Performance Evaluation of The Speaker-Independent HMM-based Speech Synthesis System "HTS-2007" for the Blizzard Challenge 2007

Junichi Yamagishi, Takashi Nose, Heiga Zen,
Tomoki Toda, Keiichi Tokuda

University of Edinburgh
Tokyo Institute of Technology
Nagoya Institute of Technology
Nara Institute of Science and Technology

ICASSP 2008

Introduction

HMM-based speech synthesis system [Yoshimura et al. '00]

- Generate speech parameters from statistics

  Spectral, excitation, and duration parameters

- Vocoded (but smooth and stable)

- Easy to change speaker characteristics

  Spectral, excitation, and duration parameters

  can easily be adapted to new speakers (or emotions)

Blizzard Challenge: open evaluation of speech synthesis systems using common database

Entry from HTS (HMM-based Triple S) working group

2005: Basic system + STRAIGHT, GV, & HSMM
2006: 2005 + full-covariance modeling
2007: 2006 + speaker-adaptive approach

History & The New HTS-2007 System

Strategy:

Speaker-dependent approach

2005

Hidden semi-Markov model (HSMM)

STRAIGHT with mixed excitation

Parameter generation algorithm considering global variance (GV) :
Diagonal covariance GV pdf

2006

Full covariance modeling:
MLLT/ Semi-tied covariance

Parameter generation algorithm considering GV :
Full covariance GV pdf

Speaker-adaptive approach

2007

Hidden semi-Markov model (HSMM)
Adaptive training & adaptation

STRAIGHT with mixed excitation

Average voice model

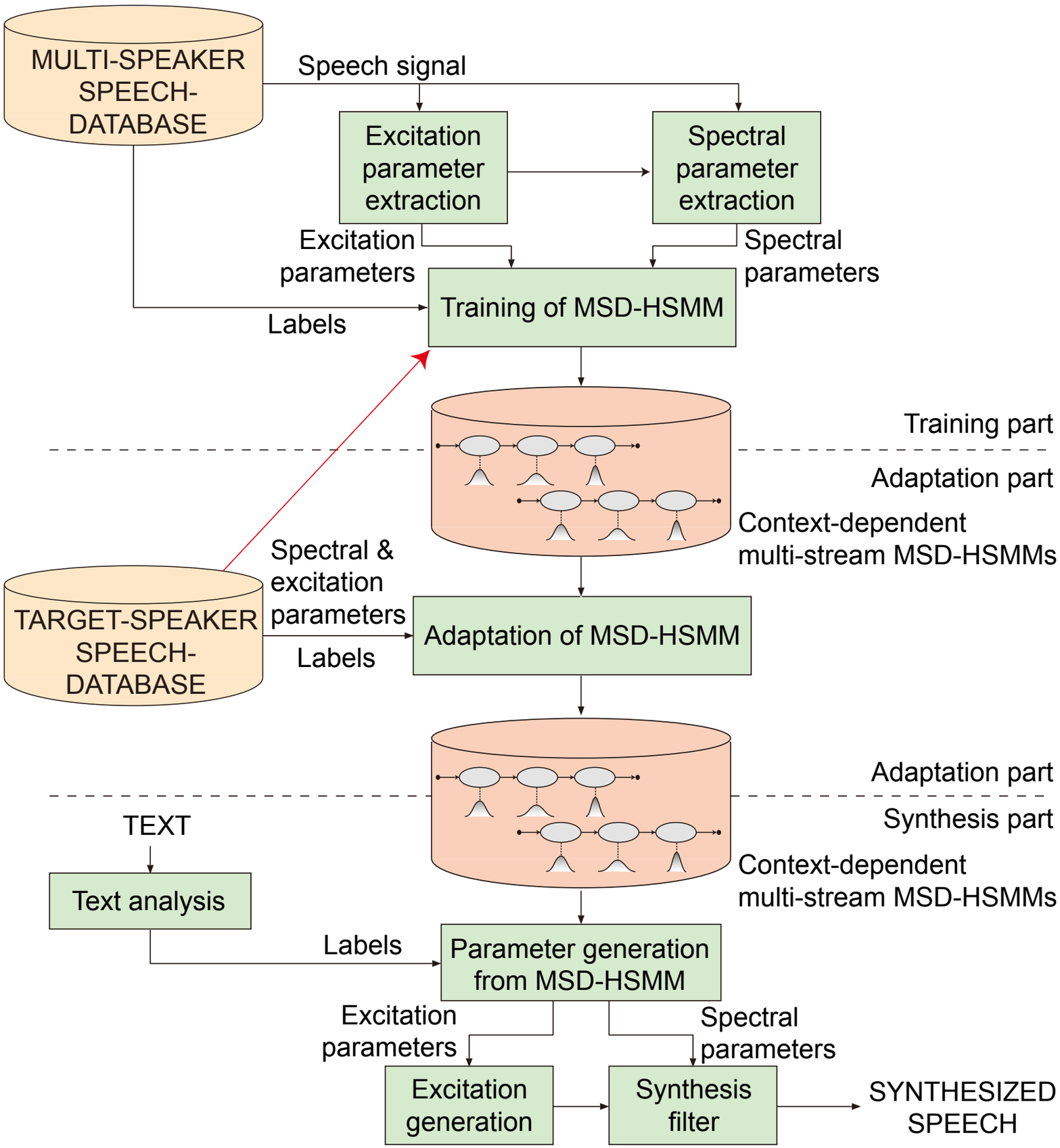
Mixed-gender acoustic modeling

CSMAPLR+MAP speaker adaptation

Full covariance modeling:
CSMAPLR transforms

Parameter generation algorithm considering GV :
Full covariance GV pdf

Overview: HTS-2007



STRAIGHT mel-cepstrum, F0, Aperiodicity measures

HSMM-based speaker adaptive training

Mixed-gender modeling

HSMM-based CSMAPLR+MAP adaptation

Full covariance modeling using CSMAPLR transforms

Parameter generation considering GV using full covariance GV pdf

STRAIGHT mel-cepstral vocoder with mixed excitation

Comparison points in this talk

Reports in previous talks

- HSMM-based adaptation and adaptive training
  [J. Yamagishi et al. IEICE Trans. 2007]
- CSMAPLR+MAP speaker adaptation
  [J. Yamagishi et al. ICASSP 2007]
- Mixed-gender modeling [J. Yamagishi et al. SSW6]
- Analysis/comparison of speaker-dependent and speaker-adaptive approaches using 3 to 30 min. of data
  [J. Yamagishi et al. ICASSP 2006]

Report in this talk

- Full-covariance modeling using CSMAPLR transforms
- Analysis/comparison of speaker-dependent and speaker-adaptive approaches using 1 to 8 hours of speech data

Full-Covariance Modeling

Diagonal covariance:

Ignore within-frame correlations

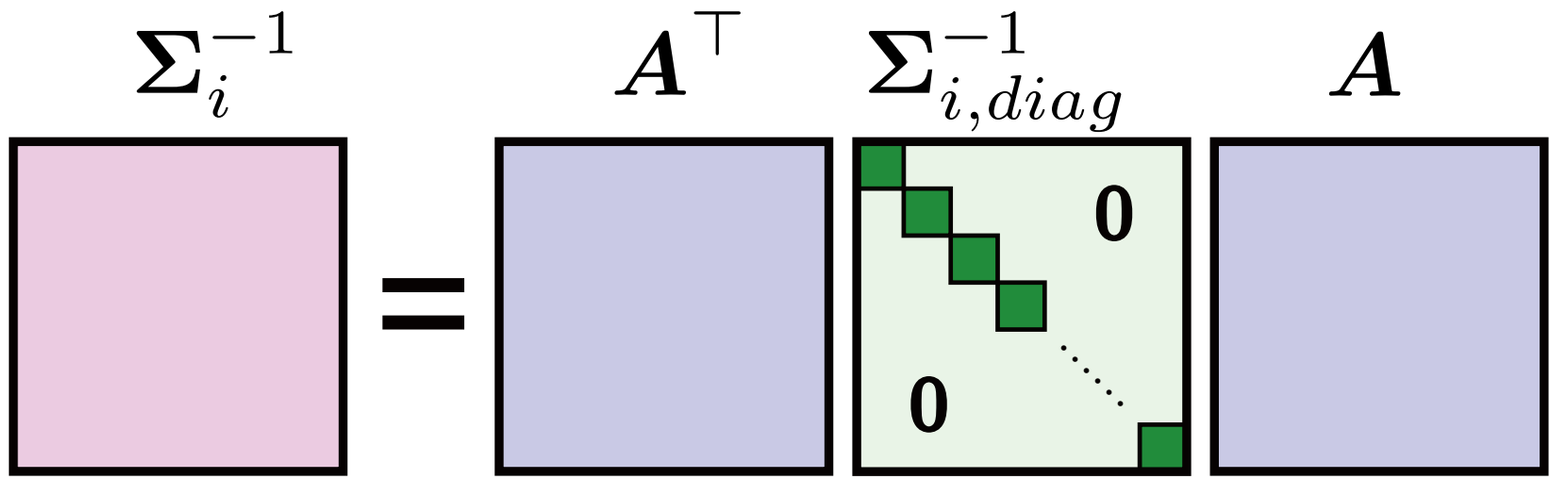
Full-covariance:

Direct modeling:

Number of model parameters drastically increases

Estimation accuracy becomes worse

An approximation method to full-covariance matrix



$\Sigma_{i,diag}^{-1}$ Diagonal precision (inverse cov.) matrix of state i

Σ_i^{-1} Approximated full precision matrix of state i

A Square transform matrix

CSMAPLR Full Covariance Modeling

CSMAPLR Transform

transform for mean

$$\widehat{\boldsymbol{\mu}}_i = \boldsymbol{A}_i^{-1}\boldsymbol{\mu}_i - \boldsymbol{A}_i^{-1}\boldsymbol{b}_i$$

transform for covariance

$$\boldsymbol{\Sigma}_i^{-1} = \boldsymbol{A}_i^{\top}\boldsymbol{\Sigma}_{i,diag}^{-1}\boldsymbol{A}_i$$

$\boldsymbol{\mu}_i$ mean vector of state i $\quad$ $\boldsymbol{b}_i$ bias vector of state i

Speaker adaptation:

Mean: (Piecewise) linear regression

Covariance: **From diagonal to full**

Advantages w.r.t. full-covariance modeling

Multiple transforms can be estimated

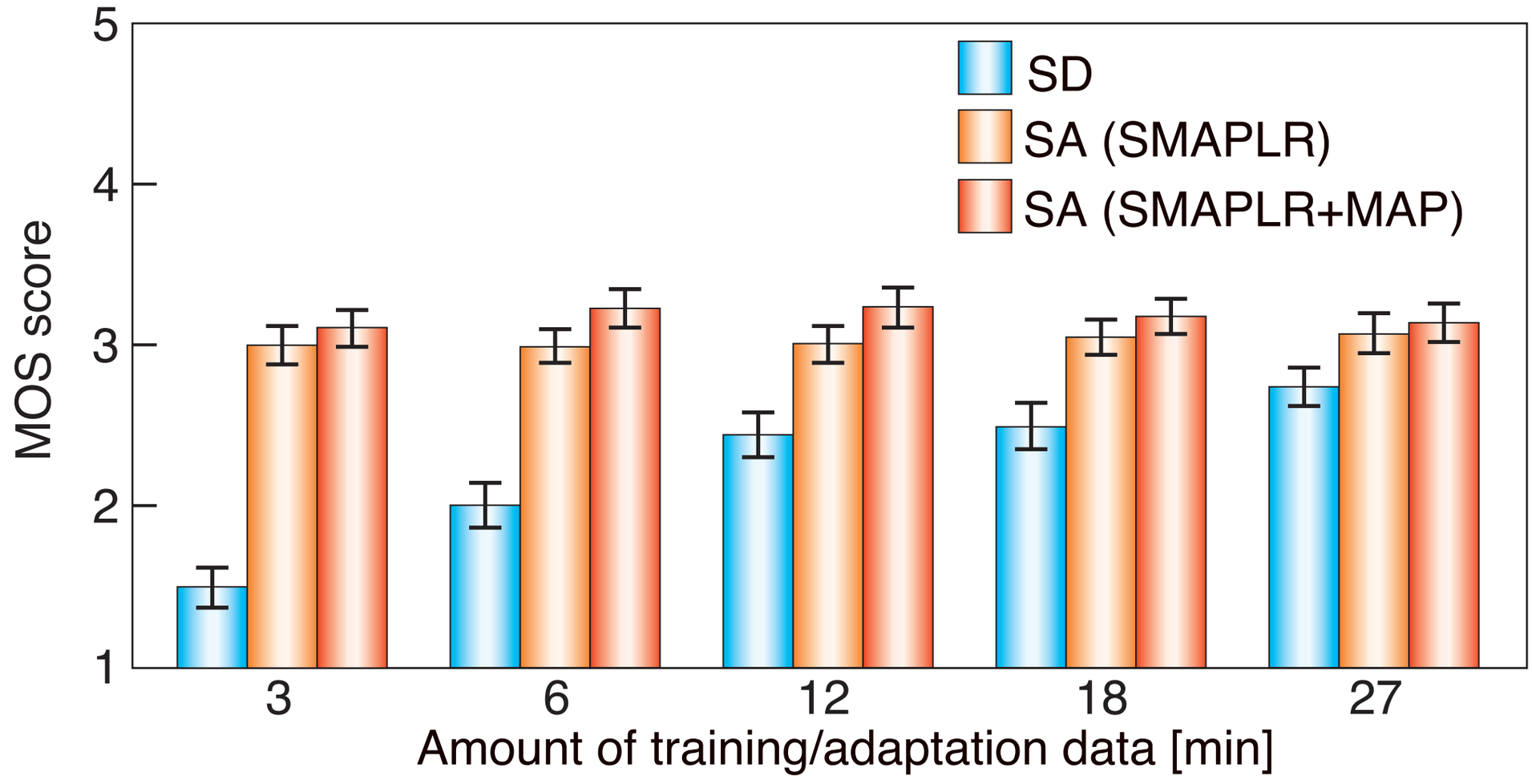
Precise approximation

Structural MAP (SMAP) criterion [K. Shinoda et al. ‘01] can be used to estimate the multiple transforms

Robust estimation even from limited amount of data

Speaker-dependent vs adaptive approach

Comparison of speaker-dependent and adaptive approaches
[J. Yamagishi et al. ICASSP 2006]



Speaker-adaptive (SA) approaches outperform speaker-dependent (SD) approach using 5 to 30 minutes of speech data.

How about more than 30 minutes of speech data?

6 minutes, 1 hour, and 8 hours

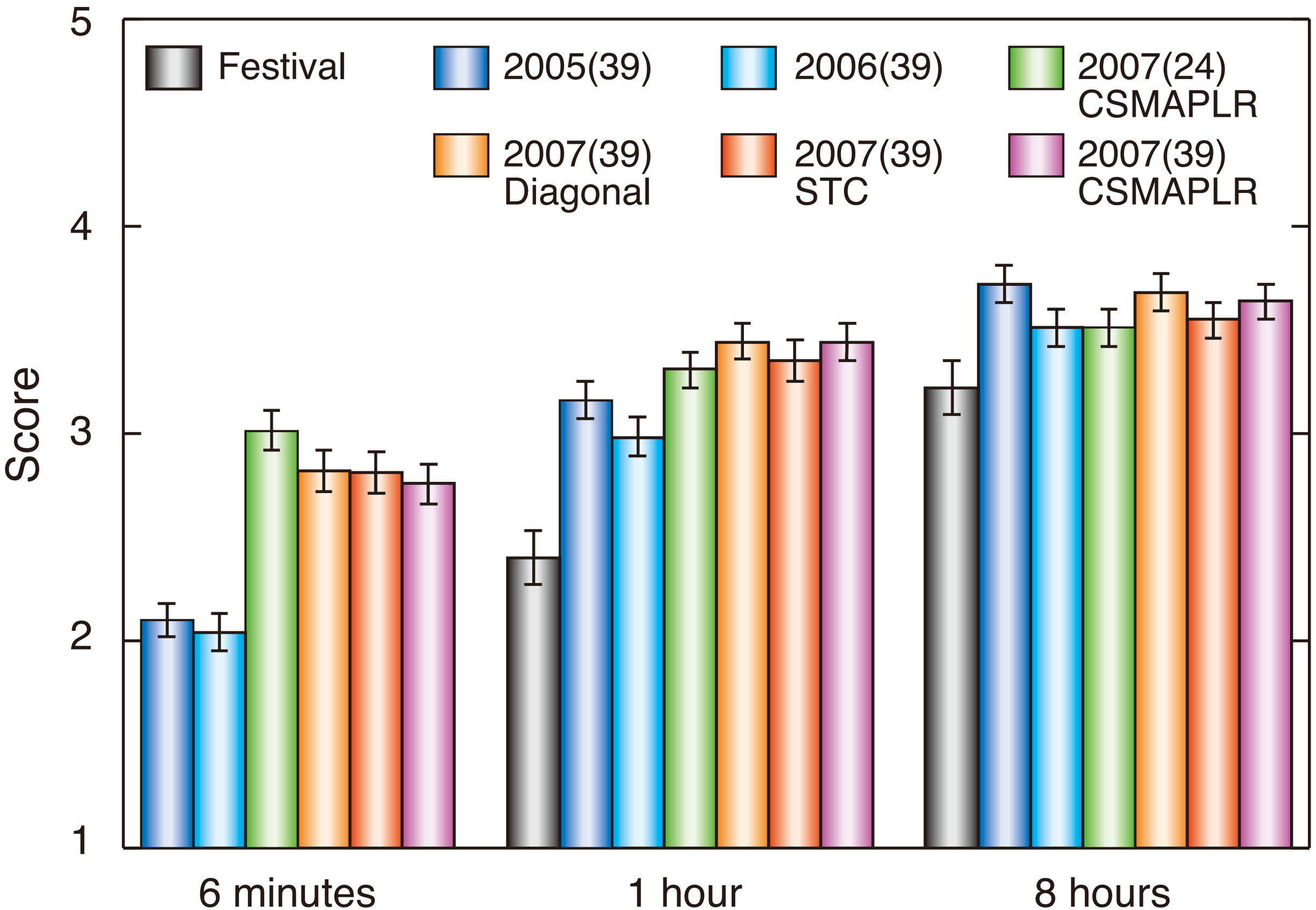
Experimental Conditions: English

Database	CMU-ARCTIC database 4 male speakers & 2 female speakers 6,780 utterances ATRECSS (Blizzard Challenge 2007) corpus 1 male speaker 6,579 utterances
Sampling rate	16 kHz
Spectral Analysis	512-order STRAIGHT analysis
Feature Vector	**0–24** or **0–40** STRAIGHT mel-cepstrum, logarithmic F0, 5 aperiodicity measures, and their delta, delta-delta parameters
Model	Context-dependent state-tied multi-stream 5-state left-to-right MSD-HSMM Gaussian pdf: Single mixture, Diagonal covariance

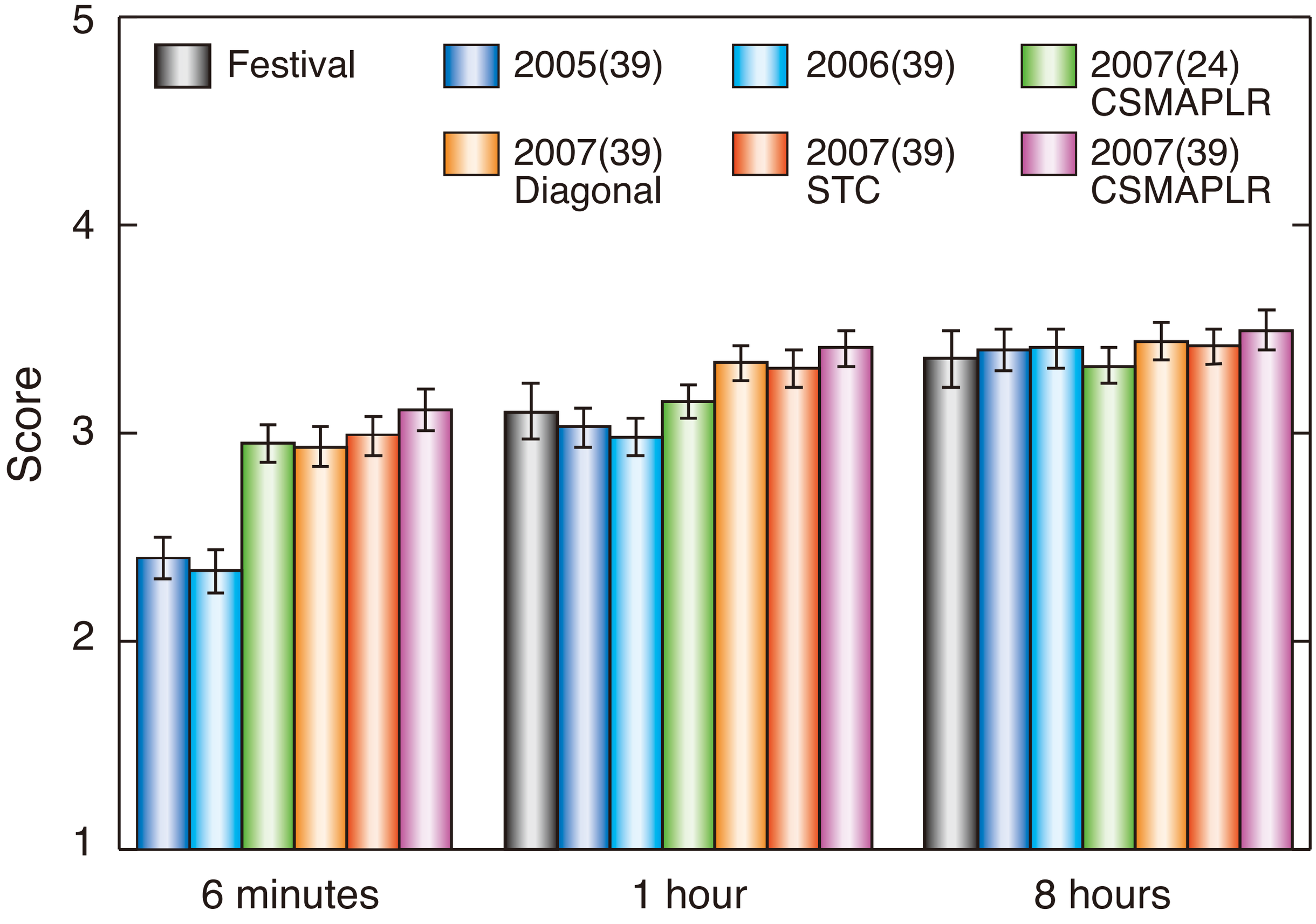
Experimental Conditions

Evaluation Methods	MOS test (naturalness) 1: poor — 5: natural CCR test (similarity) 1: very dissimilar — 5: very similar to reference
# of subjects	33 persons
# of test sentences	14 sentences randomly chosen from 50 sentences
Calibration system	Festival speech synthesis system (unit-selection)

Experimental results: MOS scores



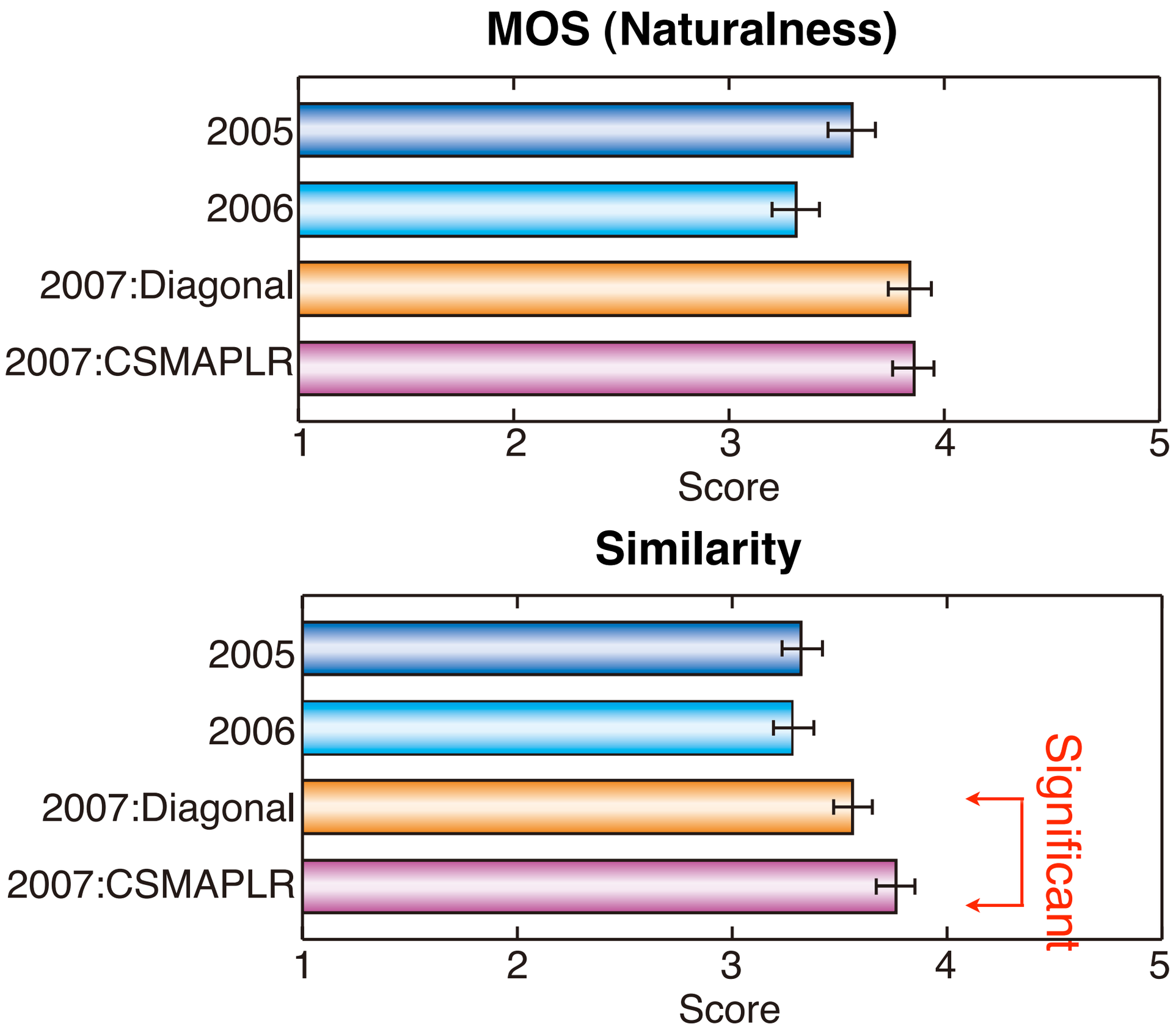
Experimental results: Similarity



Experimental Conditions: Japanese

Database	ATR Japanese speech database (B-set, C-set) 7 male speakers & 5 female speakers 5,230 utterances Japanese speech database of NIT and TIT 3 male speakers & 1 female speakers 2,012 utterances
Sampling rate	16 kHz
Spectral Analysis	512-order STRAIGHT analysis
Feature Vector	0–40 STRAIGHT mel-cepstrum, logarithmic F0, 5 aperiodicity measures, and their delta, delta-delta parameters
Model	Context-dependent state-tied multi-stream 5-state left-to-right MSD-HSMM Gaussian pdf: Single mixture, Diagonal covariance

Experimental results: Japanese



Demonstration: Various Voices

The HTS-2007 system can adapt the average voice model into ...

US English

Male

Female

Indian English

Male

Celebrity

Male

Can you guess the celebrity?
A: George W. Bush (GWB)
(another sample)

Conclusions

HTS-2007 System: High-quality speaker-adaptive speech synthesis

- significantly better than the speaker-dependent approaches in the case of realistic amount of speech data (<< 8 hours)
- comparable to the speaker-dependent approaches eve in the case of 8 hours of speech data
- significantly better than the Festival unit-selection system
  - HTS-2007 (6 min.) was comparable to Festival (1 hour)
  - HTS-2007 (1 hour) was comparable to Festival (8 hours)

Other Findings

Full-covariance modeling:

- Improves similarity of synthetic speech

High-order mel-cepstral analysis:

- Improves similarity when large amount of data is available
- Degrades naturalness when amount of speech data is limited

Online demonstration of HTS-2007

HTS-2007(39, diagonal), HTS-2005, & Festival Systems

http://www.cstr.ed.ac.uk/projects/festival/morevoices.html

Currently 5 unit-selection and 23 HTS voices are available

- 2 Scottish males
- 1 Scottish female
- 3 English males
- 1 English female
- 4 American males
- 2 American females

Please compare these systems yourselves

E-mail jyamagis@inf.ed.ac.uk