Paper Introduction

~ Modelling the Uncertainty in Recovering Articulation from Acoustics ~

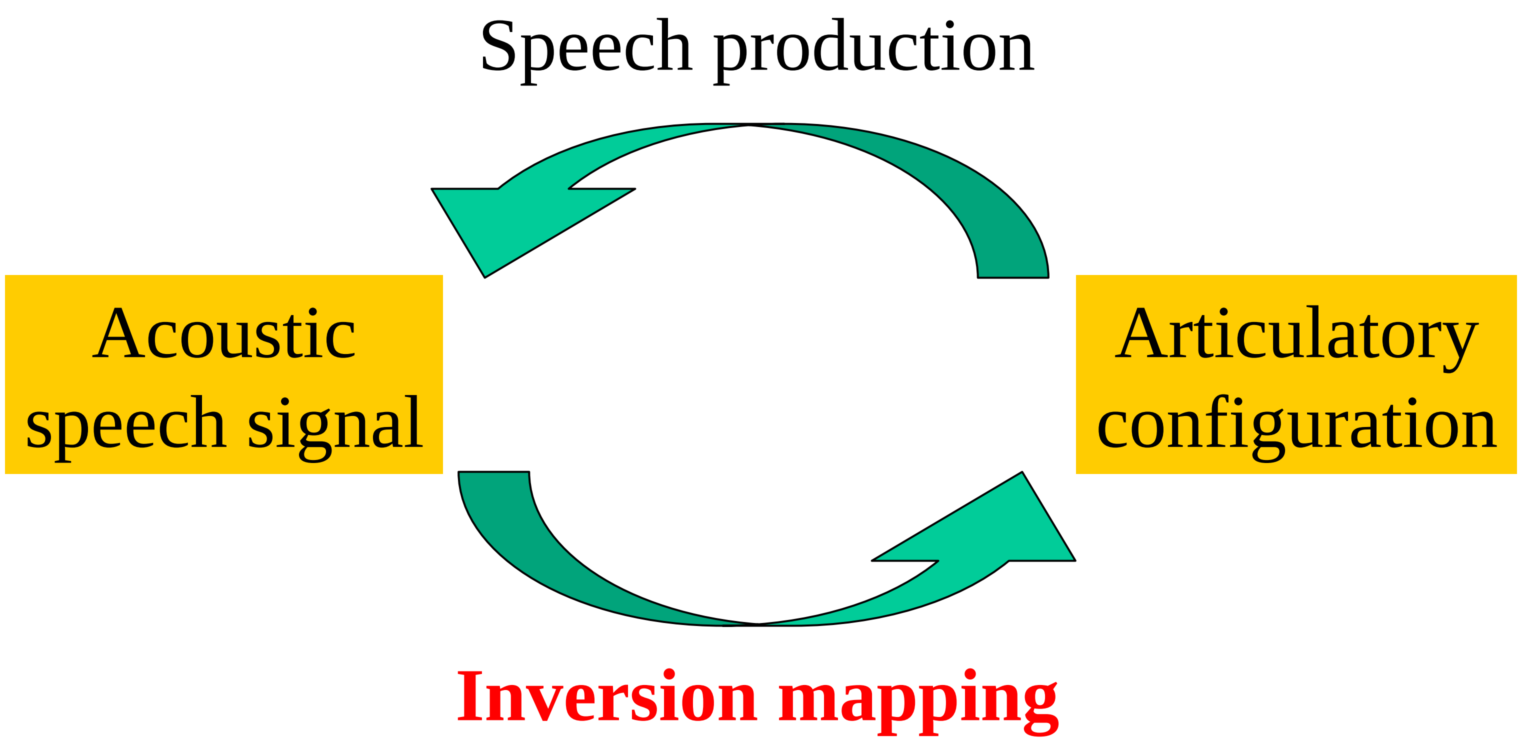
Korin Richmond, Simon King, and Paul Taylor

Tomoki Toda

November 6, 2003

Problem Addressed in This Paper

- Modelling the acoustic-to-articulatory mapping



✓Acoustic-articulatory data
✓Statistical model

Contents

- Inversion mapping
- Acoustic-articulatory data (MOCHA)
- Mapping with multilayer perceptron (MLP)
- Mapping with mixture density network (MDN)
- Comparing MLP with MDN

Inversion Mapping

- Mapping from acoustic speech signal to articulatory configuration
  - Ill-posed problem (non-unique solution)
- Applications
  - Speech coding
  - Speech training
  - Speech recognition
  - Speech synthesis

⋮

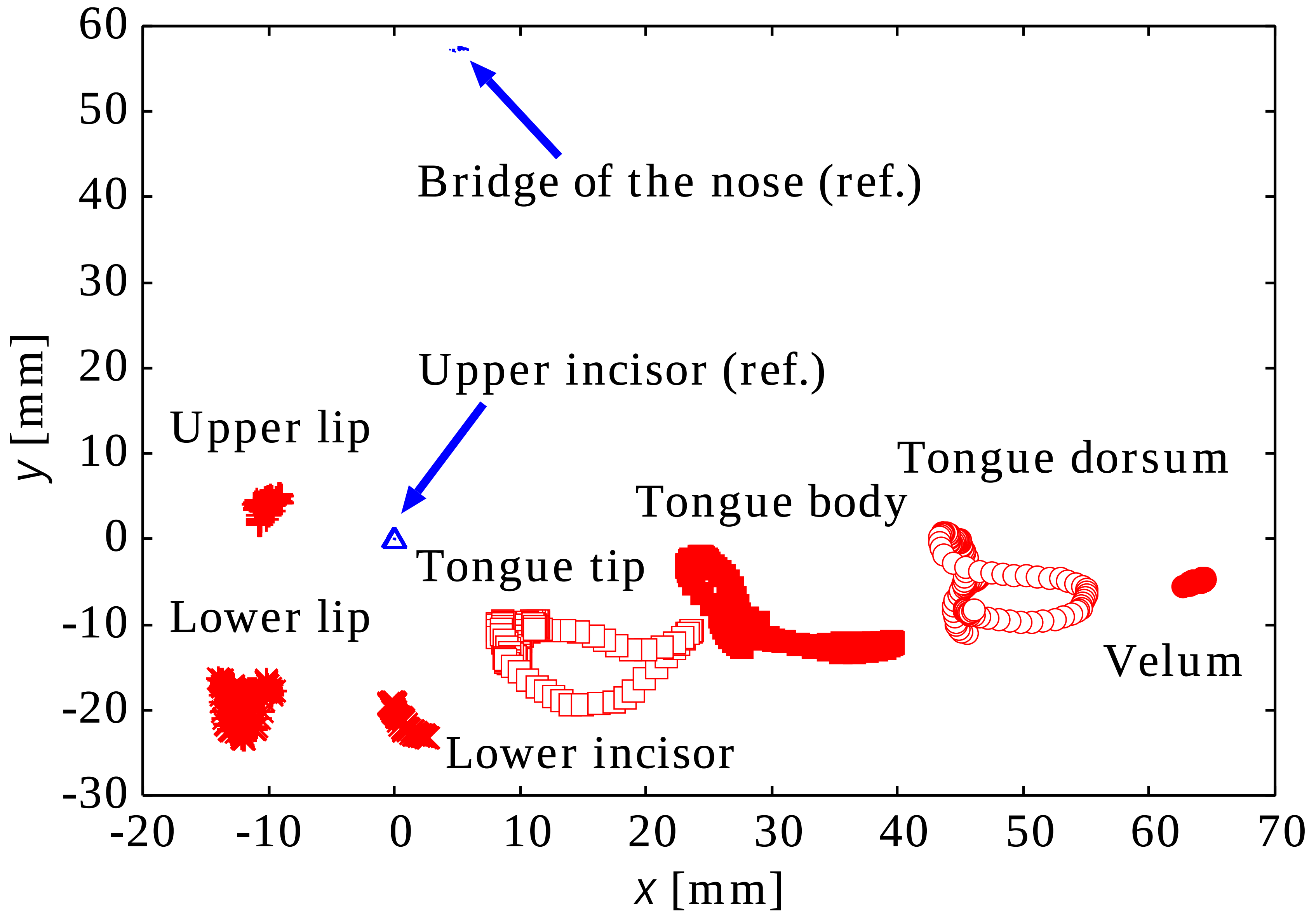
MOCHA database

- Multichannel articulatory database (MOCHA)
  - Queen Margaret University College
- Four data streams
  - Acoustic waveform (16 kHz, 16 bit)
  - Laryngograph (16 kHz, 16 bit)
  - Electropalatograph
  - Electromagnetic articulograph (EMA)
- 460 British TIMIT sentences
- 40 speakers
  - Available: 2 speakers, male and female

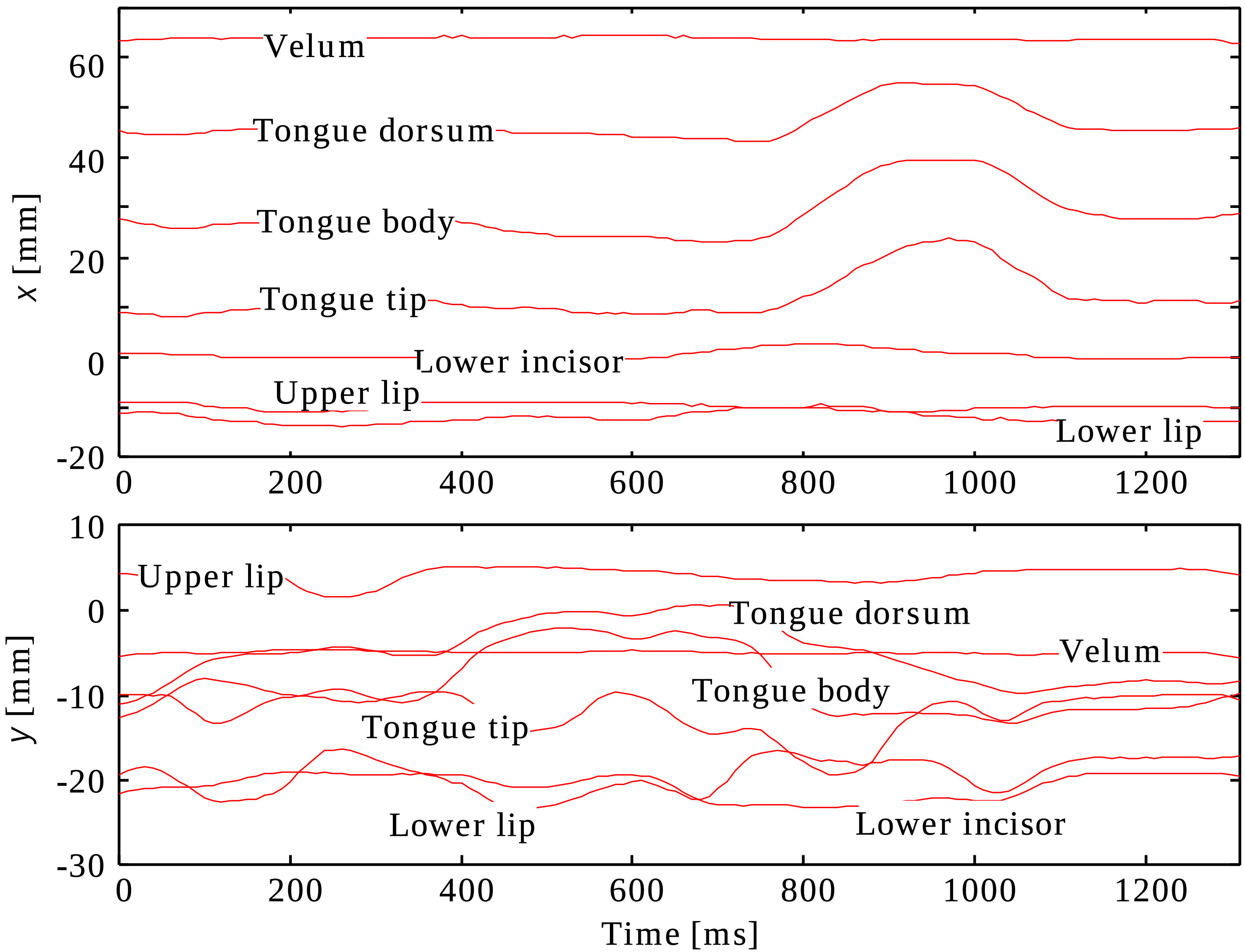
EMA: Electromagnetic Articulograph

- Sampling the movement of receiver coils attached to the articulators
  - 9 points (2 reference points)
    - Top lip, bottom lip, bottom incisor, tongue tip, tongue body, tongue dorsum, velum, (bridge of the nose, upper incisor)
  - x- and y-coordinates in the midsagittal plane
  - 14 channels
  - 500 Hz

Samples: 2-D plot

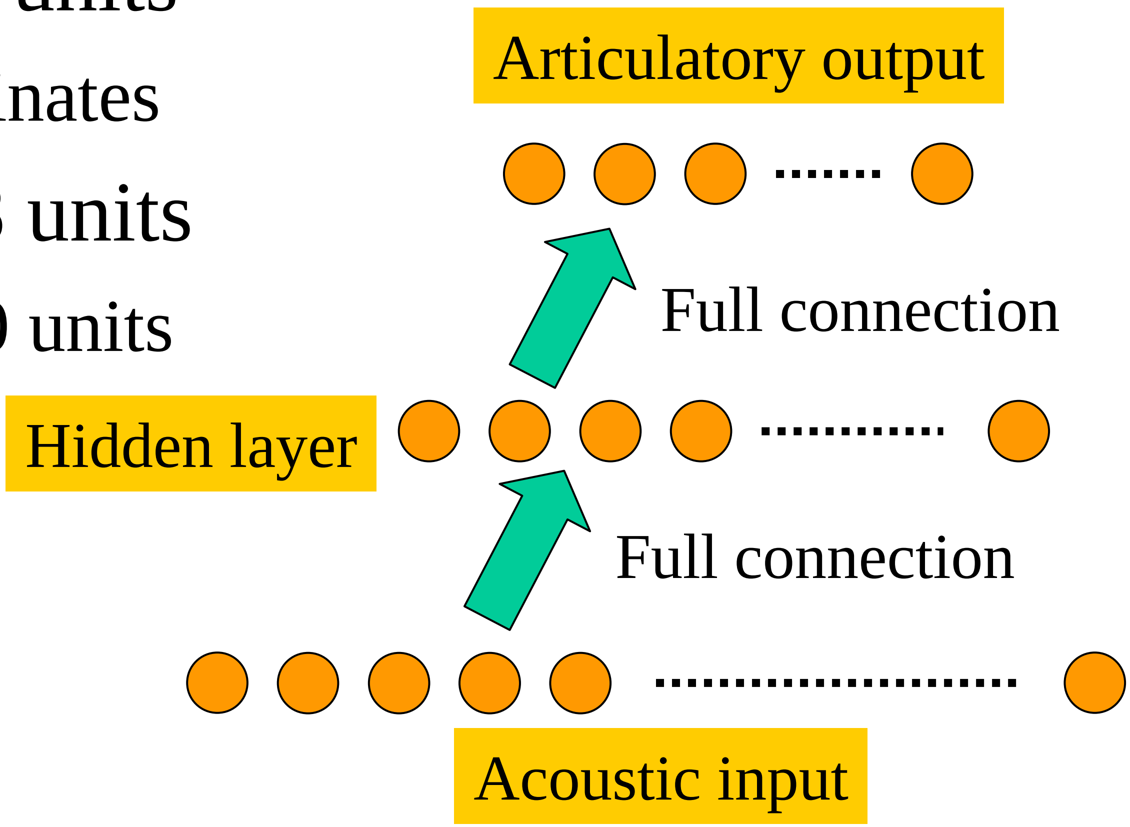


Samples: Time sequence



MLP: Multilayer Perceptron

- Input layer: 400 units
  - 20 frames of 20 filter bank coefficients
- Output layer: 14 units
  - 7 x- and y-coodinates
- Hidden layer: 38 units
  - Pruning from 50 units



Features

- Acoustic feature
  - 20 mel-scale filterbank coefficients
    - 20 ms hamming window, 10ms shift
  - Normalization: 1/(4 σ), 95% interval [0.0, 1.0]
- Articulatory feature
  - Lessening the effect of noise caused by measurement error in EMA machine
    - 10ms shift
  - Normalization: 1/(5 σ), 95% interval [0.1, 0.9]
- Removing silence frames
- Data set
  - For training: 368 sentences
  - For validation: 46 sentences
  - For test: 46 sentences

Weight Optimization

- Error function

$$E = \sum_N \sum_K (y^k(\boldsymbol{x}_n; \boldsymbol{w}) - t_n^k)^2$$

$$\Rightarrow y^k(\boldsymbol{x}_n; \boldsymbol{w}^*) = \langle t_n^k \mid \boldsymbol{x} \rangle$$

w* at the minimum of E

- Gradient descent training

$$\boldsymbol{w}^{(i+1)} = \boldsymbol{w}^{(i)} + \Delta \boldsymbol{w}^{(i)}$$

$$\Delta \boldsymbol{w}^{(i)} = -\eta \frac{\partial}{\partial \boldsymbol{w}} E^{(i)}$$

  – Scaled Conjugate Gradient (SCG)
    - Using not only first derivatives but also second derivatives

Experimental Evaluation

- Evaluation measures
  - Root mean square (RMS) error
    - Overall distance between two trajectories
  - Correlatin coefficient
    - Similarity of shape and synchrony of two trajectories
- Results
  - P. 7, Table 1
  - Average of RMS error: 1.62 mm
  - Estimated trajectories, p. 9, Figure 2

Shortcomings of MLP Mapping

- Discontinuity in estimated trajectories
  - Articulators move slowly and smoothly

  ➡ Low-pass filtering

  ✓Channel specific cutoff frequency

- Insufficient to model one-to-many mappings
  - Using context windows

  ➡ Effective but insufficient

Some limitations of using the sum-of-squares error function

MDN: Mixture Density Network

- Output: conditional probability density function

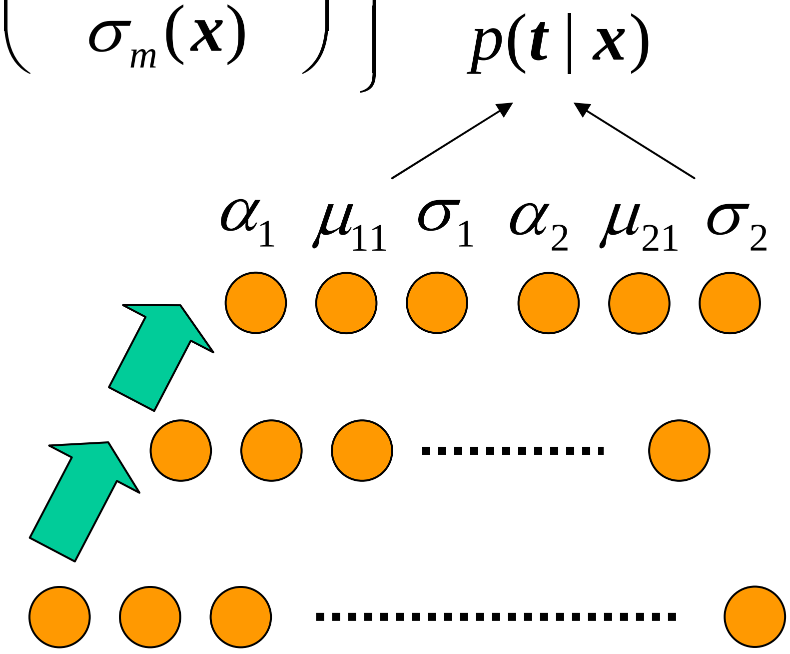
$$p(\boldsymbol{t}\,|\,\boldsymbol{x}) = \sum_{M} \alpha_m(\boldsymbol{x})\phi_m(\boldsymbol{t}\,|\,\boldsymbol{x})$$

- Kernel: Gaussian function

$$\phi_m(\boldsymbol{t}\,|\,\boldsymbol{x}) = \frac{1}{(2\pi)^{K/2}\sigma_m^K(\boldsymbol{x})} \exp\left\{ -\frac{1}{2}\sum_{K}\left(\frac{t^k - \mu_m^k(\boldsymbol{x})}{\sigma_m(\boldsymbol{x})} \right)^2 \right\}$$

  - Weight $\alpha_m(\boldsymbol{x}) = \dfrac{\exp(z_m^{(\alpha)})}{\sum_j^M \exp(z_j^{(\alpha)})}$
  - Mean $\mu_m^k(\boldsymbol{x}) = z_m^{k(\mu)}$
  - Spherical covariance

$$\sigma_m(\boldsymbol{x}) = \exp(z_m^{(\sigma)})$$



Weight Optimization in MDN

- Error function

$$E = -\sum_{N} \ln\left\{ \sum_{M} \alpha_m(\boldsymbol{x}_n)\phi_m(\boldsymbol{t}_n \mid \boldsymbol{x}_n) \right\}$$

$$\pi_m(\boldsymbol{x}_n, \boldsymbol{t}_n) = \frac{\alpha_m(\boldsymbol{x}_n)\phi_m(\boldsymbol{t}_n \mid \boldsymbol{x}_n)}{\sum_{j=1}^{M} \alpha_j(\boldsymbol{x}_n)\phi_j(\boldsymbol{t}_n \mid \boldsymbol{x}_n)}$$

$$\frac{\partial E_n}{\partial z_m^{(\alpha)}} = \alpha_m(\boldsymbol{x}_n) - \pi_m(\boldsymbol{x}_n, \boldsymbol{t}_n)$$

$$\frac{\partial E_n}{\partial z_m^{k(\mu)}} = \pi_m(\boldsymbol{x}_n, \boldsymbol{t}_n)\left\{ \frac{\mu_m^k(\boldsymbol{x}_n) - t_n^k}{\sigma_m^2(\boldsymbol{x}_n)} \right\}$$

$$\frac{\partial E_n}{\partial z_k^{(\sigma)}} = -\pi_m(\boldsymbol{x}_n, \boldsymbol{t}_n)\sum_{K} \left\{ \left(\frac{t_n^k - \mu_m^k(\boldsymbol{x}_n)}{\sigma_m(\boldsymbol{x}_n)} \right)^2 - 1 \right\}$$

- K-means based initialization
  - Unconditional density of the target data

Experimental Evaluation

- Results
  - Output probability, p. 14, Figure 4, 5
    - Low variance: accuracy is higher
    - High variance: accuracy is lower
  - Phonetic dependent variances, p. 16, Table 2.
    - Low variance of critical articulator
      - Tongue tip y [s, z, θ]
      - Upper lip x [m, w, b, p]
    - Exceptional example
      - Velum x [m, n, ŋ]
  - Using characteristics of variance

Comparing MDN with MLP

- MLP
  - Single Gaussian probability density function
    - Mean: varied according to input (output of MLP)
    - Variance: fixed (global variance in training data)
- MDN
  - Multiple Gaussian probability density function
    - Weight: varied according to input
    - Mean: varied according to input
    - Variance: varied according to input

Comparison with Mean Likelihood

- Results
  - P. 18, Table. 3
  - Y-coordinates is more improved except for velum
    - Y: 13.3%, X: 4.8%
- Comparing probability density functions
  - P. 19, Figure 6
  - Effectiveness of multi density for modelling one-to-many mappings

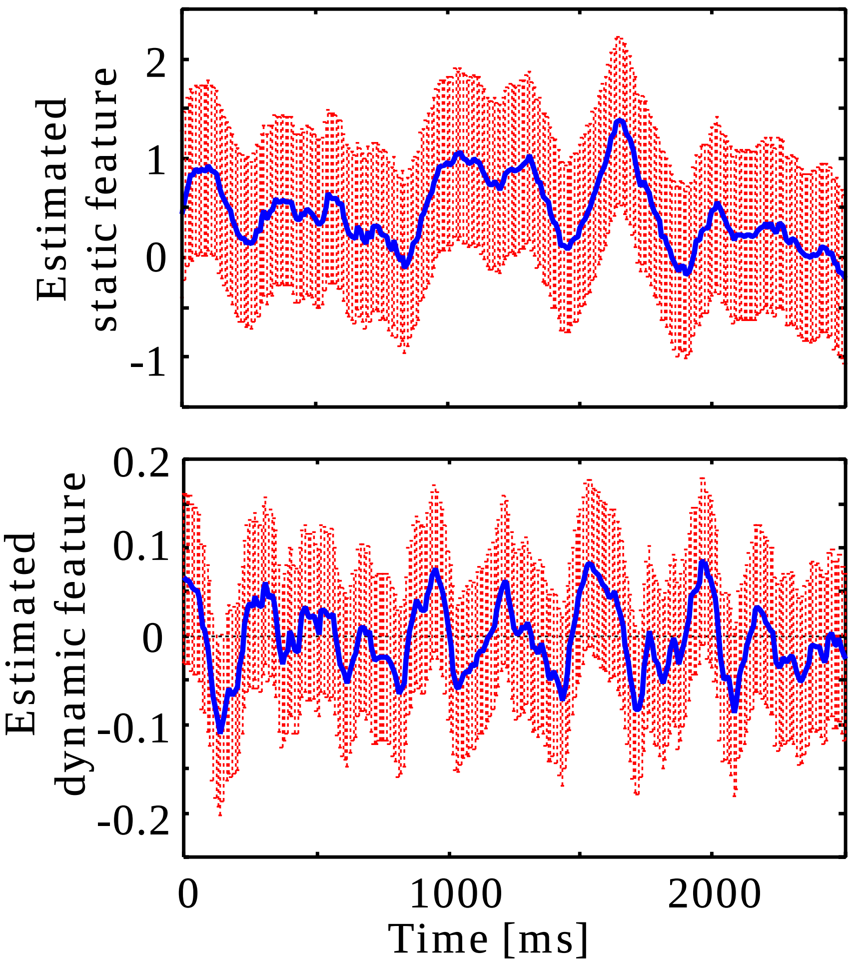
Conclusion

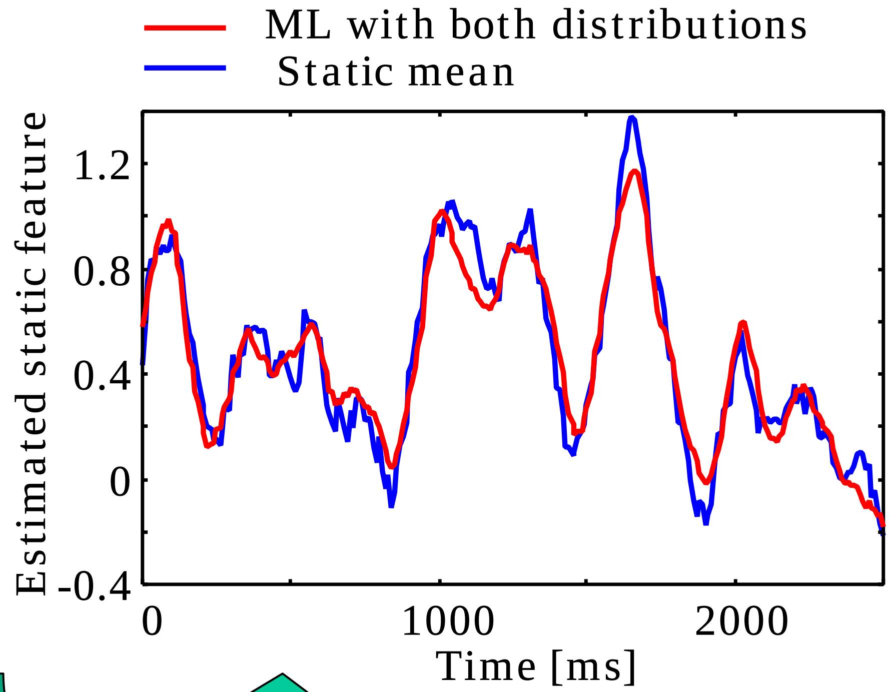
- Acoustic-to-articulatory mapping
  – Ill-posed problem
- Using MOCHA database
  – Speech waveform
  – EMA
- Mapping with MLP or MDN
  – MDN: more flexible and accurate model
    - Effective to one-to-many mapping
    - Variance varied according to input signal

Future Work

- Continuous trajectory
  - Kalman smoothing with variance
- Using articulatory information
  - Cost function in concatenative speech synthesis
  - Spectral estimation
  - Speech recognition

Proposed Algorithm





1. Estimating not only static but also dynamic probability

2. Feature generation with ML for static and dynamic distributions