Paper Introduction

~ Non-Parallel Training for Voice Conversion by Maximum Likelihood Constrained Adaptation ~

Athanasios Mouchtaris, Jan Van der Spiegel, and Paul Mueller

Tomoki Toda

July 8, 2004

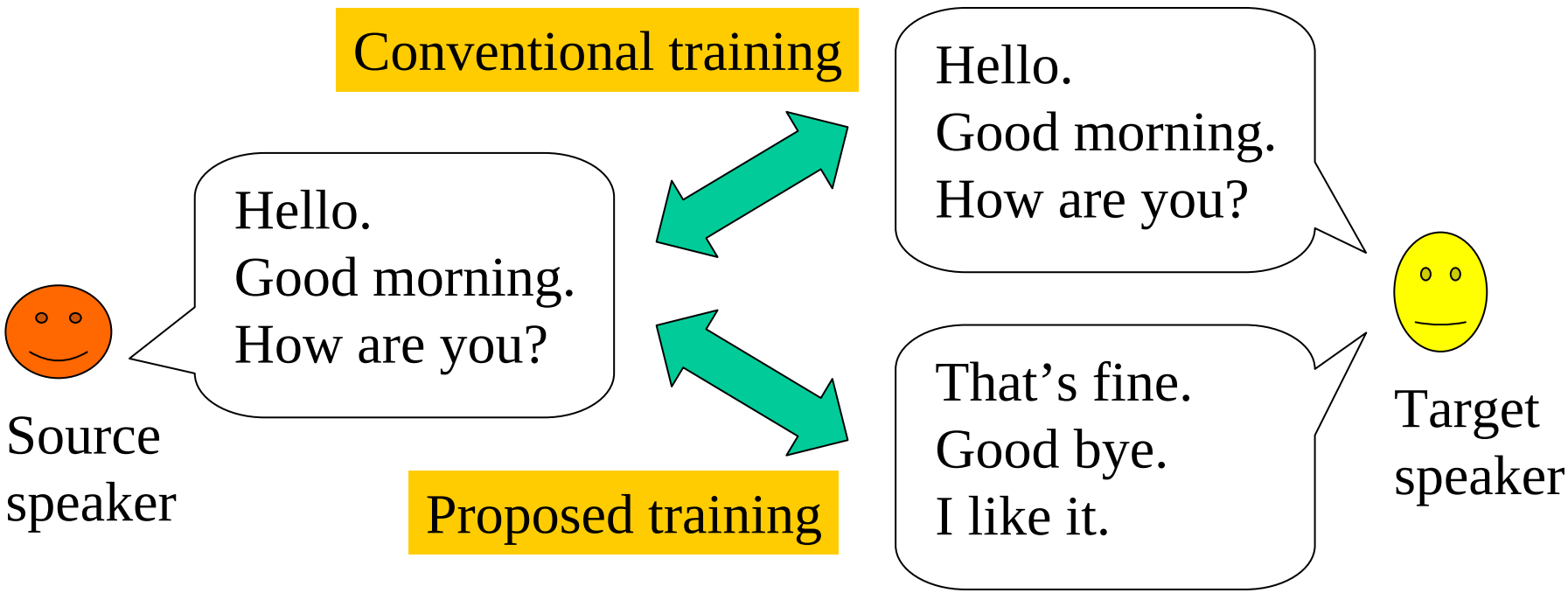
Introduced Paper

Novel techniques for flexible speech synthesis

- Voice conversion
  - Training with non-parallel speech corpus (ICASSP2004)
- Paper list for HMM-based speech synthesis
  - Adaptation, eigenvoices, and speaker interpolation (ICASSP, EUROSPEECH, ICSLP, …)

Problem Addressed in This Paper

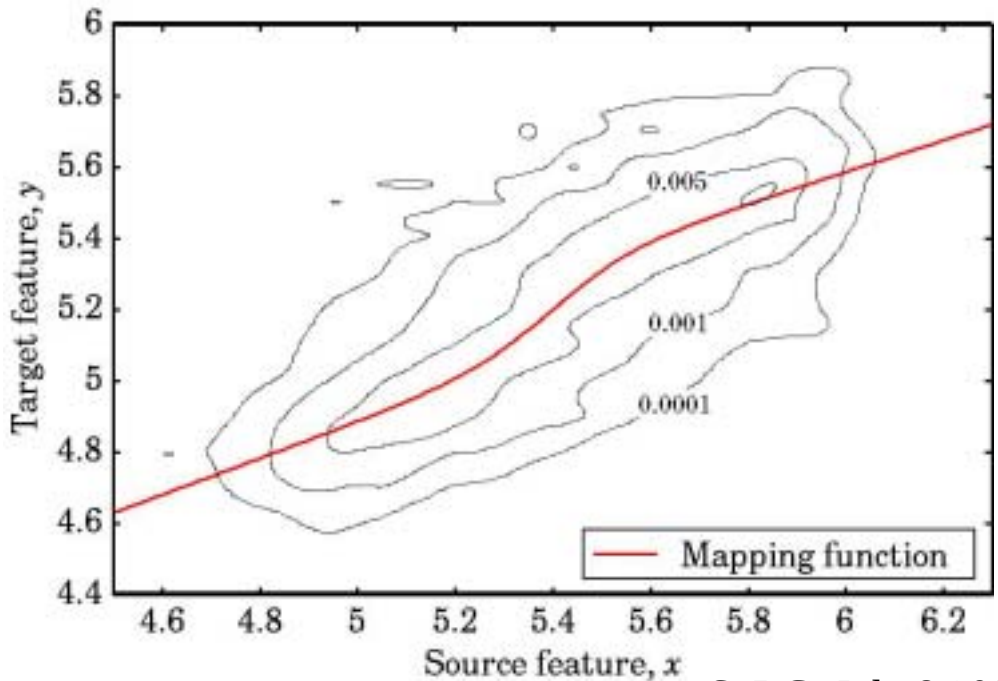
- Training of conversion function from a non-parallel corpus



Conventional Approach

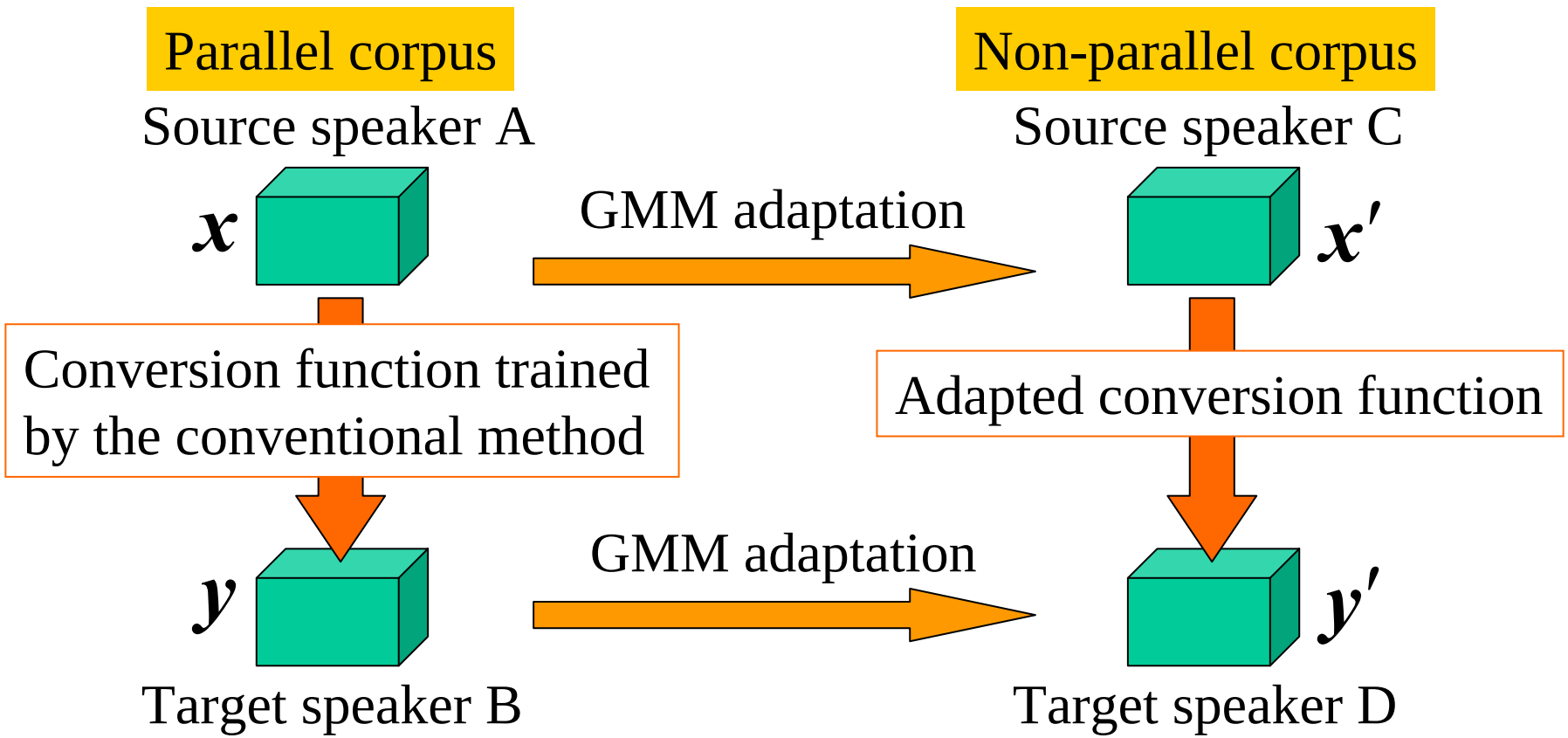
- Using a parallel corpus consisting of the same sentences uttered by source and target speakers
- Modeling a joint probability by a GMM with time-aligned source and target features

Correlation between features can be modeled directly.



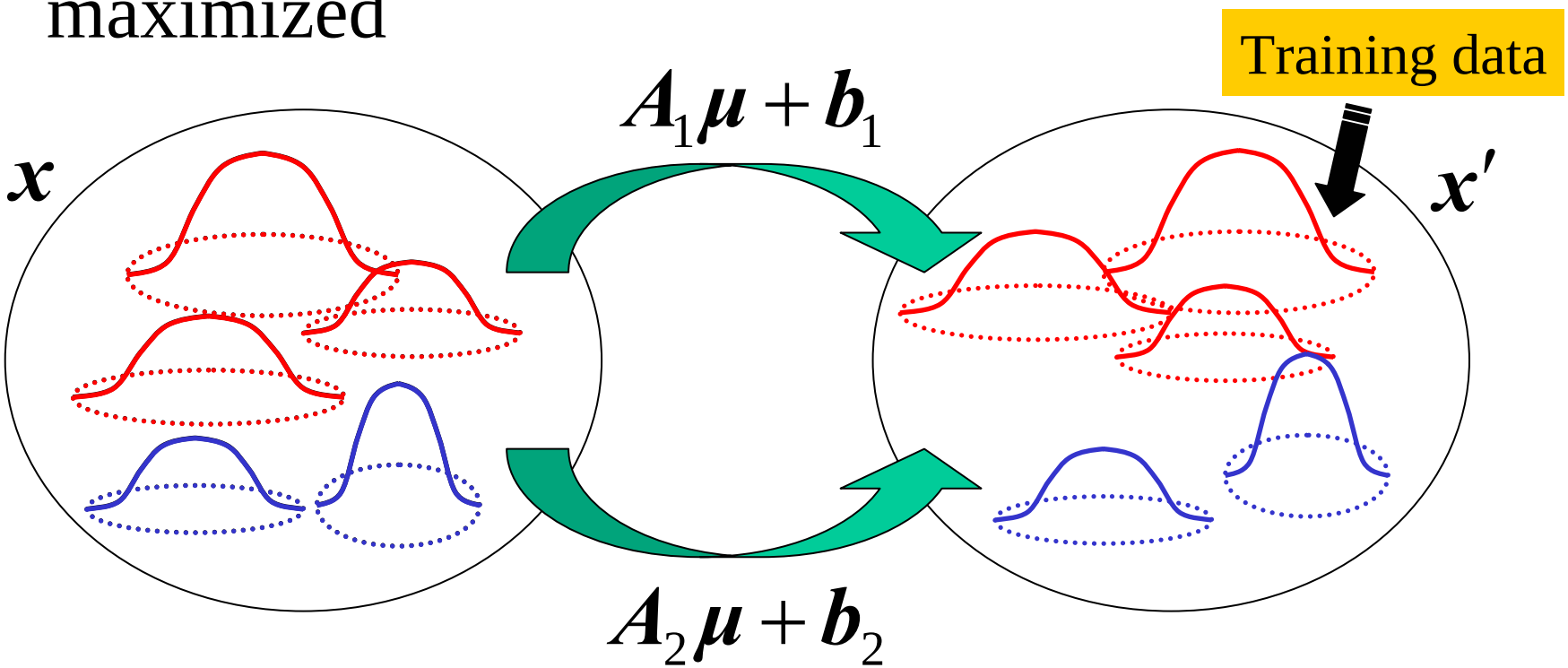
Proposed Approach

- Estimating conversion function from statistics of a parallel corpus with ML adaptation



Adaptation: MLLR (Leggetter et al.)

- Estimating linear transformation for each group of mixtures so that a likelihood function is maximized



* Covariance matrices can also be transformed.

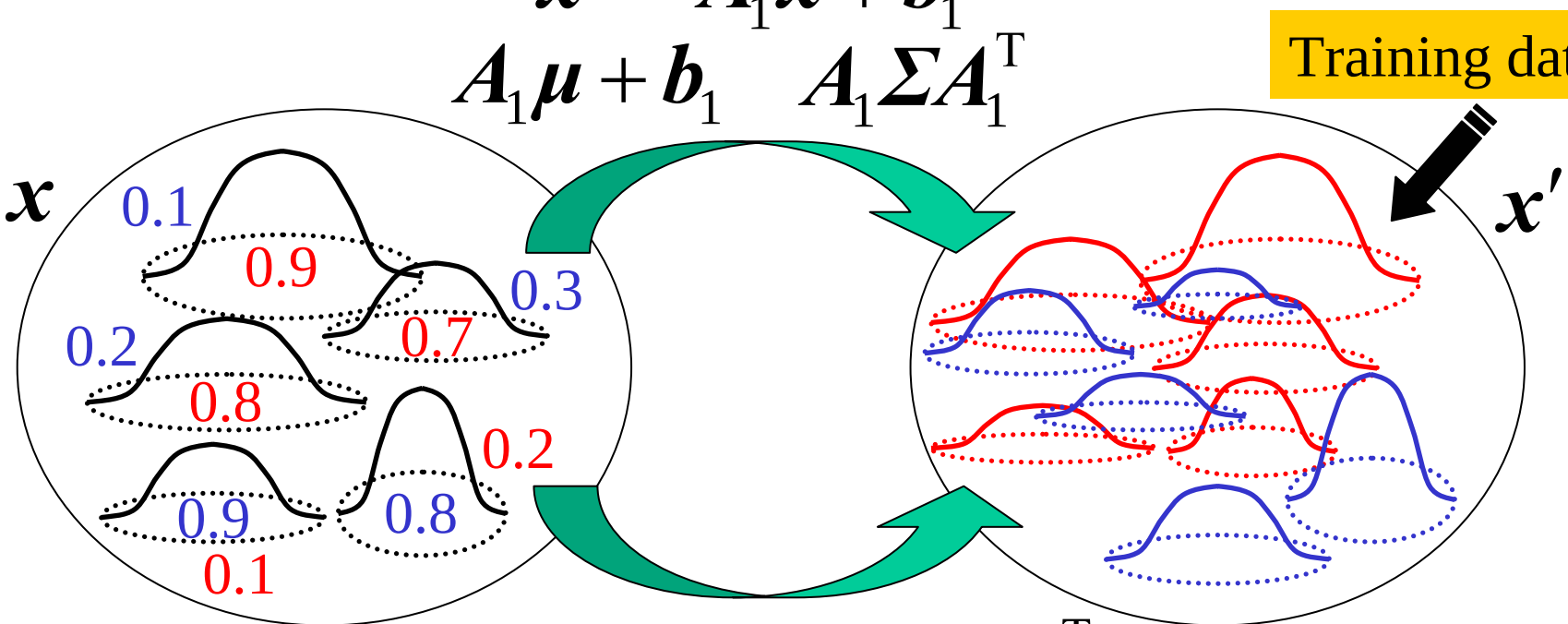
ML Stochastic-Transformation (MLST)

(Diakoloukas et al.)

- Estimating linear transformation for each transformation-component under linear constraints

$$x' = A_1 x + b_1$$

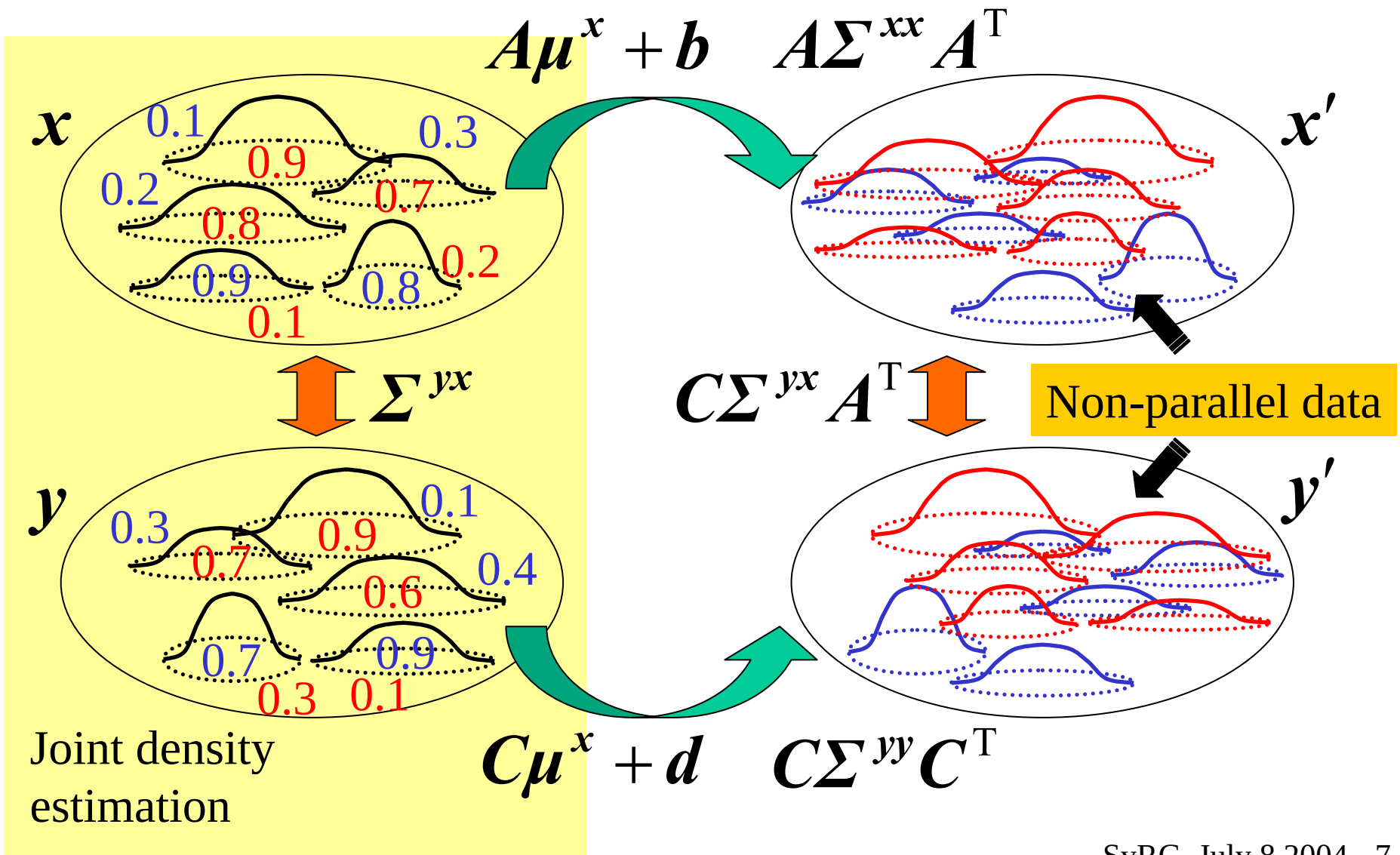
$$A_1\mu + b_1 \quad A_1 \Sigma A_1^{\mathrm{T}}$$



$$A_2\mu + b_2 \quad A_2 \Sigma A_2^{\mathrm{T}}$$

$$x' = A_2 x + b_2$$

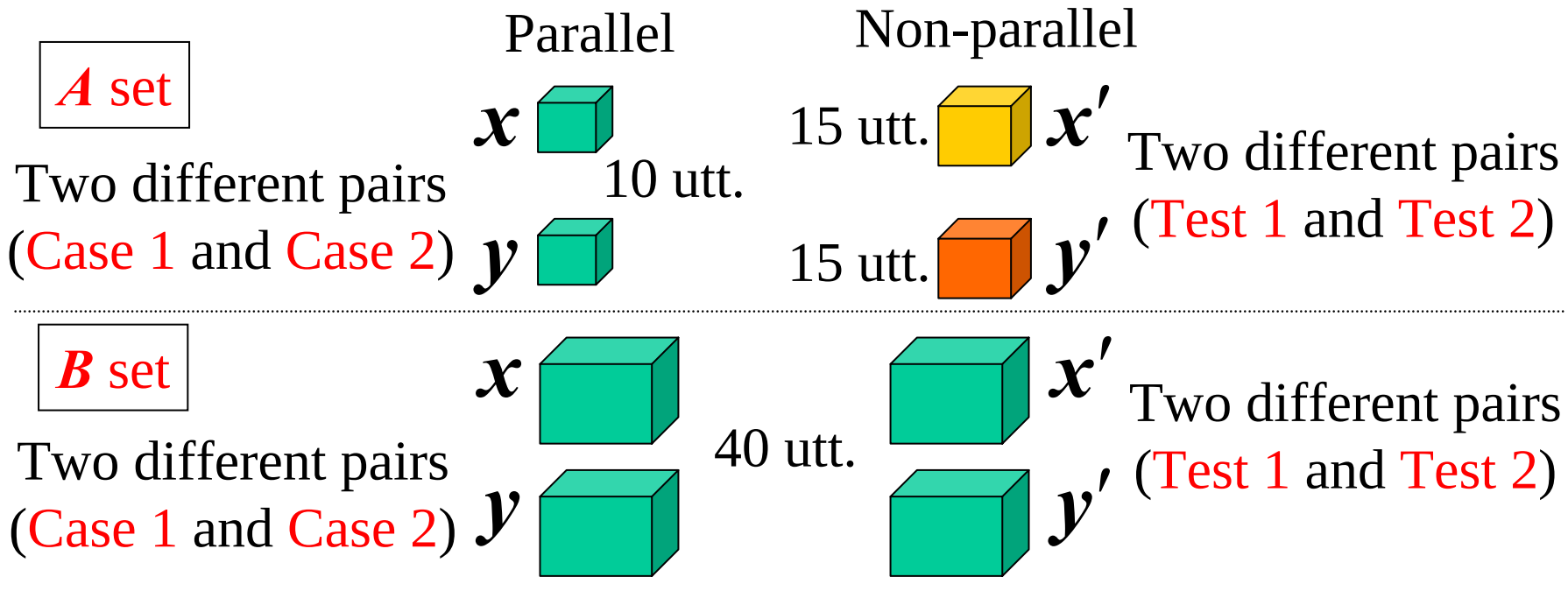
MLST with Non-Parallel Corpus



Objective Experimental Evaluations

- Evaluation with an objective measure based on spectral distance
- Investigation
  - Effectiveness of the proposed method
  - Effect of the number of adaptation parameters
  - Effect of the amount of training data
- Experimental conditions
  - Training set: 40 sentences
  - Evaluation set: 10 sentences
  - Number of mixtures: 16

Experimental Results (Table 1)



- ✓ The proposed method can reduce the error in all cases.
- ✓ Parallel training is better than non-parallel training.
- ✓ ***B*** set has better results than ***A*** set because of using a larger number of training data.

Experimental Results (Fig. 2)

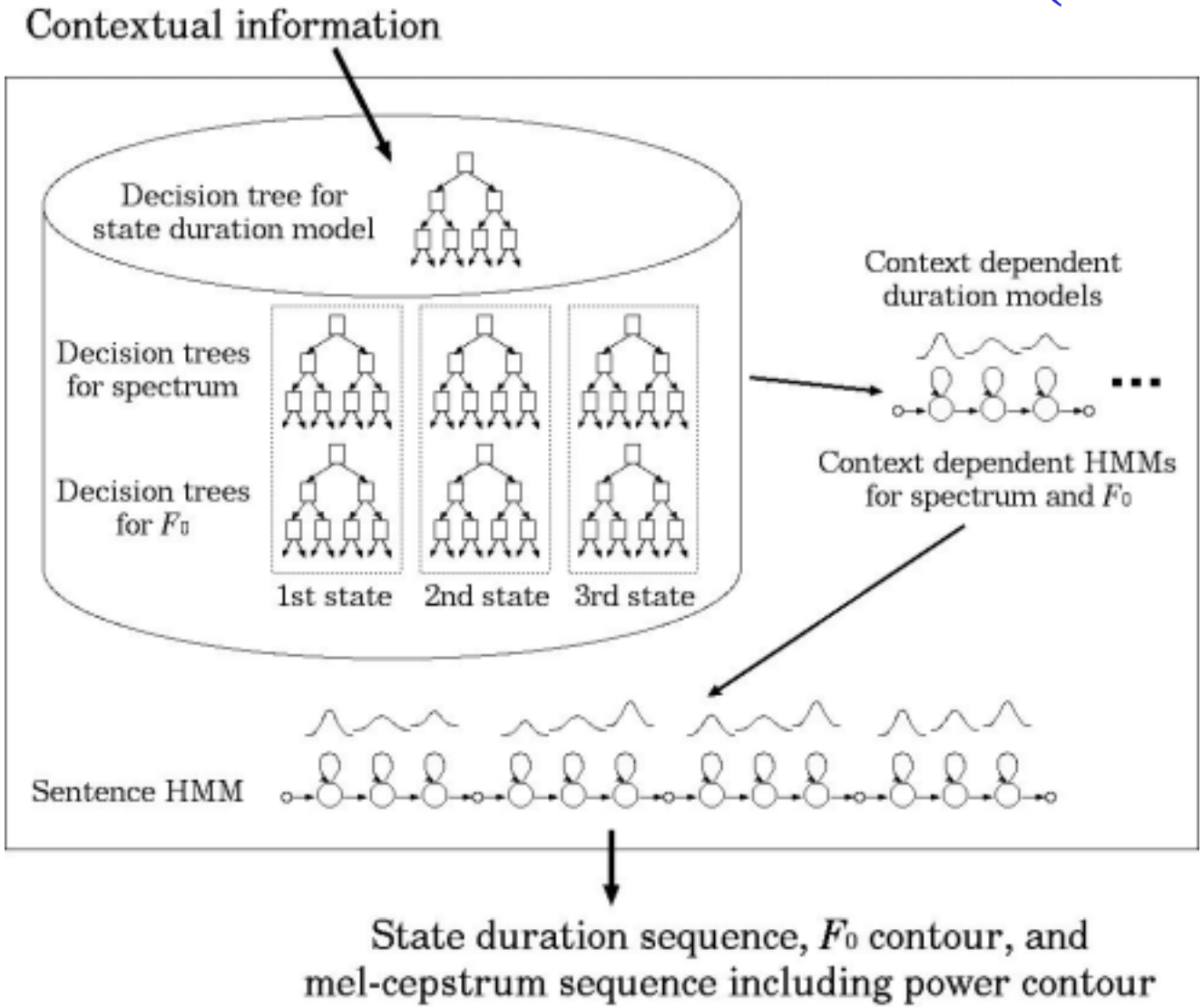
- Fig. 2(a): effect of the number of adaptation parameters
  - The error decreases when increasing the number of adaptation parameters.
  - The improvement tendency is saturated.
- Fig. 2(b): effect of the amount of training data
  - The error decreases when increasing the amount of training data.
  - The improvement tendency is saturated.

Conclusions

- Proposing a training method with a non-parallel corpus
  - Using a GMM trained with a parallel corpus as an initial model
  - Adaptation based on ML under linear constraints
- Objective evaluations
  - Showing the effectiveness of the proposed method
  - Investigating some combinations of speakers
  - Investigating effects of the number of adaptation parameters and the amount of training data

HMM-Based Speech Synthesis

(Tokuda et al.)



Reference list

- Adaptation
  - MLLR
    - Spectrum: Tamura et al., Proc. ESCA/COCOSDA Workshop pp. 273-276, Nov. 1998
    - F_0: Tamura et al., Proc. ICASSP, pp. 805-808, May. 2001
    - Duration: Tamura et al., Proc. EUROSPEECH, pp. 345-348, Sep. 2001
  - Average voice
    - STC: Yamagishi et al., Proc. ICSLP, pp. 133-136, Sep. 2002
    - STC+SAT: Yamagishi et al., Proc. ICASSP, pp. 716-619, Apr. 2003
  - Context clustering decision tree
    - Yamagishi et al., Proc. ICASSP, pp. 5-8, May 2004
- Eigenvoices
  - Shichiri et al., Proc. ICSLP, pp. 1269-1272, Sep. 2002
- Speaker interpolation
  - Yoshimura et al., Proc. EUROSPEECH, pp. 2523-2526, Sep. 1997