A SPEAKER ADAPTATION TECHNIQUE FOR GAUSSIAN PROCESS REGRESSION BASED SPEECH SYNTHESIS USING FEATURE SPACE TRANSFORM

Tomoki Koriyama, Syohei Oshio, Takao Kobayashi
Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology, Japan

Abstract

– Propose model adaptation technique for statistical parametric speech synthesis based on Gaussian process regression (GPR)
– Incorporate acoustic-feature-space linear transform that converts acoustic features of source speakers in training data into those of a target speaker
– Objective and subjective evaluations show that the proposed approach outperformed HMM-based speaker adaptation technique

Background

– GPR-based speech synthesis outperforms HMM-based one and gives comparable with, or higher performance than DNN-based one
– Conventional study focused on speaker dependent model

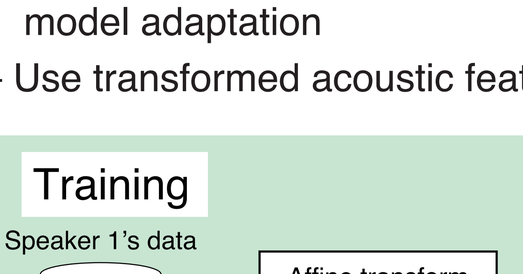
Purpose of this work

– Examine speaker adaptation technique for GPR-based synthesis

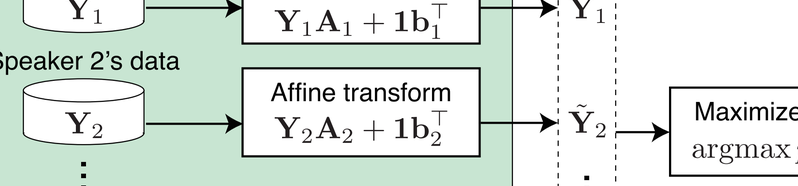
GPR-based speech synthesis

$$\mathbf{Y}_T|\mathbf{Y}_N \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

$$\boldsymbol{\mu} = \mathbf{K}_{TN}(\mathbf{K}_N + \sigma^2\mathbf{I})^{-1}\mathbf{Y}_N$$



$$\boldsymbol{\Sigma} = \mathbf{K}_T + \sigma^2\mathbf{I} - \mathbf{K}_{TN}(\mathbf{K}_N + \sigma^2\mathbf{I})^{-1}\mathbf{K}_{NT}$$



$\mathbf{Y}_N$: Acoustic feature sequence of synthetic data
$\mathbf{Y}_T$: Output feature sequence of training data
$\mathbf{K}$: Gram matrix
σ^2: Noise power

– Kernel function is defined to represent the similarity of frames
– GPR can make use of raw speech data characteristics without parameterization using means and variances

Proposed method

– Employ affine transform to target speaker's feature space as the model adaptation
– Use transformed acoustic features as training data of GPR

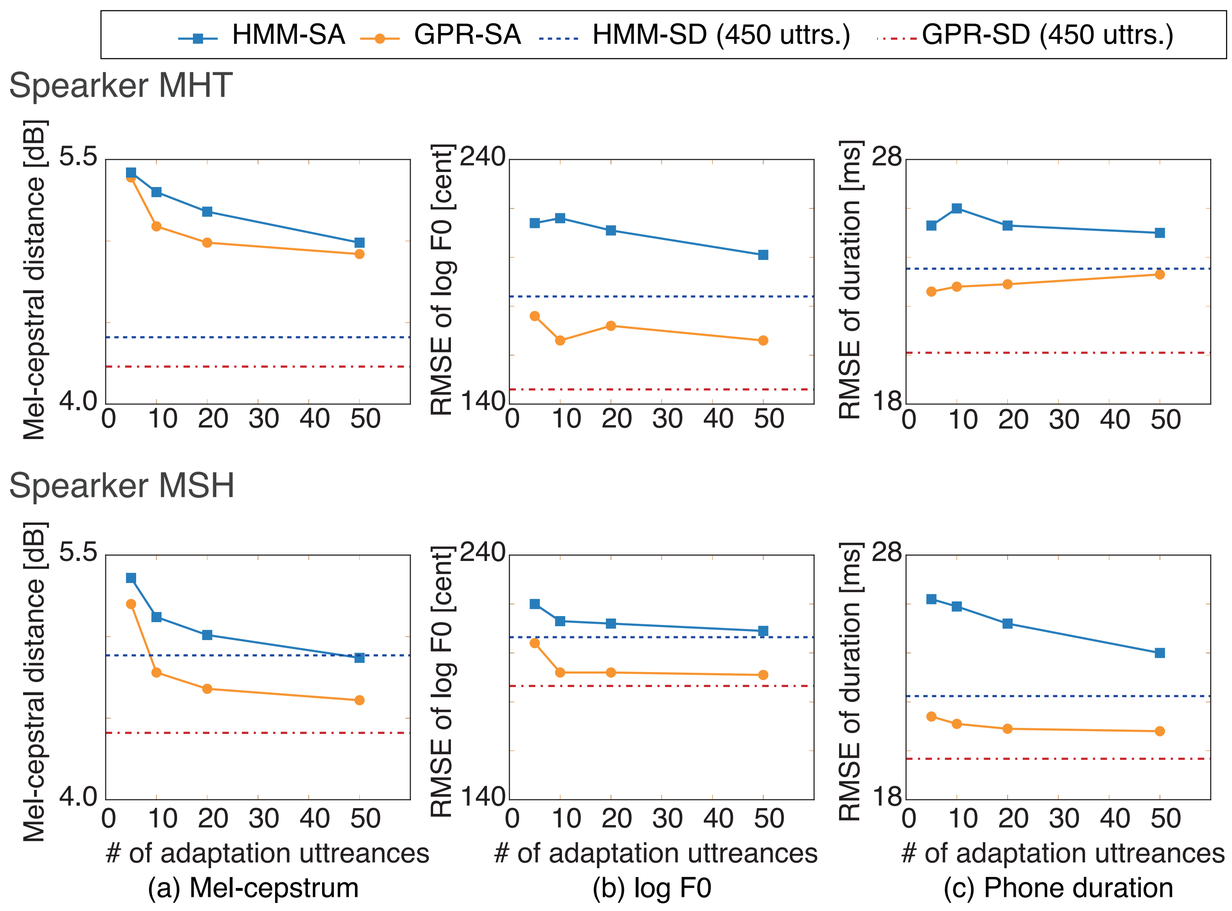
Experimental conditions

Database	ATR Japanese speech database set B
Training data	2 males (MHO, MMY), 450 utterances (approximately 40 min)
Adaptation data	2 males (MHT, MSH), 5, 10, 20, or 50 utterances (40sec. to 4min.)
Test data	53 utterances
Feature vector	0-39th mel-cepstrum, log F0, 5-band aperiodicity with their delta and delta-delta

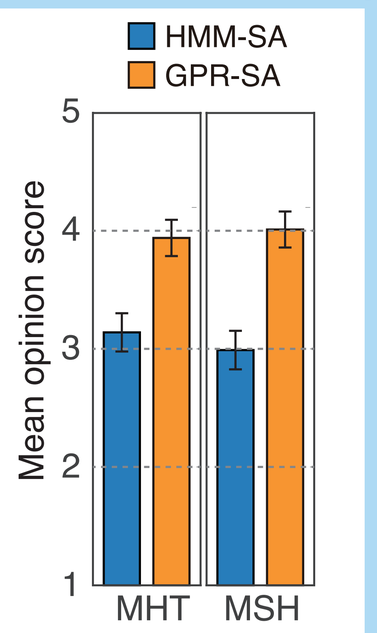
Methods

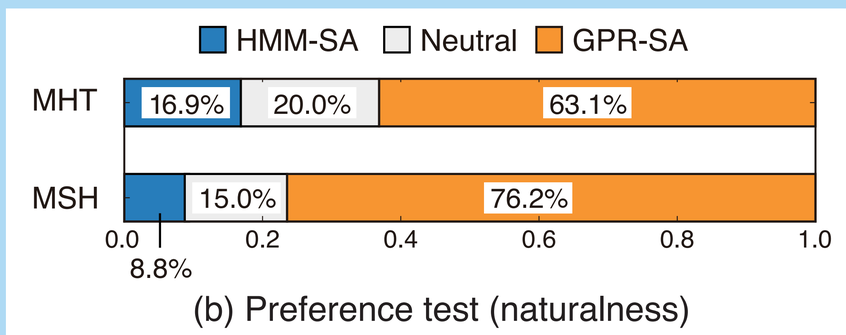
– **HMM-SA**: Adaptation using CSMAPLR+MAP
# of transforms was determined by state occupation count
– **GPR-SA**: Proposed method
Global tarnsform was used
– **HMM-SD**: Speaker dependent HMM trained by 450 utterances
– **GPR-SD**: Speaker dependent GPR trained by 450 utterances

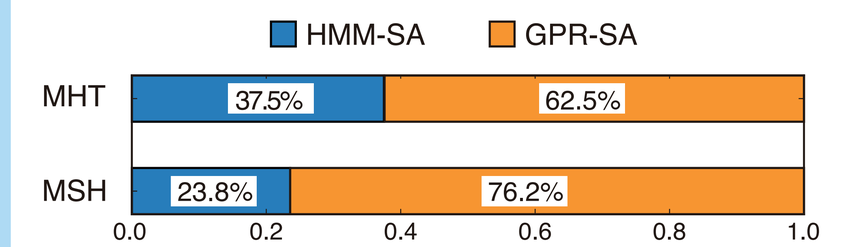
Results of objective evaluation



Results of subjective evaluation







Conclusion and future work

– Introduced feature-space transform matrices to target speaker's acoustic feature space
– Objective and subjective evaluation results showed that the proposed method outperformed the conventional HMM-based adaptation
– Future work will investigate the effect of the use of more speakers