Acoustic Modeling of Speaking Styles and Emotional Expressions in HMM-Based Speech Synthesis

Source: IEICE TRANS. INF. & SYST.

Author : Junichi YAMAGISHI, Takao KOBAYASHI

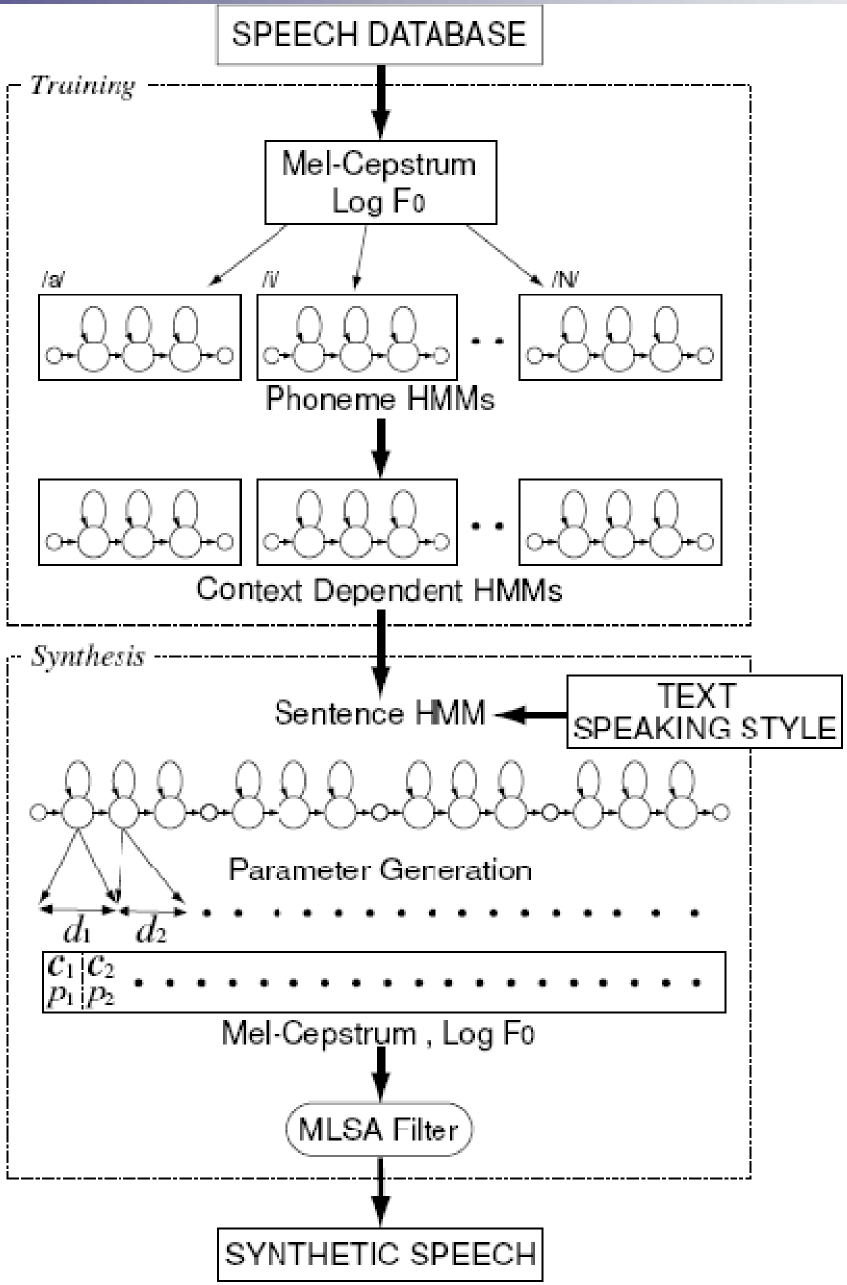
Professor : 陳嘉平

Reporter : 楊治鏞

Introduction

- Recent research on speech synthesis has focused on generating emotional expressiveness and various speaking styles in synthesized speech.

- In this paper, we describe an alternative approach that enables expressing various emotions and/or speaking styles easily and effectively in synthetic speech by using an HMM-based speech synthesis framework.


SPEECH DATABASE

Training

Mel-Cepstrum
Log F0

Phoneme HMMs

Context Dependent HMMs

Synthesis

Sentence HMM

TEXT
SPEAKING STYLE

Parameter Generation

Mel-Cepstrum , Log F0

MLSA Filter

SYNTHETIC SPEECH

Style-dependent modeling

- In the style-dependent modeling method, each style is modeled individually by using an acoustic model.
- A pseudo root node is added to the resulting decision trees of each style to combine the models for all styles into a single acoustic model.

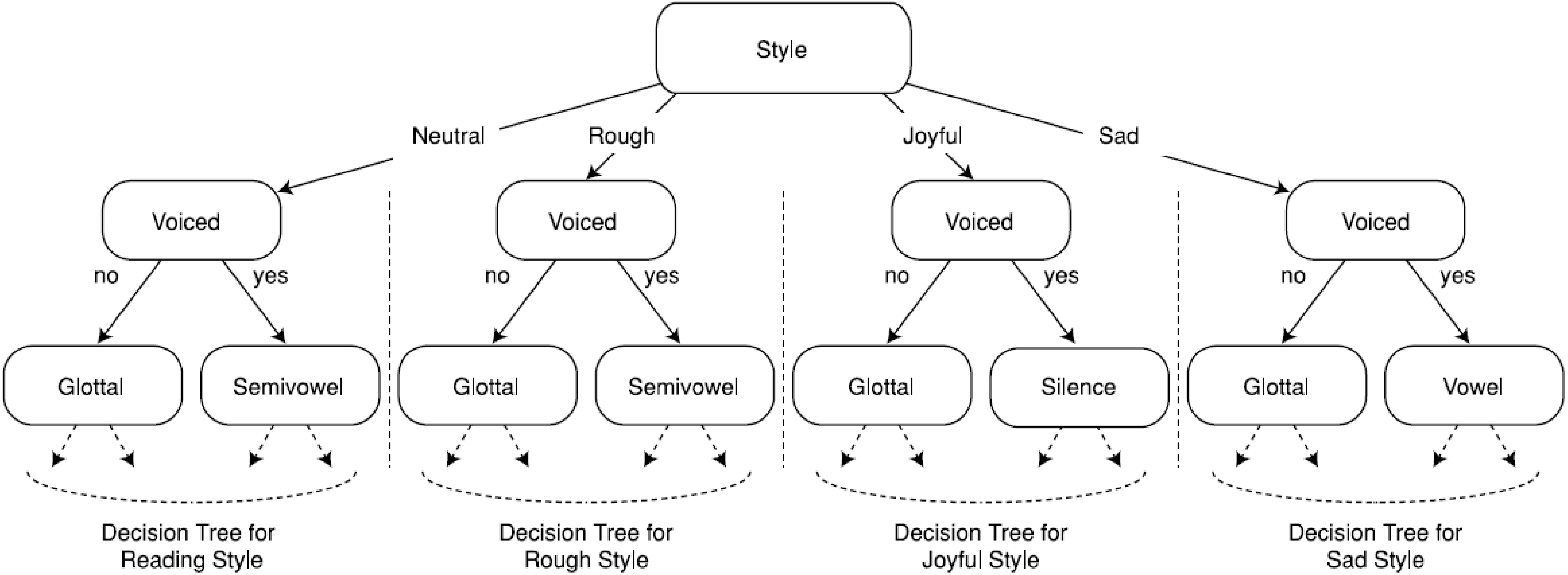


Fig. 2 Part of a constructed decision tree in style-dependent modeling. A pseudo root node is added to decision trees of each style to combine models for all styles into a single acoustic model.

Style-mixed modeling

- In the style-mixed modeling method, each style is treated as one of contexts, and the tree-based context clustering technique is applied to all styles at the same time.

- As a result, all styles are modeled by using a single acoustic model as shown in Fig. 3.

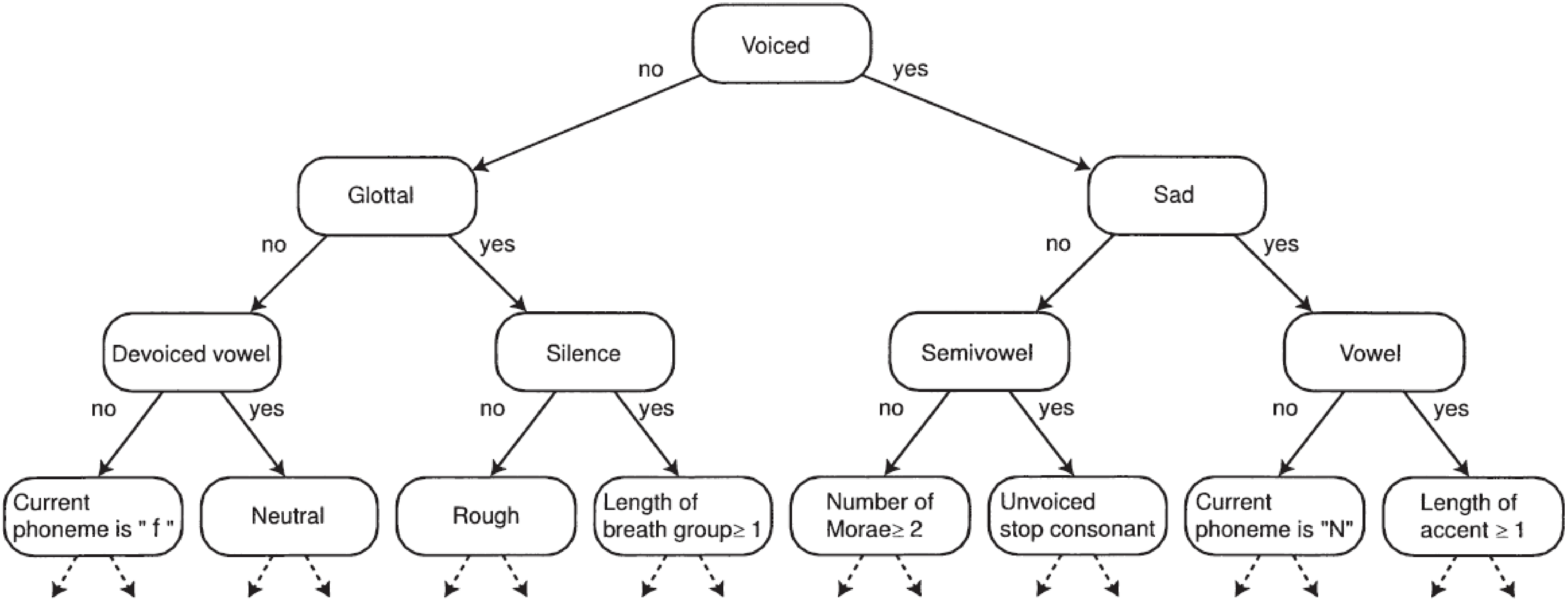


Fig. 3 Part of a constructed decision tree in style-mixed modeling. Styles are split by using style-related questions as well as other contexts.

Experiments

- To compare the proposed modeling methods, we chose four styles of read speech — ***polite***, ***rough***/***impolite***, ***joyful***, and ***sad*** — and constructed speech database, which were composed of **503** phonetically balanced sentences obtained from the ATR Japanese speech database.

- All the sentences were uttered by a male speaker, **MMI**, and a female speaker, **FTY**, in all the styles.

Experiments

- The feature vectors consisted of **25 mel-cepstral coefficients** including the zeroth coefficient, the logarithm of the fundamental frequency, and their delta and delta-delta coefficients.

- Both the style-dependent and style-mixed models were trained using **450** sentences for each style.

- We also used speech samples uttered by the same speakers in a ***neutral*** style for reference purposes.

Speech Database

- We first evaluated whether the **recorded speech samples** were perceived by listeners as being uttered in the intended styles.

- **Nine male** subjects were presented with all **503** sentences uttered in each of the styles and then asked whether they perceived the speech samples as having been uttered in the intended styles.

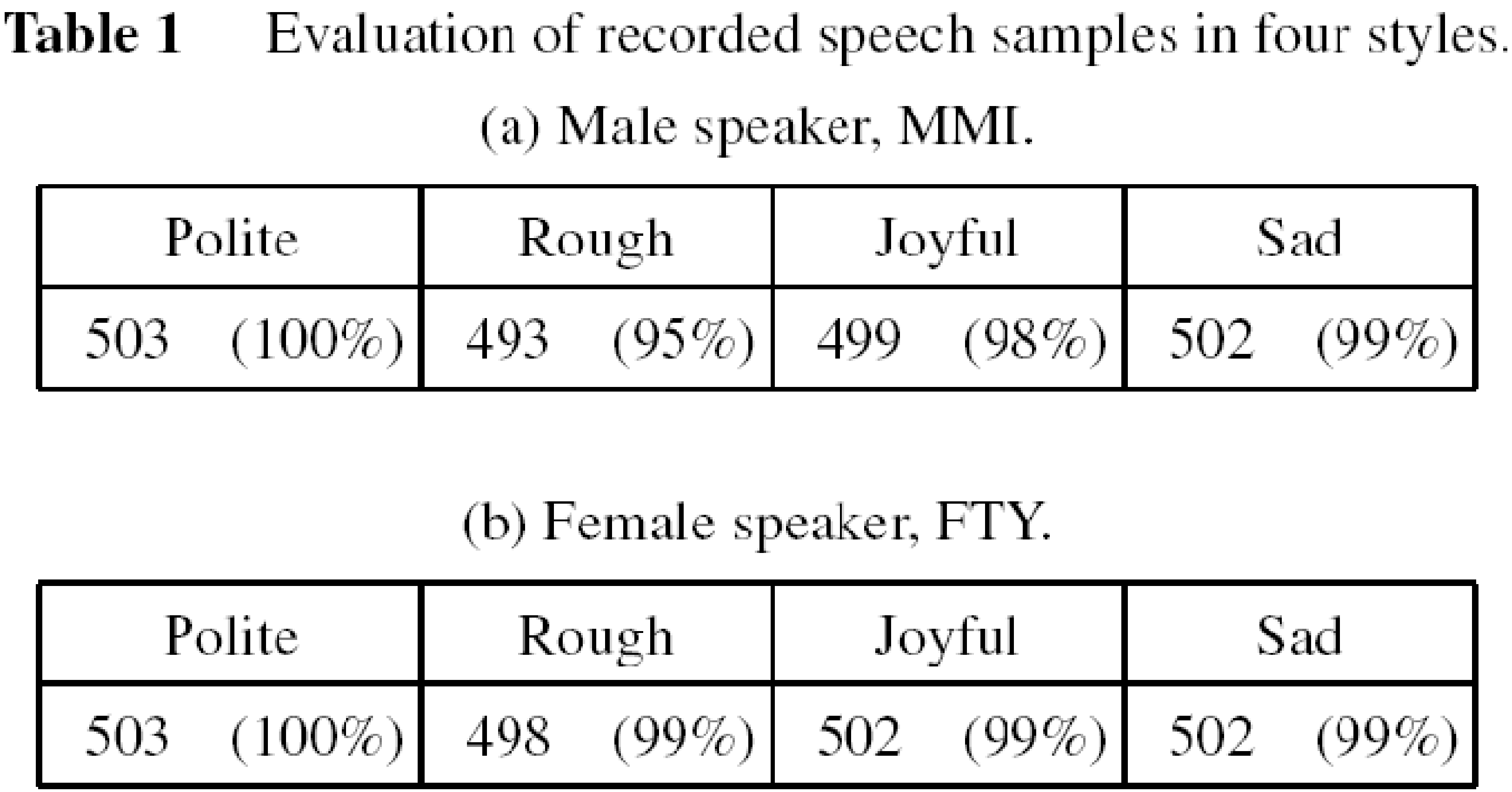


Table 1 Evaluation of recorded speech samples in four styles.

(a) Male speaker, MMI.

Polite	Rough	Joyful	Sad
503 (100%)	493 (95%)	499 (98%)	502 (99%)

(b) Female speaker, FTY.

Polite	Rough	Joyful	Sad
503 (100%)	498 (99%)	502 (99%)	502 (99%)

Speech Database

- **Nine male** subjects were asked to assign eight test sentences chosen at random from 53 test sentences to a ***neutral***, ***polite***, ***rough***, ***joyful***, or ***sad*** group.

- Speech samples that were not put by the subjects into one of these groups were classified as “other”.

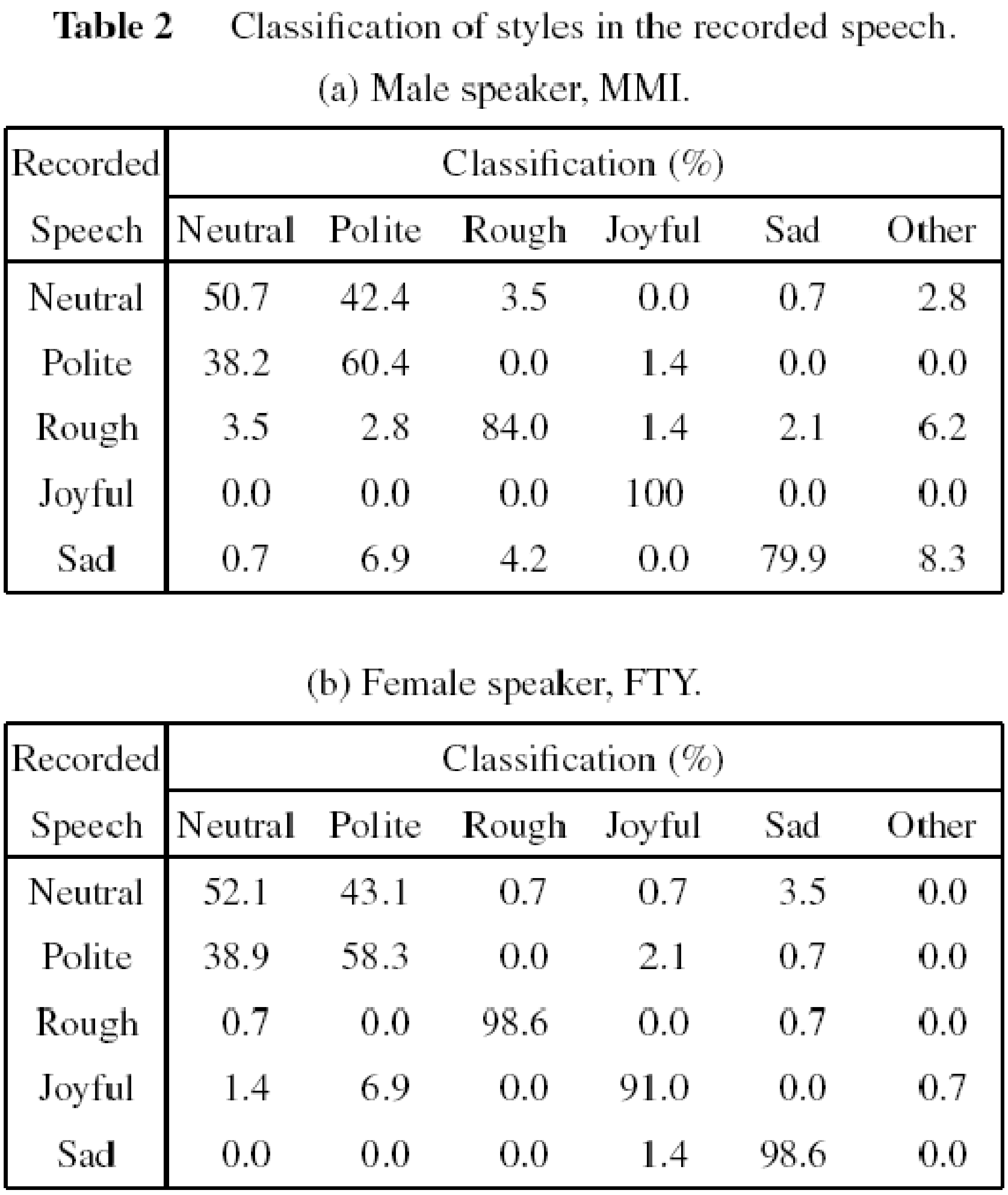


Table 2 Classification of styles in the recorded speech.

(a) Male speaker, MMI.

Recorded	Classification (%)					
Speech	Neutral	Polite	Rough	Joyful	Sad	Other
Neutral	50.7	42.4	3.5	0.0	0.7	2.8
Polite	38.2	60.4	0.0	1.4	0.0	0.0
Rough	3.5	2.8	84.0	1.4	2.1	6.2
Joyful	0.0	0.0	0.0	100	0.0	0.0
Sad	0.7	6.9	4.2	0.0	79.9	8.3

(b) Female speaker, FTY.

Recorded	Classification (%)					
Speech	Neutral	Polite	Rough	Joyful	Sad	Other
Neutral	52.1	43.1	0.7	0.7	3.5	0.0
Polite	38.9	58.3	0.0	2.1	0.7	0.0
Rough	0.7	0.0	98.6	0.0	0.7	0.0
Joyful	1.4	6.9	0.0	91.0	0.0	0.7
Sad	0.0	0.0	0.0	1.4	98.6	0.0

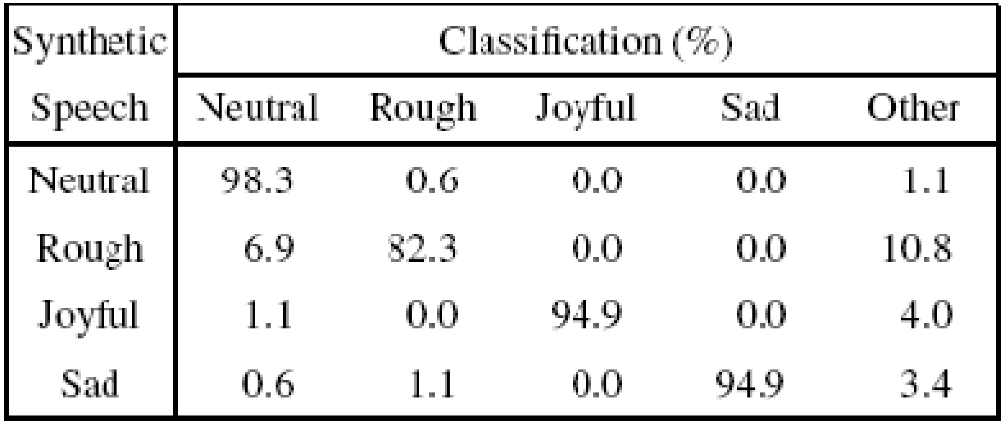
Subjective Evaluations of Styles in Synthesized Speech

- **Eleven male** subjects were asked to classify **eight test sentences** chosen at random from **53 test sentences** not included in the training data as being ***neutral***, ***rough***, ***joyful***, or ***sad*** depending on the style of speech.

- In these experiments, more than **80%** of speech samples generated using both models were judged to be similar to those in the target styles.

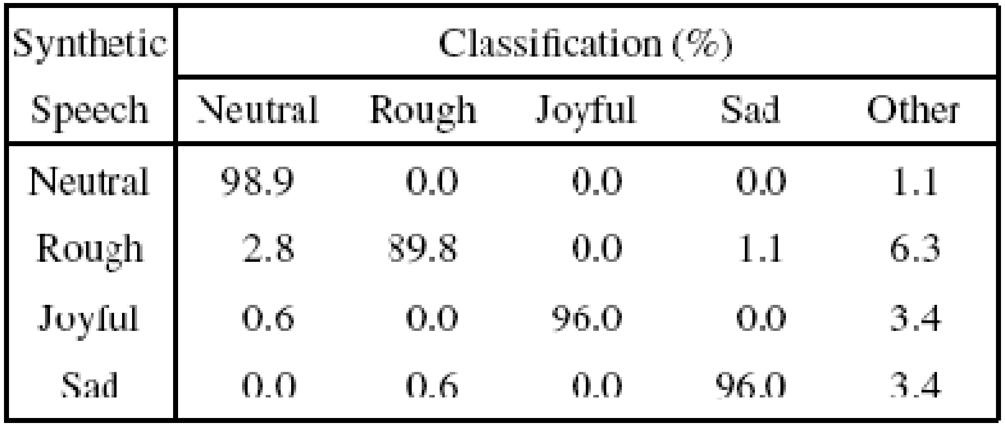
MMI

(a) Style-Dependent Model.



Synthetic	Classification (%)				
Speech	Neutral	Rough	Joyful	Sad	Other
Neutral	98.3	0.6	0.0	0.0	1.1
Rough	6.9	82.3	0.0	0.0	10.8
Joyful	1.1	0.0	94.9	0.0	4.0
Sad	0.6	1.1	0.0	94.9	3.4

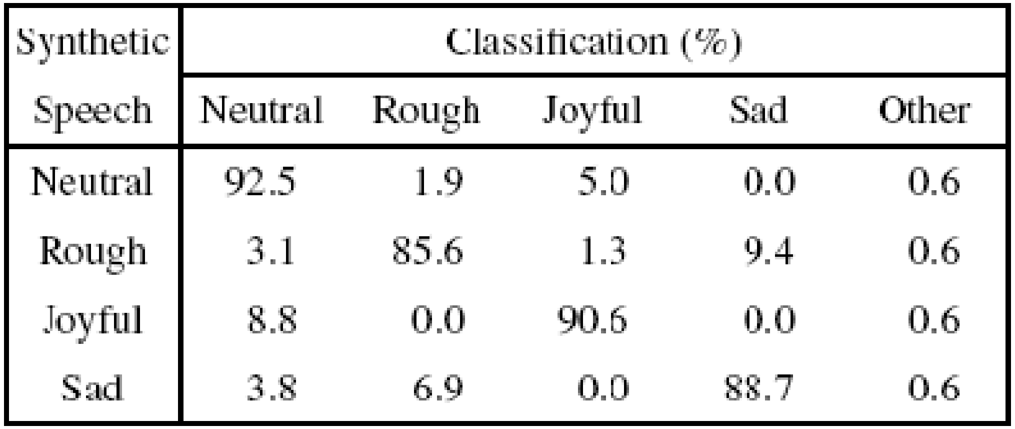
(b) Style-Mixed Model.



Synthetic	Classification (%)				
Speech	Neutral	Rough	Joyful	Sad	Other
Neutral	98.9	0.0	0.0	0.0	1.1
Rough	2.8	89.8	0.0	1.1	6.3
Joyful	0.6	0.0	96.0	0.0	3.4
Sad	0.0	0.6	0.0	96.0	3.4

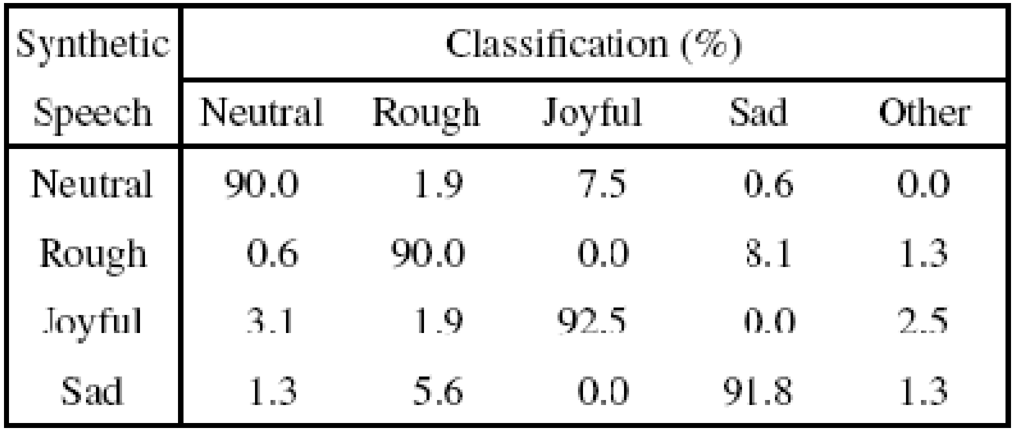
FTY

(a) Style-Dependent Model.



Synthetic	Classification (%)				
Speech	Neutral	Rough	Joyful	Sad	Other
Neutral	92.5	1.9	5.0	0.0	0.6
Rough	3.1	85.6	1.3	9.4	0.6
Joyful	8.8	0.0	90.6	0.0	0.6
Sad	3.8	6.9	0.0	88.7	0.6

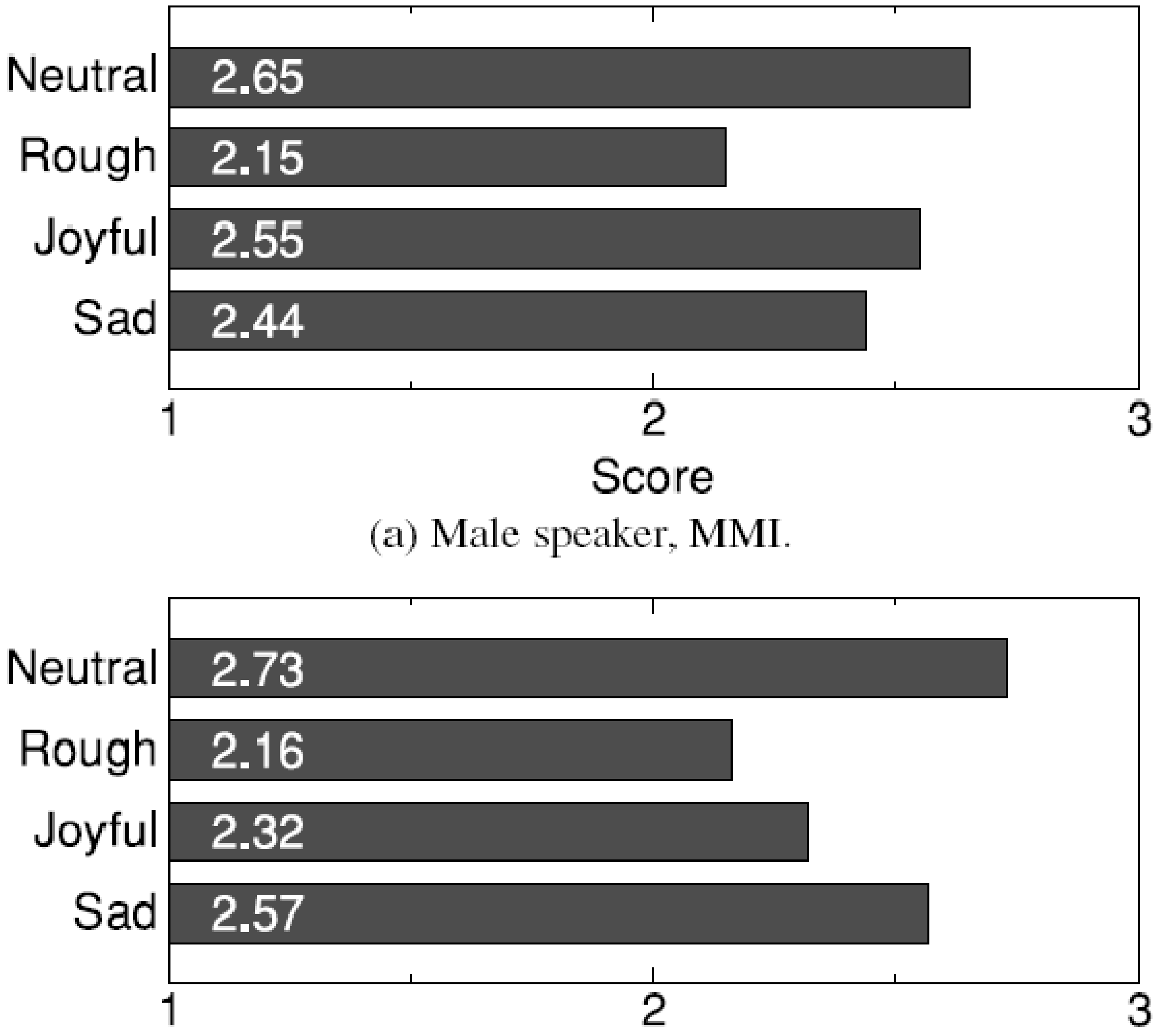
(b) Style-Mixed Model.



Synthetic	Classification (%)				
Speech	Neutral	Rough	Joyful	Sad	Other
Neutral	90.0	1.9	7.5	0.6	0.0
Rough	0.6	90.0	0.0	8.1	1.3
Joyful	3.1	1.9	92.5	0.0	2.5
Sad	1.3	5.6	0.0	91.8	1.3

Subjective Evaluations of Naturalness

- We conducted a subjective evaluation test to rate the naturalness of the speech synthesized by using the **style-dependent model**.
- Ten subjects listened to **eight** sentences chosen randomly from **53** test sentences and then they rated the naturalness of the synthesized speech.
- A 3-point scale was used with **3 for** “**good**”, **2 for** “**acceptable**”, and **1 for** “**bad**”.



(a) Male speaker, MMI.

(b) Female speaker, FTY.

Fig. 4 Subjective evaluation of naturalness of speech synthesized using style-dependent modeling.

Subjective Evaluations of Naturalness

- **Sixteen male** subjects were presented, in random order, with a pair of same-style speech samples synthesized using the two models, and then **they were asked which synthesized speech sounded more natural**.

- For each subject, **four** test sentences were chosen at random from **53** test sentences not included in the training data.

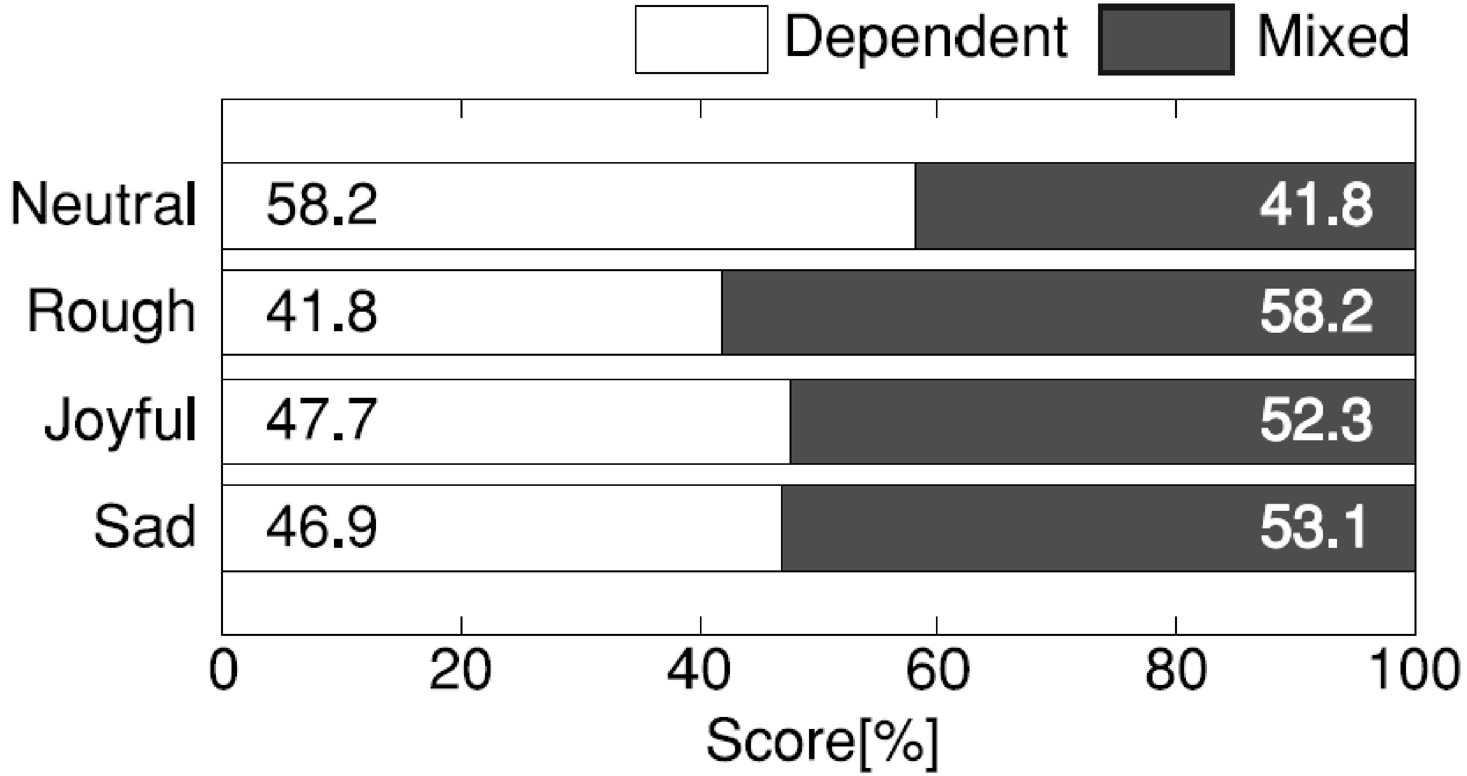


Fig. 5 Paired comparison test to assess the naturalness of synthesized speech generated using the style-dependent and style-mixed models for the male speaker, MMI.