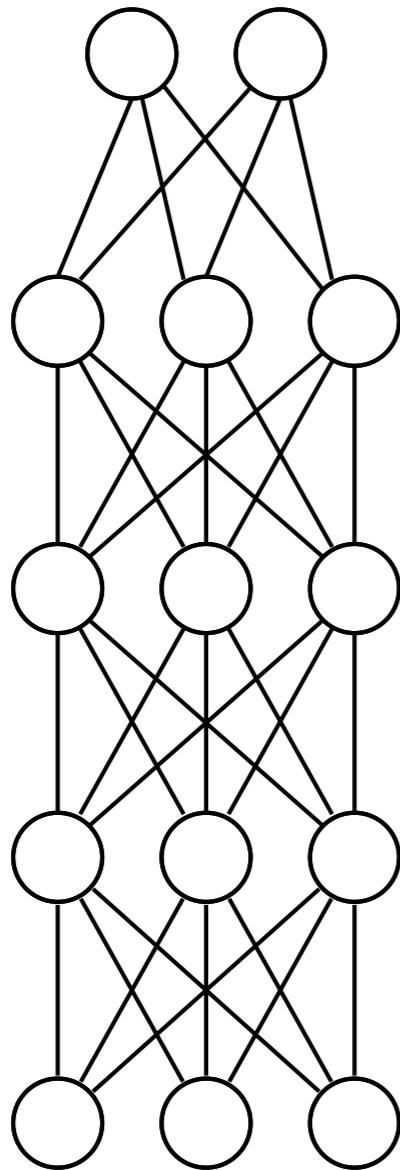
RMS Error
1.37mm

K. Richmond. Trajectory mixture density networks with multiple mixtures for acoustic-articulatory inversion. International Conference on Non-Linear Speech Processing, NOLISP 2007.

RMS Error
0.99mm

K. Richmond. Preliminary inversion mapping results with a new EMA corpus. In Proc. Interspeech, pages 2835–2838, Brighton, UK, September 2009.

Improvements with deeper ANN models

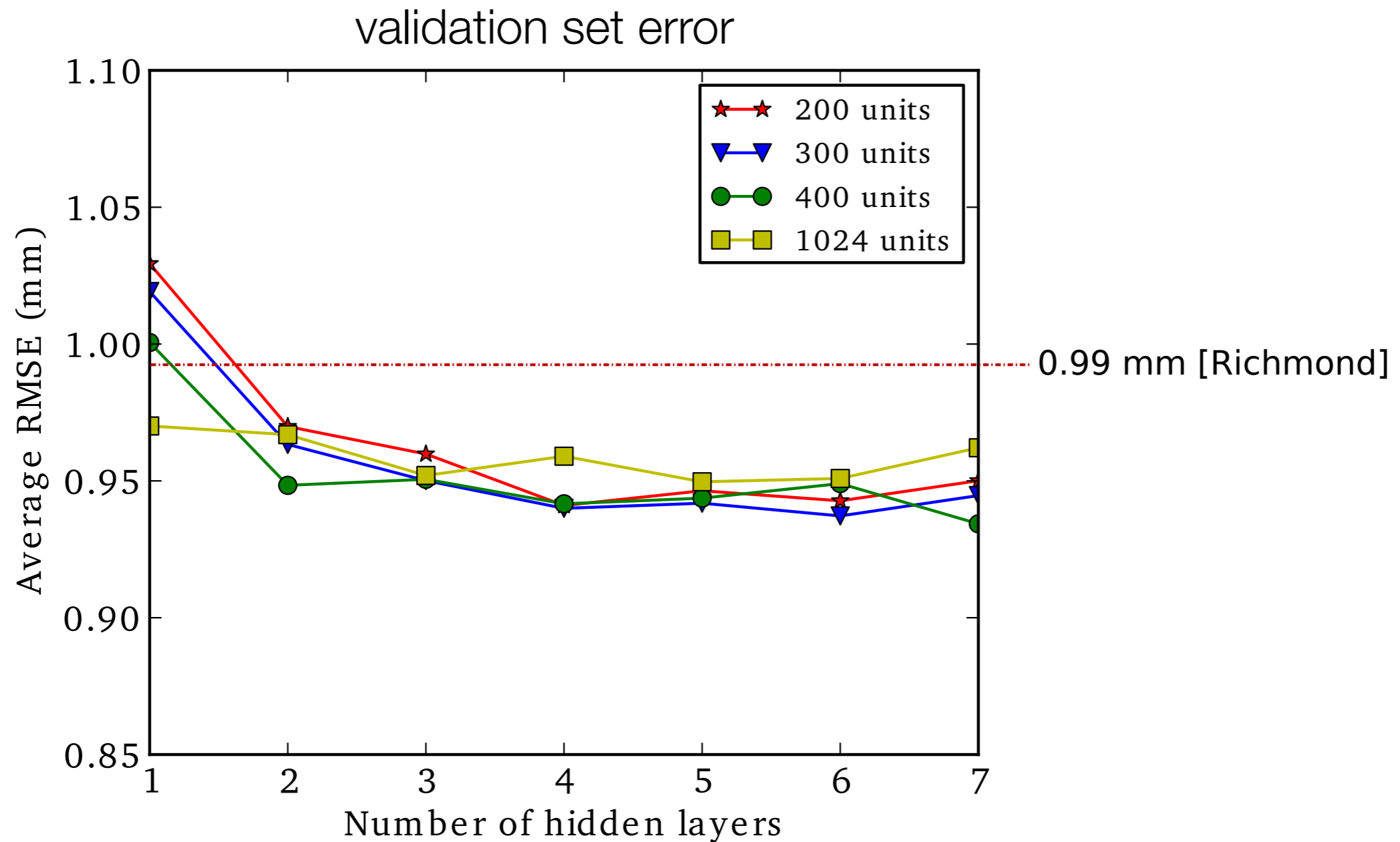


- The target - deep multilayer ANN
- To build this network:
 - stacking RBM pretraining
 - add final layer + optimize
 - fine tune all weights with standard backpropagation

Deep MLP test - experiment settings

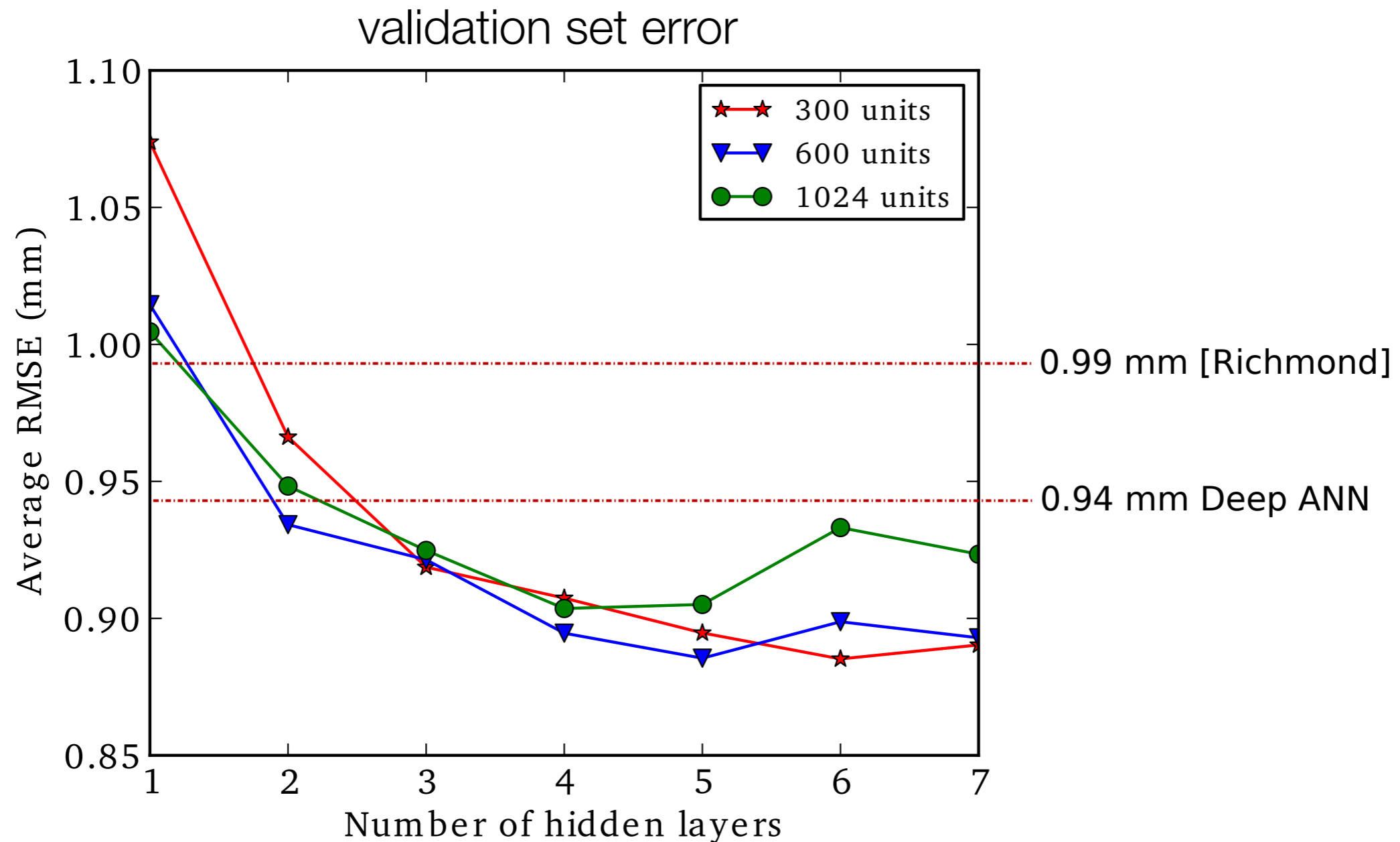
- Input: PLP features 9 frames (100ms)
- Output: 12 linear units (x,y per articulator)
- Sigmoidal hidden units
- 1 to 7 hidden layers
- 200, 300, 400 or 1024 units per hidden layer

Deep MLP results - Error versus depth



test set RMS error = 0.94 mm (mngu0 day1)

Deep TMDN results - Error versus depth



test set RMS error = 0.885 mm (mngu0 day1)

Is this (or any) inversion mapping any good?!

Two central questions:

- Is the articulatory representation adequate?
- What is the best performance we can hope to achieve?

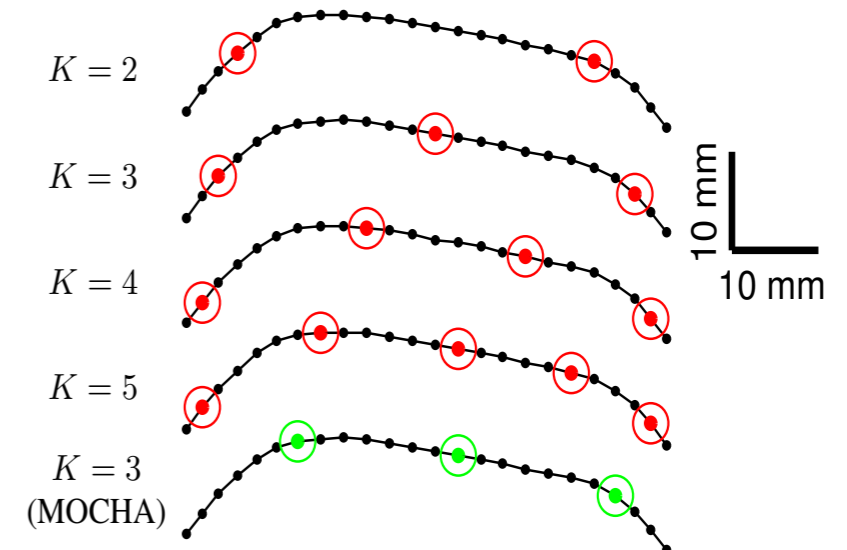
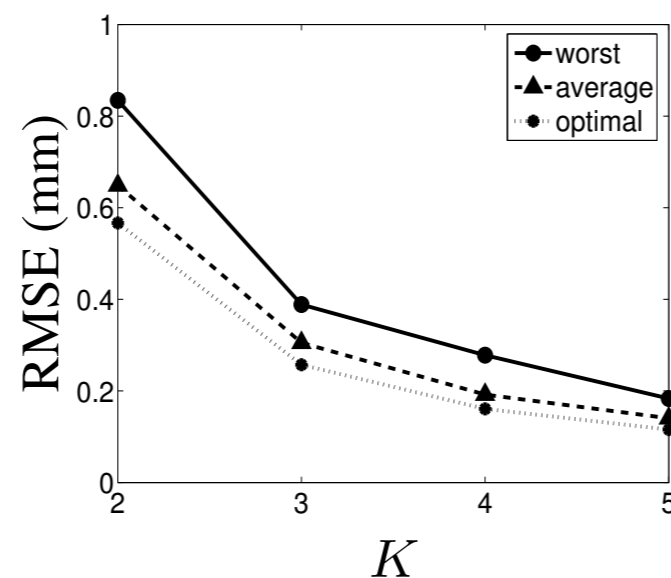
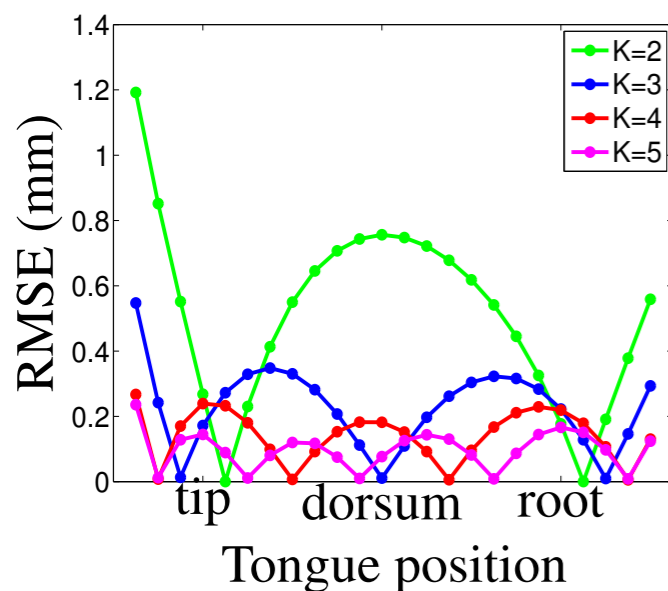
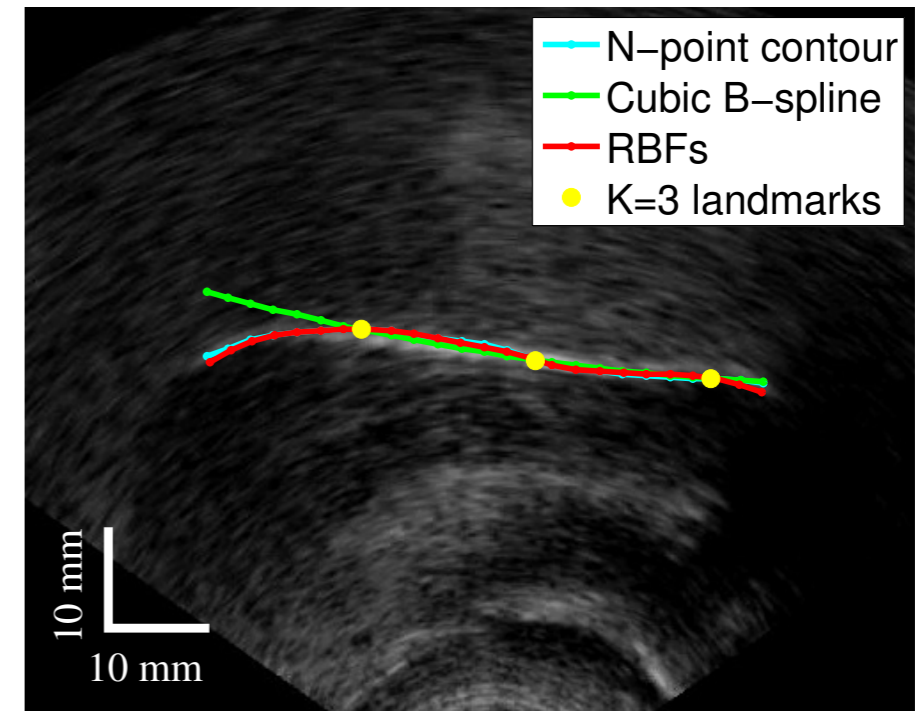
Adequate articulatory representation?

- Specifically, is EMA enough?
- Some indications from:
 - Tongue contour modelling work
 - Synthesis work
 - Articulatory controllable HMM-based synthesis
 - Direct articulatory-to-acoustic mapping

Tongue contour prediction from limited points

C. Qin, M. Carreira-Perpiñán, K. Richmond, A. Wrench, and S. Renals. Predicting tongue shapes from a few landmark locations. In Proc. Interspeech, 2008.

- Database of hand-labelled ultrasound images for training + testing data
- RBF used to predict tongue contour from varying number of points on contour (e.g. EMA locations)
- Also evaluate optimal “EMA” point placement
- For 3 tongue points, average error = 0.3mm

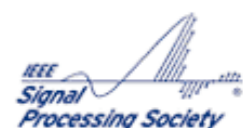
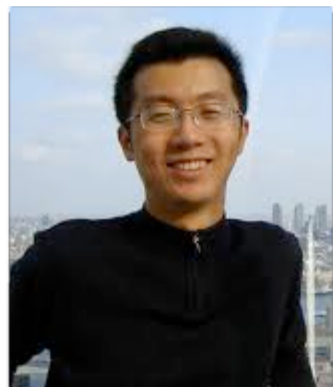


Articulatorily controllable synthesis

- Aim => To change speech produced by hidden Markov model-based speech synthesiser using articulatory controls.

Z. Ling, K. Richmond, J. Yamagishi, and R. Wang. Integrating articulatory features into HMM-based parametric speech synthesis. *IEEE Transactions on Audio, Speech and Language Processing*, 17(6):1171-1185, August 2009.

Z. Ling, K. Richmond, and J. Yamagishi. Articulatory control of HMM-based parametric speech synthesis using feature-space-switched multiple regression. *Audio, Speech, and Language Processing, IEEE Transactions on*, 21(1):207-219, 2013.



Motivation

Hidden Markov model-based synthesis is state-of-the-art.

- Pros:
 - Flexible (parametric rather than concatenative)
 - Trainable (data-driven rather than expert-intensive rules)
 - Adaptable (speaker, style, emotion...)
- Cons:
 - “Black box”
 - Modification requires more data
- So, aim is to gain even more flexible control over synthesis

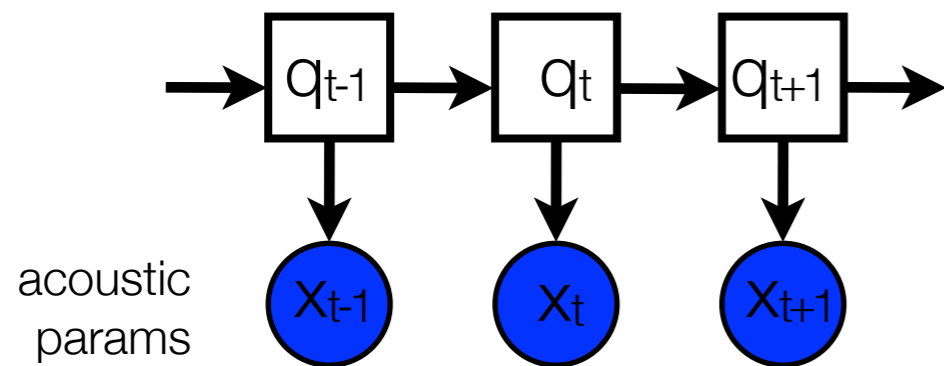


Interpolate:
normal->angry

Introducing articulation into HMM synthesis model

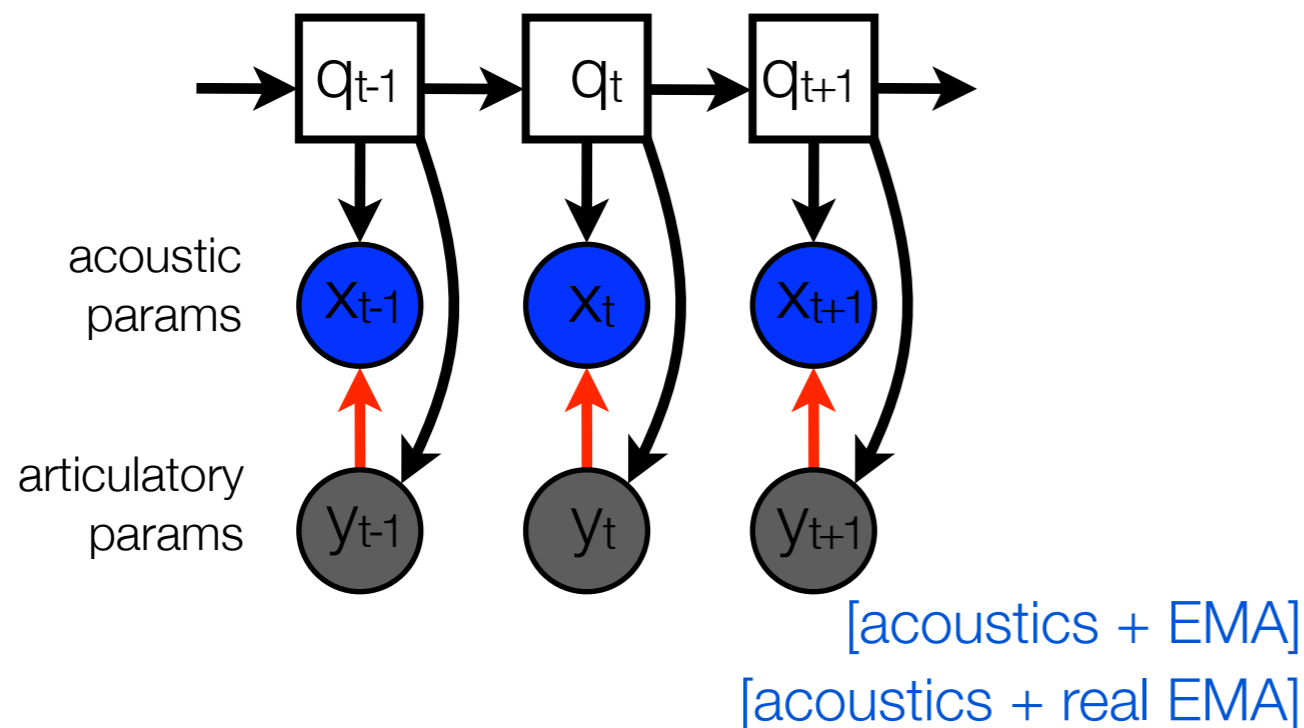
(First attempt)

[acoustic only]



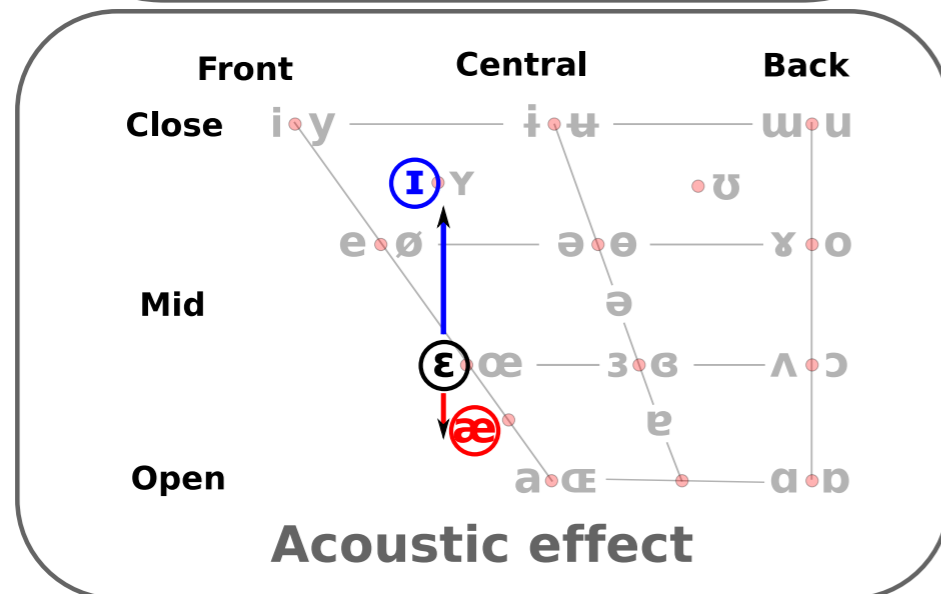
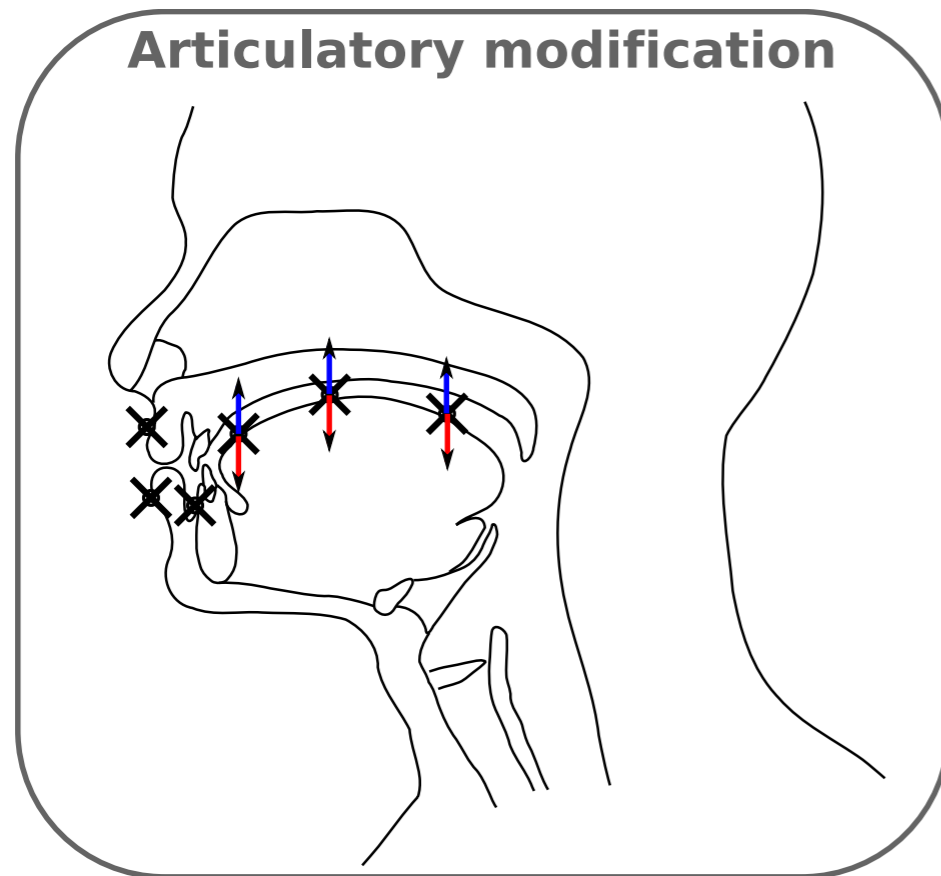
- Model joint distribution of acoustic and articulatory parameters
- Acoustic distribution is *dependent on* articulation
- Dependency = linear transform

$$b_j(\mathbf{x}_t | \mathbf{y}_t) = \mathcal{N}(\mathbf{x}_t; \mathbf{A}_j \mathbf{y}_t + \mu_{x_j}, \Sigma_{x_j})$$



- A_j is (tied) linear transform matrix for state j ...
- ... = **Global piecewise linear mapping**
- No loss of quality
- NOTE: can use arbitrary function to modify $\mathbf{y}_t \Rightarrow$ articulatory control!

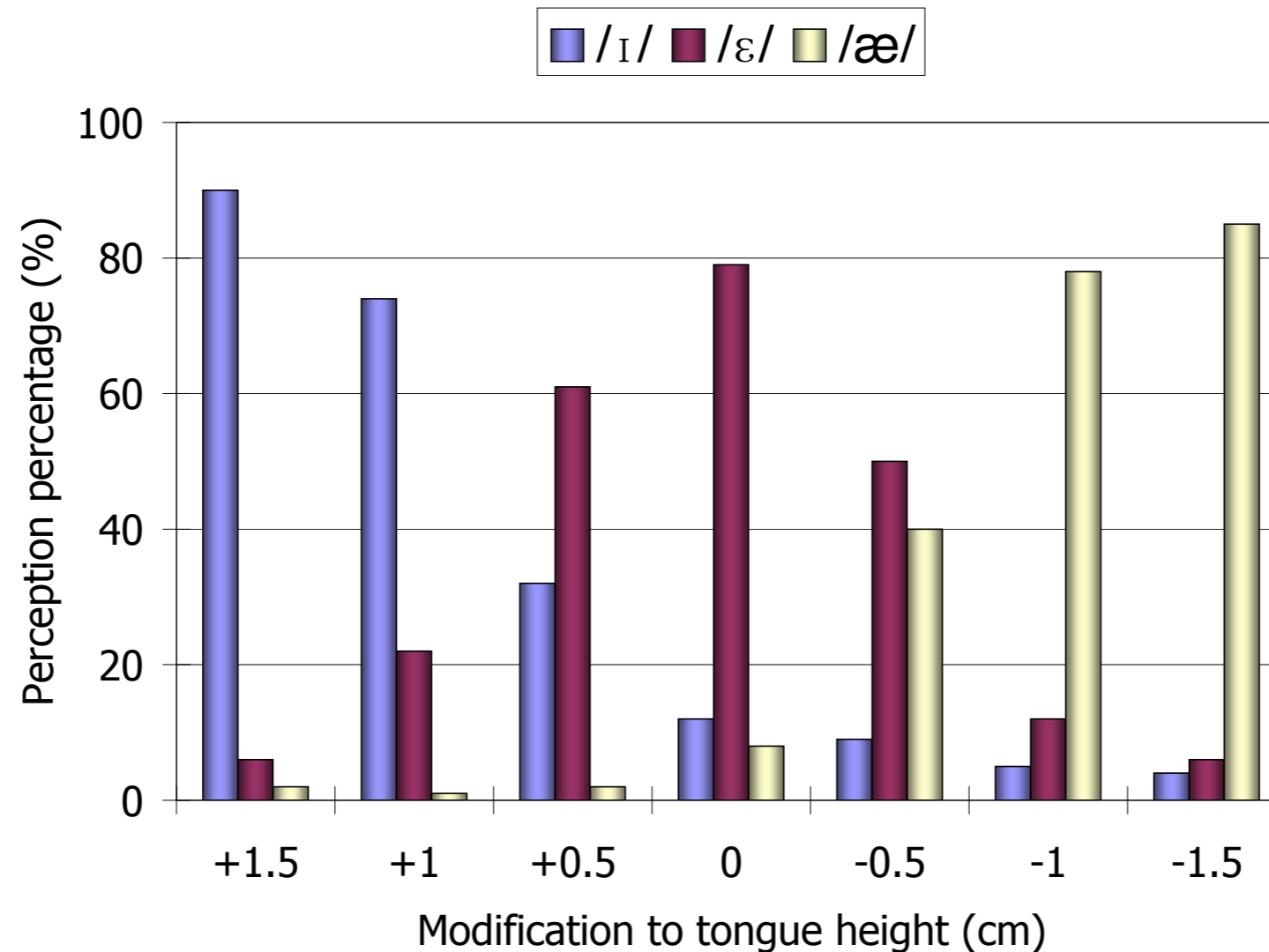
Change model tongue height => change vowel



Raise tongue (cm)	
	+1.5
	+1.0
	+0.5
➔	bet
	-0.5
	-1.0
	-1.5
Lower tongue (cm)	

Perceptual test results

- 20 listeners, lab conditions, results pooled across speakers and words



•Z. Ling, K. Richmond, J. Yamagishi, and R. Wang. Integrating articulatory features into HMM-based parametric speech synthesis. *IEEE Transactions on Audio, Speech and Language Processing*, 17(6): 1171-1185, **2010 IEEE Best Young Author paper award**, August 2009.

Direct articulatory-acoustic mapping

- “Articulatory synthesis” - but data-driven (no physiological model)
- Nonlinear regression from articulation to acoustic synthesis parameters
- Some examples of simple baseline system (MLP mapping)
 - Input = EMA+Gain+F0
 - Output = LSF vocoder parameters
 - Training data = 720 utts of mngu0 day2 EMA



“However, that optimism now seems premature.”

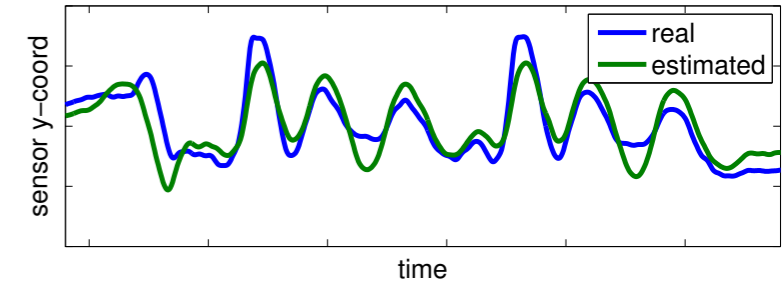


“But this yard.”



“There is a huge amount of data overload.”

How good is an inversion mapping?



Korin Richmond, Zhenhua Ling, Junichi Yamagishi, and Benigno Uría. On the evaluation of inversion mapping performance in the acoustic domain. In Proc. Interspeech, 2013.

Inversion mapping methods standardly evaluated using RMS error and correlation. **Is this good enough?**

- Zero RMSE and perfect correlation will not happen
- Non-uniqueness (c.f. “(non-)critical” articulators ignored)
- No indication how close **optimal inversion** is

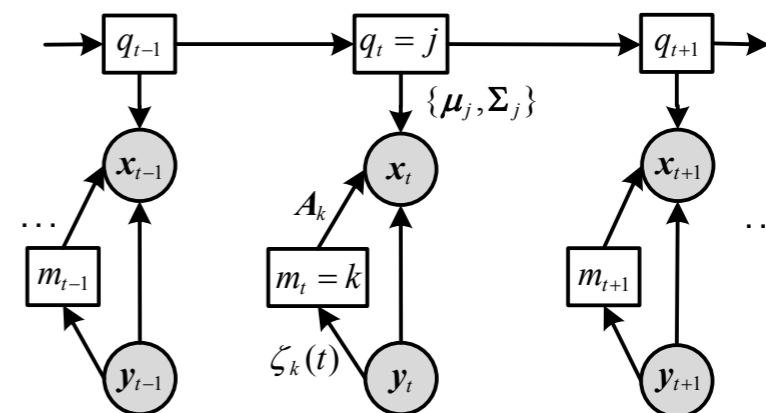
Question: Can “task-based” evaluation add new insight?

- Explore with task = articulatory controlled TTS
- Compare standard articulatory error measures with error calculated in the **acoustic** domain

Articulatory-controlled HMM-based TTS

“Feature-space-switched Multiple Regression HMM” (FSS-MRHMM)

- Spectral distributions (\mathbf{x}_t) depend on state + external articulation (\mathbf{y}_t)
- Separate Single GMM for transform tying (matrices \mathbf{A} below)
- Context feature tailoring



probability for (separate) GMM component $m_t = k$ given \mathbf{y}_t

expanded articulatory vector $[\mathbf{y}_t^T, 1]^T$

$$b_j(\mathbf{x}_t | \mathbf{y}_t) = \sum_{k=1}^M \zeta_k(t) \mathcal{N}(\mathbf{x}_t; \mathbf{A}_k \boldsymbol{\xi}_t + \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$$

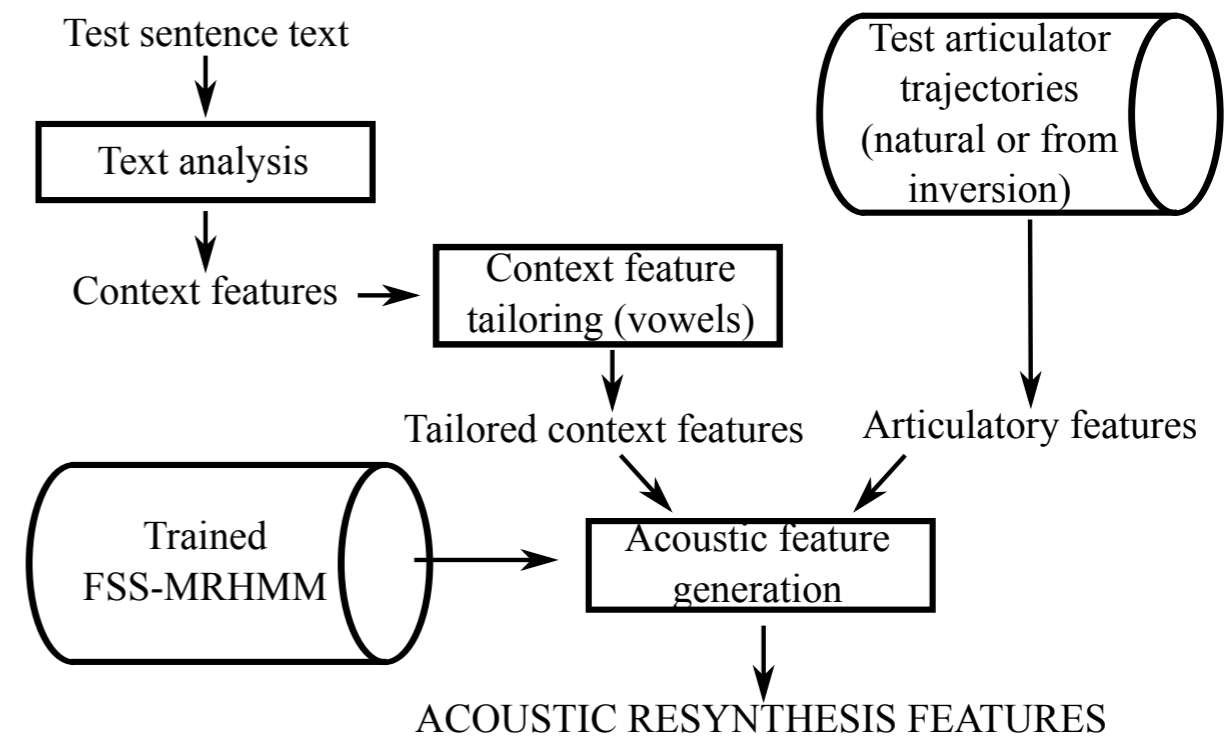
transform matrix for GMM component k

state-dependent acoustic mean and variance

Experiment method

Test a range of inversion methods using:

1. **articulatory** evaluation using standard RMS error and correlation
2. **acoustic** evaluation using an articulator controlled text-to-speech synthesiser:
 - i. *acoustic* RMS error
 - ii. human perceptual test

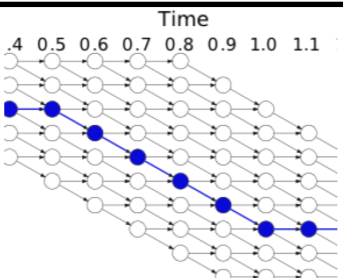
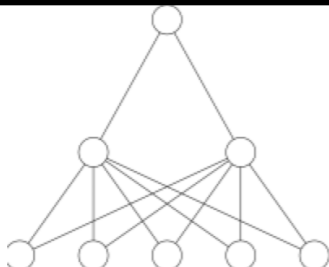
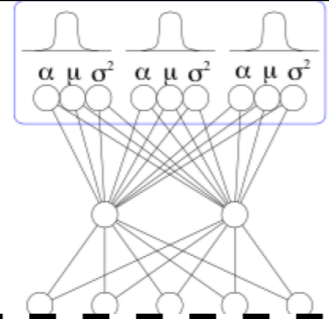
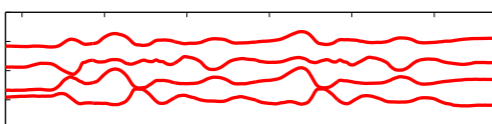


articulatory synthesis flowchart:

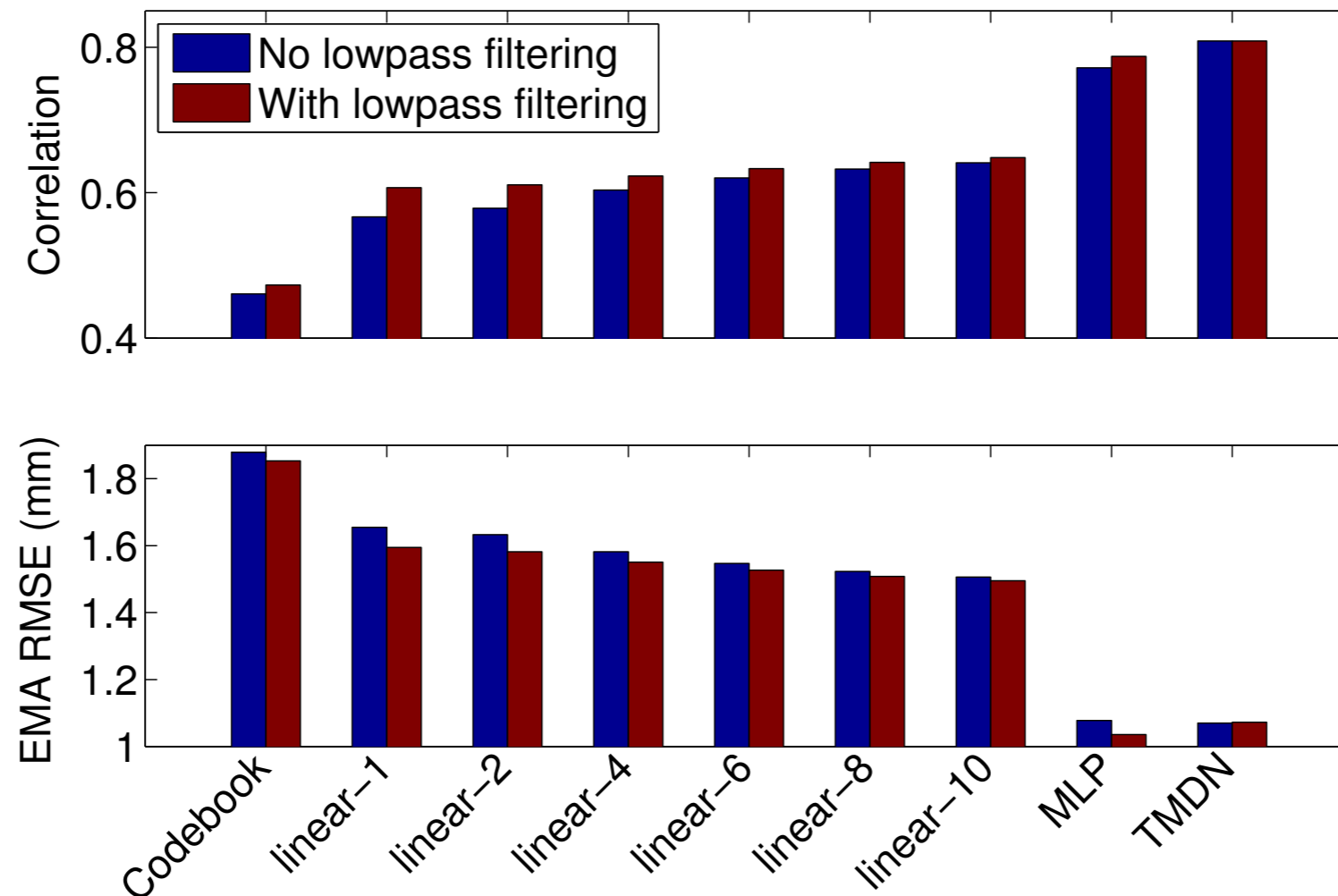
Data processing

- Day 1 EMA set of mngu0 corpus (for TTS + inversion methods)
 - one British male, single session, 1263 prompts total
 - 3D EMA (AG500), 6 coils (lips, jaw, 3 tongue)
- Acoustics -> LSF (STRAIGHT), order 40+gain, 5msec frameshift (= EMA sample rate)
- Dataset sizes (num. utterances): Validation (63), EMA Test (63), Train (1137)
- All data z-score normalised (not for FSS-MRHMM)

4 mapping methods tested (+smoothed versions)

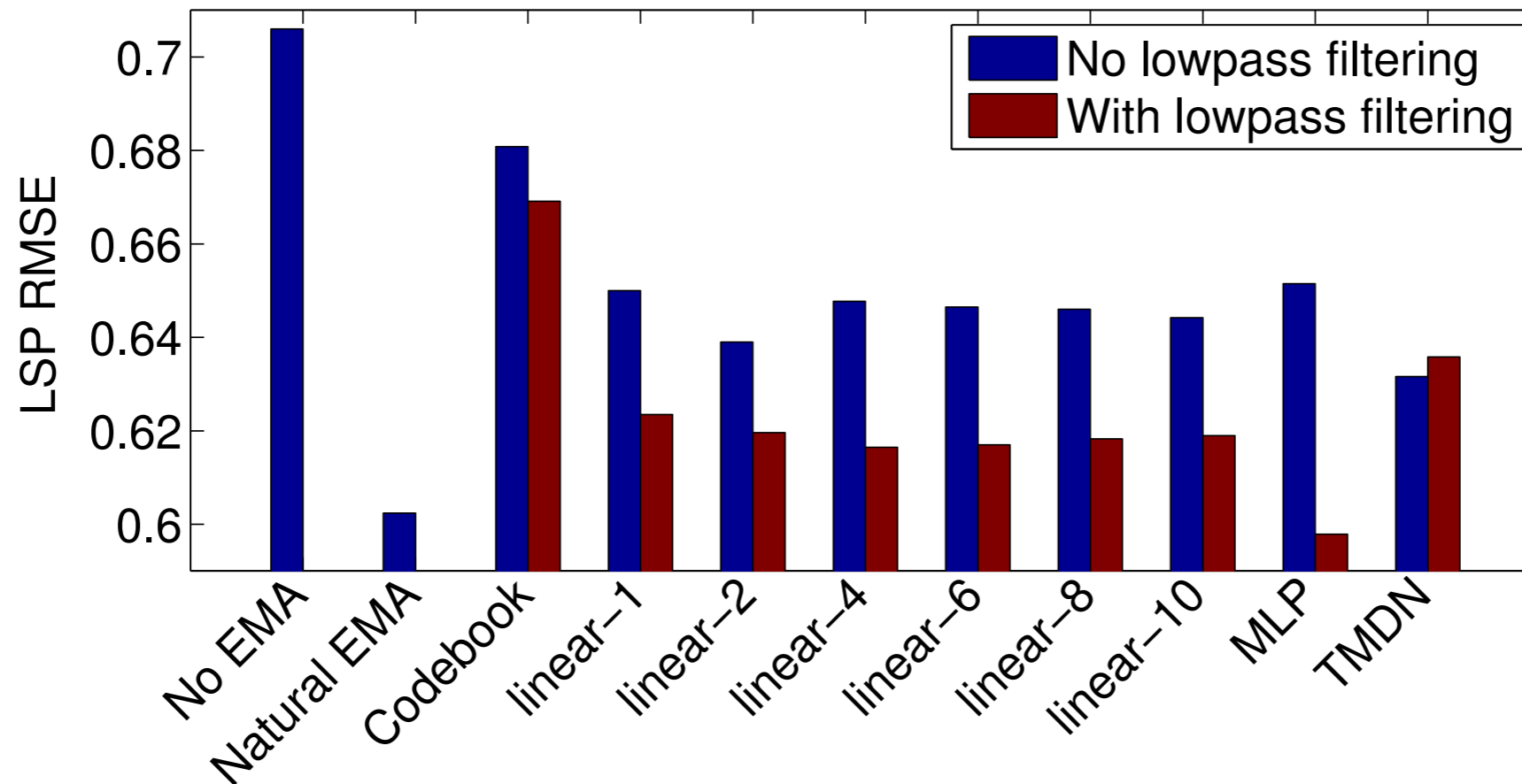
Inversion methods tested	
Linear	Simple linear projection, with 1,2,4,6,8 or 10 acoustic context frames
Codebook	 KD-Tree to find 5000 candidates each frame, then Viterbi search with unweighted Euclidean <i>target</i> and <i>join</i> costs
MLP	 1 per channel: 1 hidden layer, 100 units with tanh activation function, 10 acoustic context frames (alternate frames selected)
TMDN	 1 per channel: 1 hidden layer, 100 units tanh activation function, 10 context frames, [1,2 or 4] GMM components, static/ $\Delta/\Delta\Delta$ PDFs + MLPG
<div style="display: flex; align-items: center;"> <div style="margin-right: 10px;">+ Filter</div>  <div style="margin-left: 10px;"> Output additionally smoothed: 2nd order Butterworth filter, 10Hz lowpass cutoff </div> </div>	

Standard articulatory error measure results



- Codebook < linear < MLP ≤ TMDN
- +Filter improves all results, except TMDN
- Reasonable spread of performance
- NOTE: This is **not** a fair comparison of methods!

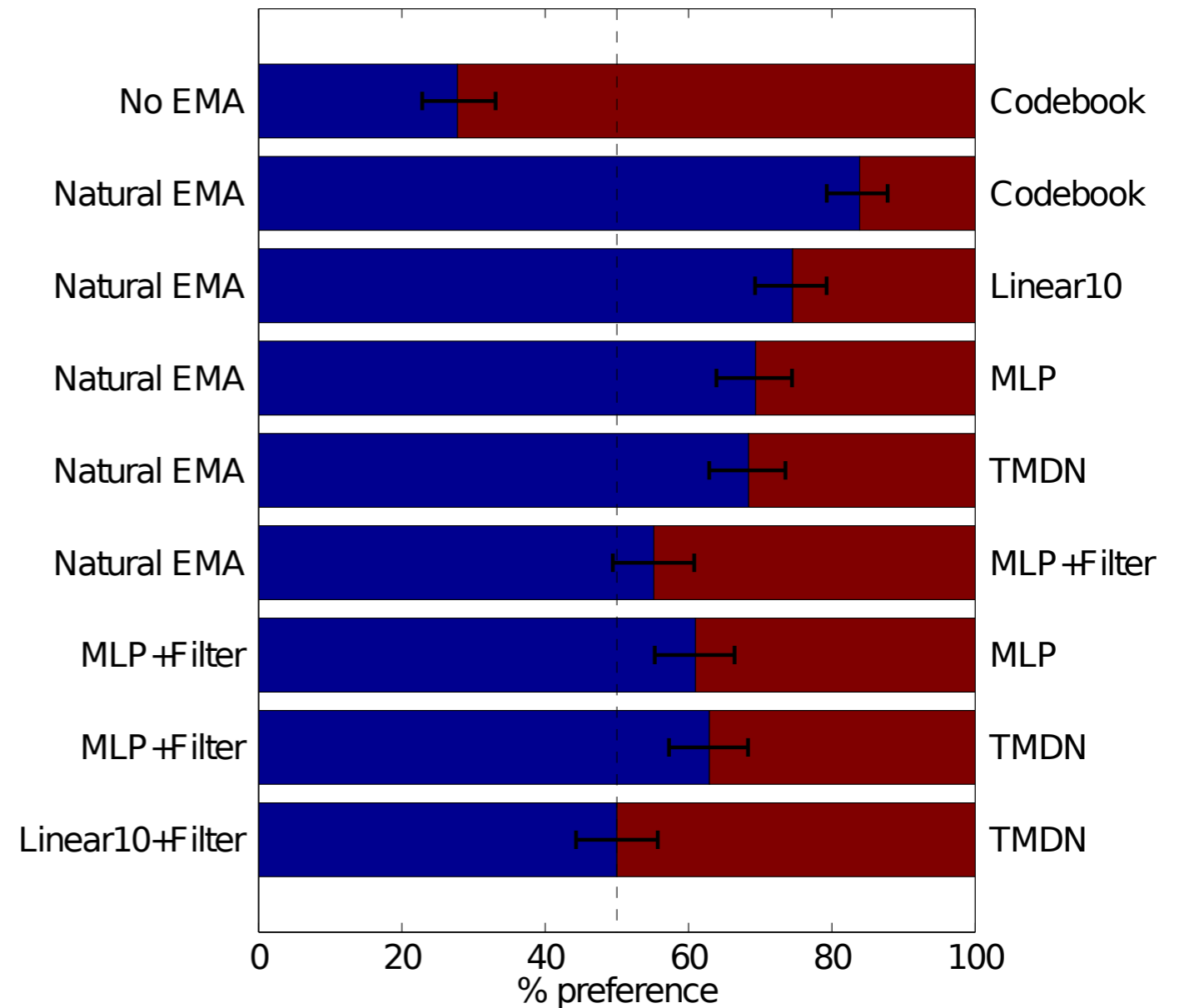
Acoustic evaluation results: LSF RMSE



- Synthesise 63 test utts, calculate LSF error
- Perceptually weighted Euclidean distance
- Some interesting differences
 - MLP+Filter much better than TMDN
 - TMDN appears worse than linear+Filter

Acoustic evaluation: listening test results

- 30 paid native British English listeners, lab conditions
- 9 preference tests, each with 10 pairs of stimuli
- Results generally like LSF RMSE results
- Some LSF RMSE differences are imperceptible
- MLP+Filter \cong Natural EMA !!



No EMA < codebook < [MLP,linear10, TMDN] < Natural EMA

Conclusions from this study

Two questions addressed:

1. Can acoustic task-based evaluation give useful info?

- YES!
- Interesting differences from standard articulatory RMSE and correlation

2. Can we get insight into “optimal inversion” performance?

- MLP+LPFiltering performed as well as natural EMA...
 - ...BUT we cannot yet claim this is “**optimal** inversion”
- Simply, sufficient inversion for this task

Summary of talk

On the inversion mapping:

- Shown ANNs are a viable model for inversion mapping
- Deep ANN models give the state-of-the-art performance (probably)
- Given some indications of level of “information” in EMA data
 - This is far from conclusive... open question up for debate and further study
- Raised question of “**optimal inversion**”
 - Some interesting results - RMSE and correlation don't give full picture
 - Found invertedEMA = naturalEMA, but too early to judge if this is “optimal”

Thanks for listening!