Voice Conversion Using Duration-Embedded Bi-HMMs for Expressive Speech Synthesis

Source: TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING

Author : Chung-Hsien Wu, Chi-Chun Hsia, Jhing-Fa Wang

Professor : 陳嘉平

Reporter : 楊治鏞

Introduction

- Voice conversion methods have previously been proposed to convert the speech signals uttered by a speaker to the other speaker with limited speech data.

- With a limited size of speech database, this study proposes a spectral and a prosodic voice conversion model as the post-process of the TTS system for expressive speech synthesis.

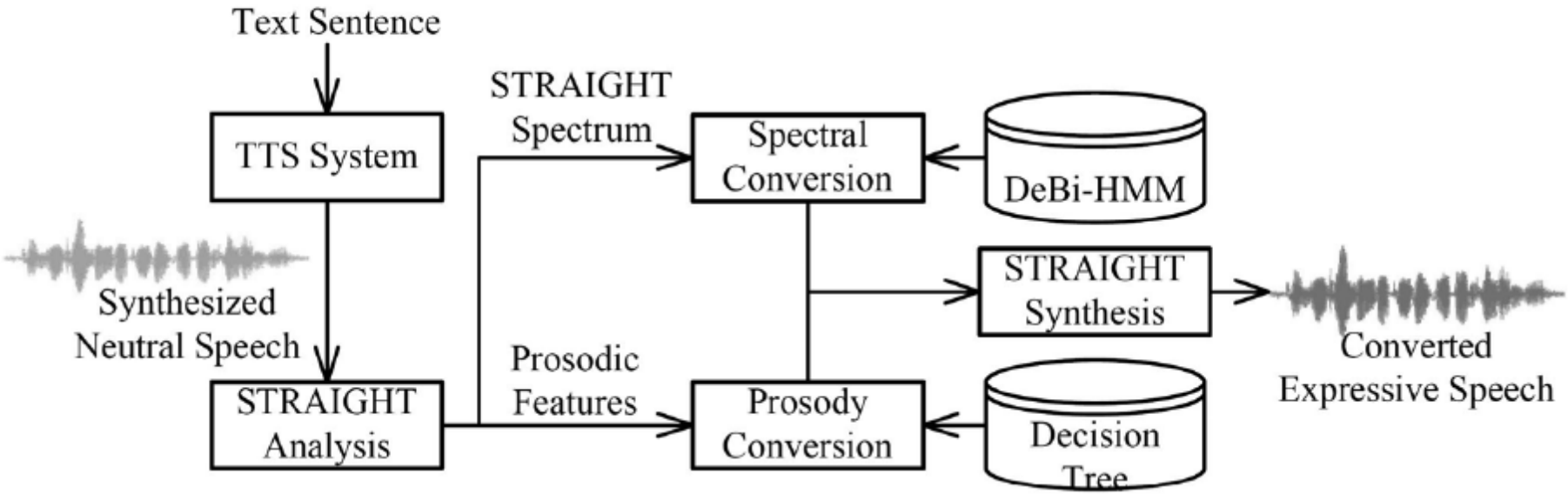


Fig. 1. Architecture of the proposed expressive speech synthesis system.

PROPOSED CONVERSION METHOD

- The Bi-HMM is adopted to model the spectral envelope evolution of source and target speech, separately but synchronously.

- The conversion function for each state is estimated with MMSE criterion under the conditional normal assumption.

Bi-HMM Spectral Conversion Model

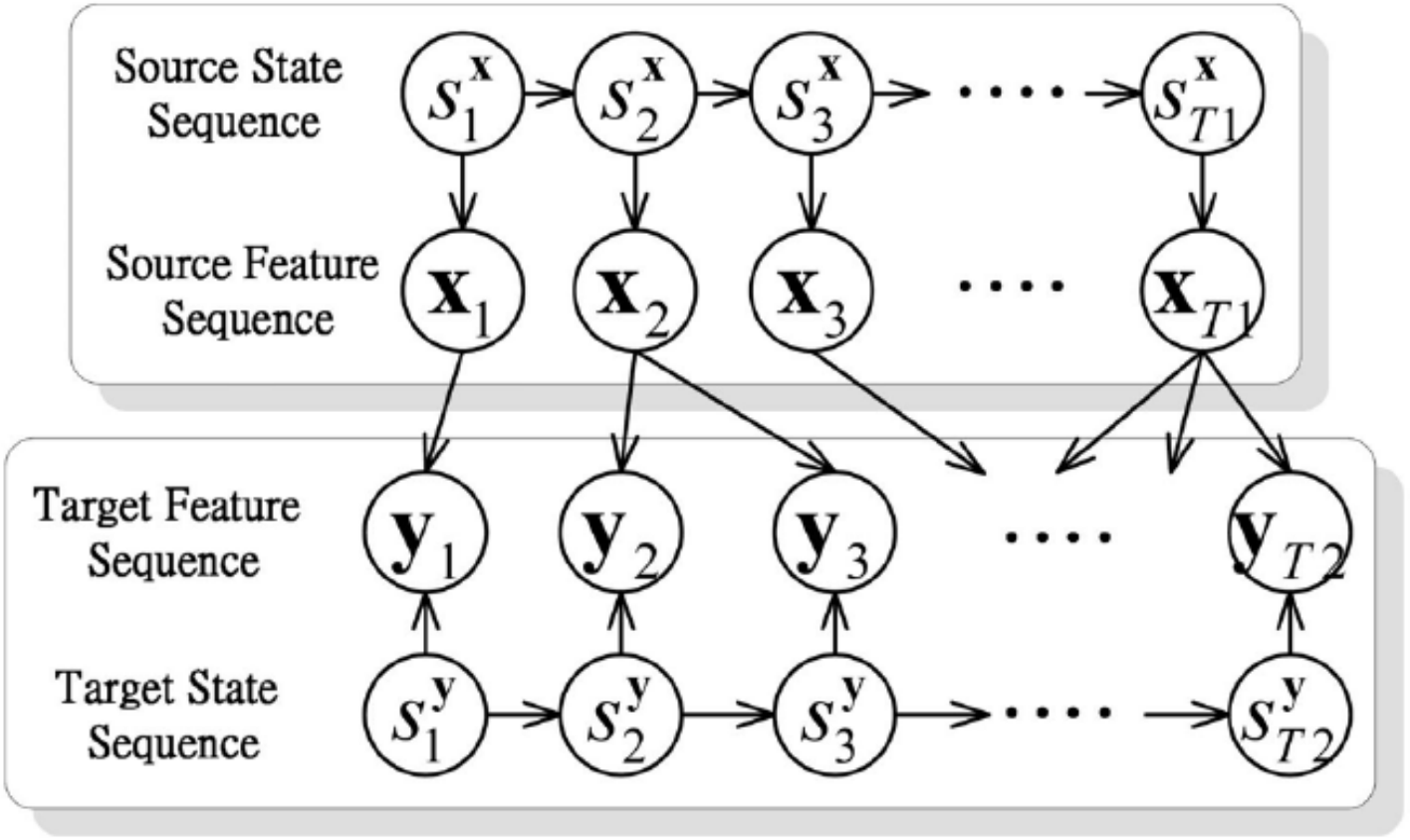


Fig. 2. State alignment between the source and target HMMs which compose the Bi-HMM.

Bi-HMM Spectral Conversion Model

- The upper portion depicts the relationship between the source feature vector sequence $\mathrm{X} = \{\mathrm{x}_1, \mathrm{x}_2, \ldots, \mathrm{x}_{T1}\}$ with length $T1$ and its corresponding state sequence $S_{\mathrm{X}} = \{s_{\mathrm{x}_1}, s_{\mathrm{x}_2}, \ldots, s_{\mathrm{x}_{T1}}\}$.

- The lower portion depicts the relationship between the target feature vector sequence $\mathrm{Y} = \{\mathrm{y}_1, \mathrm{y}_2, \ldots, \mathrm{y}_{T2}\}$ and its corresponding state sequence $S_{\mathrm{y}} = \{s_{\mathrm{y}_1}, s_{\mathrm{y}_2}, \ldots, s_{\mathrm{y}_{T2}}\}$.

Bi-HMM Spectral Conversion Model

- For a given state $s_{\mathrm{X}} = i$ and an input source feature vector x, the converted target vector $\tilde{\mathrm{y}}$ is predicted using the following conversion function:

$$\tilde{\mathrm{y}} = F_{s_{\mathrm{X}}=i}(\mathrm{x}) = \sum_{j=1}^{J} \left\{ p_i(j \mid \mathrm{x}, \Lambda) \times \left[\mu_{i,j}^{\mathrm{Y}} + \Sigma_{i,j}^{\mathrm{YX}} \left(\Sigma_{i,j}^{\mathrm{XX}} \right)^{-1} \left(\mathrm{x} - \mu_{i,j}^{\mathrm{X}} \right) \right] \right\}$$

- where J denotes the number of mixture components; $\mu_{i,j}^{\mathrm{X}}$ and $\mu_{i,j}^{\mathrm{Y}}$ denote the mean vectors of source and target feature vector sequences in mixture j of state i .

Bi-HMM Spectral Conversion Model

- $p_i(j \mid \mathrm{x}, \Lambda)$ represents the conditional probability where belongs to the mixture j in state i and is estimated as

$$p_i(j \mid \mathrm{x}, \Lambda) = \frac{w_{i,j} N\left(\mathrm{x}; \mu_{i,j}^{\mathrm{X}}; \mu_{i,j}^{\mathrm{XX}}\right)}{\sum_{k=1}^{J} w_{i,k} N\left(\mathrm{x}; \mu_{i,k}^{\mathrm{X}}; \mu_{i,k}^{\mathrm{XX}}\right)}$$

- where $w_{i,j}$ denotes the weight of mixture j in state i .

Bi-HMM Spectral Conversion Model

- Given the Bi-HMM parameter set $\Lambda = \{\Lambda_x, \Lambda_y\}$ and the source and target feature vector sequences X and Y, the state sequences S_X and S_Y with maximum joint probability $p(\mathrm{Y}, S_\mathrm{Y}, \mathrm{X}, S_\mathrm{X} \mid \Lambda)$ is obtained as follows:

$$
\begin{aligned}
\left(S_\mathrm{Y}^*, S_\mathrm{X}^*\right) &= \arg\max_{S_\mathrm{Y}, S_\mathrm{X}} p\left(\mathrm{Y}, S_\mathrm{Y}, \mathrm{X}, S_\mathrm{X} \mid \Lambda\right) \\
&= \arg\max_{S_\mathrm{Y}, S_\mathrm{X}} p\left(\mathrm{Y}, S_\mathrm{Y} \mid \mathrm{X}, S_\mathrm{X}, \Lambda_\mathrm{Y}\right) p\left(\mathrm{X}, S_\mathrm{X} \mid \Lambda_\mathrm{X}\right) \\
&= \arg\max_{S_\mathrm{Y}, S_\mathrm{X}} p\left(\mathrm{Y} \mid \mathrm{X}, S_\mathrm{X}, S_\mathrm{Y}, \Lambda_\mathrm{Y}\right) \\
&\qquad \times p\left(S_\mathrm{Y} \mid \Lambda_\mathrm{Y}\right) p\left(\mathrm{X} \mid S_\mathrm{X}, \Lambda_\mathrm{X}\right) p\left(S_\mathrm{X} \mid \Lambda_\mathrm{X}\right)
\end{aligned}
$$

Bi-HMM Spectral Conversion Model

- Given the input X , a candidate target $\tilde{\mathrm{Y}}$ is predicted for a specific state sequence S_{X} . The optimal state sequence pair $\left(S_{\mathrm{X}}^{*}, S_{\tilde{\mathrm{Y}}}^{*}\right)$ in the conversion phase is obtained by

$$\left(S_{\mathrm{X}}^{*}, S_{\tilde{\mathrm{Y}}}^{*}\right) = \arg\max_{S_{\mathrm{Y}}, S_{\mathrm{X}}} p\left(\tilde{\mathrm{Y}} \mid S_{\tilde{\mathrm{Y}}}, \Lambda_{\mathrm{Y}}\right) p\left(S_{\tilde{\mathrm{Y}}} \mid \Lambda_{\mathrm{Y}}\right)$$

$$\times p\left(\mathrm{X} \mid S_{\mathrm{X}}, \Lambda_{\mathrm{X}}\right) p\left(S_{\mathrm{X}} \mid \Lambda_{\mathrm{X}}\right)$$

Bi-HMM Spectral Conversion Model

- The target vector sequence with maximum probability in Bi-HMM is obtained by the following conversion function:

$$\tilde{\mathrm{Y}} = F_{s_{\mathrm{X}}^{*}}(\mathrm{X}) = \left\{ F_{s_{\mathrm{x}_1}^{*}}(\mathrm{x}_1), F_{s_{\mathrm{x}_2}^{*}}(\mathrm{x}_2), \ldots F_{s_{\mathrm{x}_{\mathrm{T}1}}^{*}}(\mathrm{x}_{\mathrm{T}1}) \right\}$$

Embedded Duration Model

- When the speech signal stays in state i, $i = 1, \ldots, I$ with self-transition probability a_{ii}, for τ frames, the implicit duration probability density is a geometric distribution $d_i(\tau) = a_{ii}^{\tau-1}(1 - a_{ii})$.

- This exponential state duration density is inappropriate for most speech signals.

Embedded Duration Model

- In this study, a Gamma duration model is adopted as follows:

$$d_i\left(\tau \mid \eta_i, \nu_i\right) = \frac{\eta_i^{\nu_i}}{\Gamma(\nu_i)} \tau^{\nu_i - 1} \exp\left(-\eta_i \tau\right)$$

- where $\Gamma(\cdot)$ is gamma function; $\eta_i > 0$ and $\nu_i > 0$ are parameters of the gamma distribution.

Embedded Duration Model

- The E-step calculates the expectation of the log likelihood of the complete data given new estimates, $\hat{\eta}_i$ and $\hat{\nu}_i$, after having the following current estimates, given by η_i and ν_i

$$Q\left(\hat{\eta}_i, \hat{\nu}_i \mid \eta_i, \nu_i\right) = \sum_{t=1}^{T} \left\{ \xi_{t \in t_s}(i) \right.$$

$$\cdot \left[-\log \Gamma\left(\hat{\nu}_i\right) + \hat{\nu}_i \log \hat{\eta}_i + \left(\hat{\nu}_i - 1\right) \log \tau_t - \hat{\eta}_i \tau_t \right]$$

- where t_s is the starting frame of a state, T represents the total number of frames for a training utterance

Embedded Duration Model

$$\xi_{t \in t_s}(i) = \delta(s_t - i)\,\delta(t - t_s) = P\left(s_{t \in t_s} = i \mid \mathrm{X}, \Lambda\right)$$

where $\delta(\cdot)$ is the Kronecker delta function.

- Unfortunately, no closed-form solution to new estimate was derived.
- Since gamma distribution $d_i(\tau \mid \eta_i, \nu_i)$ contains the mean ν_i / η_i and variance ν_i / η_i^2, the parameters are then derived empirically from the sample mean and variance.

Prosody Conversion by Decision Tree

- The prosody conversion decision tree (PDT) is derived from the prosodic features of a syllable.
- This work adopted the ratio of syllable duration, pitch mean, pitch dynamic range, energy mean, and energy dynamic range for prosody conversion.
- Table 1 shows the linguistic question set used to train the decision tree, which includes the linguistic features in word and sentence levels.

Linguistic Question Set Used to Train the Decision Tree for Prosody Conversion

Word level features	Tone type of the syllable
	Word length represented in syllable number
	Part of speech of the word
	Punctuation marks ahead and behind the word
Sentence level features	Word position in the phrase (front, middle, end)
	Number of phrases in the sentence
	Word position in the sentence

EXPERIMENTS

- A total number of distinct tonal syllables of 1410 were used as the basic units to synthesize neutral speech.

- Six expressive styles were applied, namely happiness, sadness, anger, confusion, apology, and question.

STATISTICS OF EXPRESSIVE SPEECH DATABASES

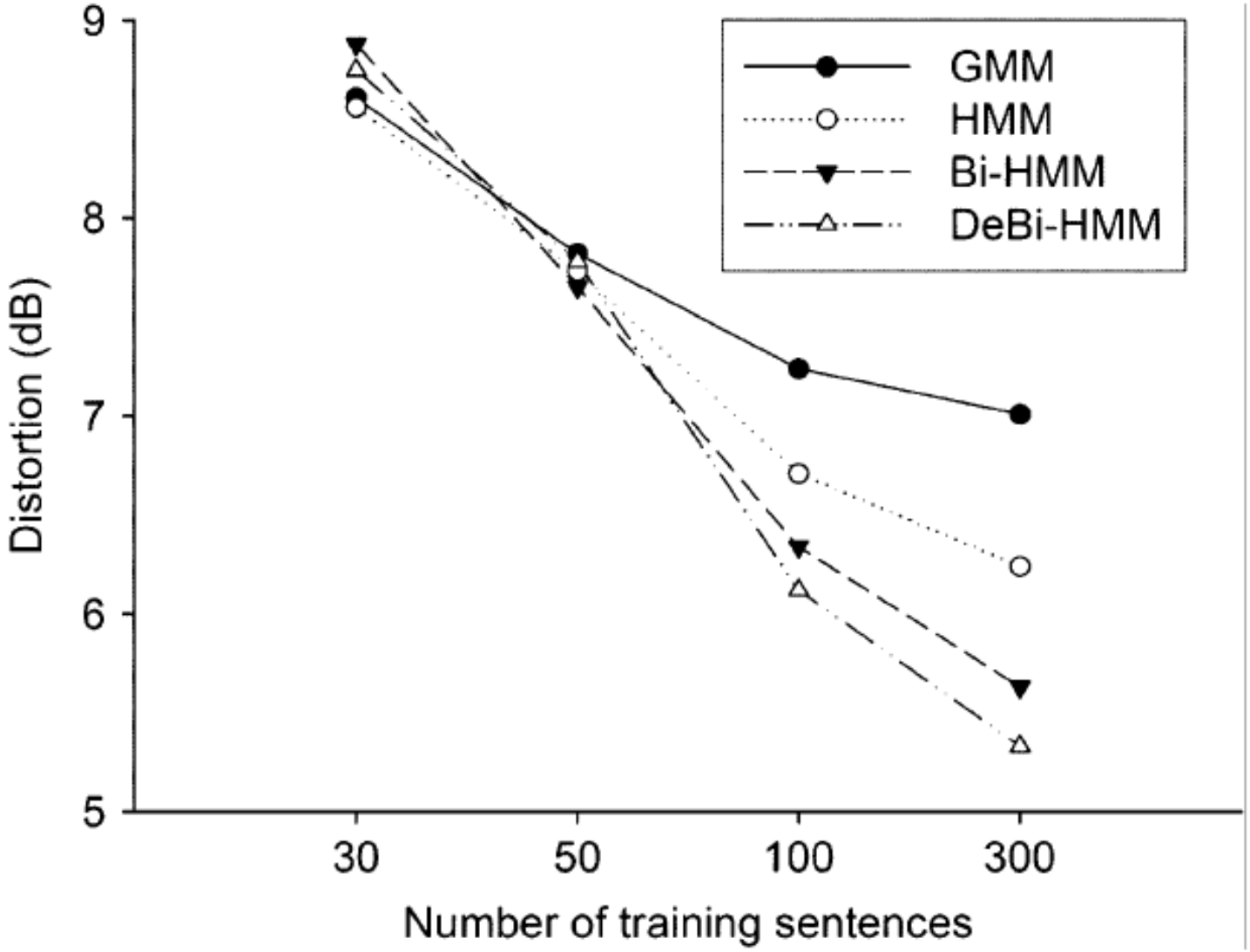
Expressive Style	Happiness				Sadness				Anger			
Number of sentences	30	50	100	300	30	50	100	300	30	50	100	300
Avg. appearance number of context indep. sub-syllables	14.24	21.81	47.00	124.62	16.24	24.10	41.78	136.33	12.28	22.52	46.28	119.53
Number of syllables	459	680	1420	3930	501	760	1310	4140	369	680	1408	3637
Number of words	270	395	855	2246	277	425	728	2262	216	386	776	1951
Length (min:sec)	1:31	2:28	4:56	15:15	1:46	3:02	5:51	17:41	1:11	1:58	4:02	11:54
Expressive Style	Confusion				Question				Apology			
Number of sentences	30	50	100	300	30	50	100	300	30	50	100	300
Avg. appearance number of context indep. sub-syllables	13.07	21.19	42.03	113.21	13.95	22.33	43.66	123.50	11.50	20.14	35.48	131.86
Number of syllables	393	630	1290	3420	426	685	1312	3750	342	602	1072	3963
Number of words	244	366	705	1839	241	374	737	1963	212	381	656	2359
Length (min:sec)	1:25	2:23	4:46	14:06	1:09	1:56	3:46	12:31	1:15	2:10	4:08	13:12

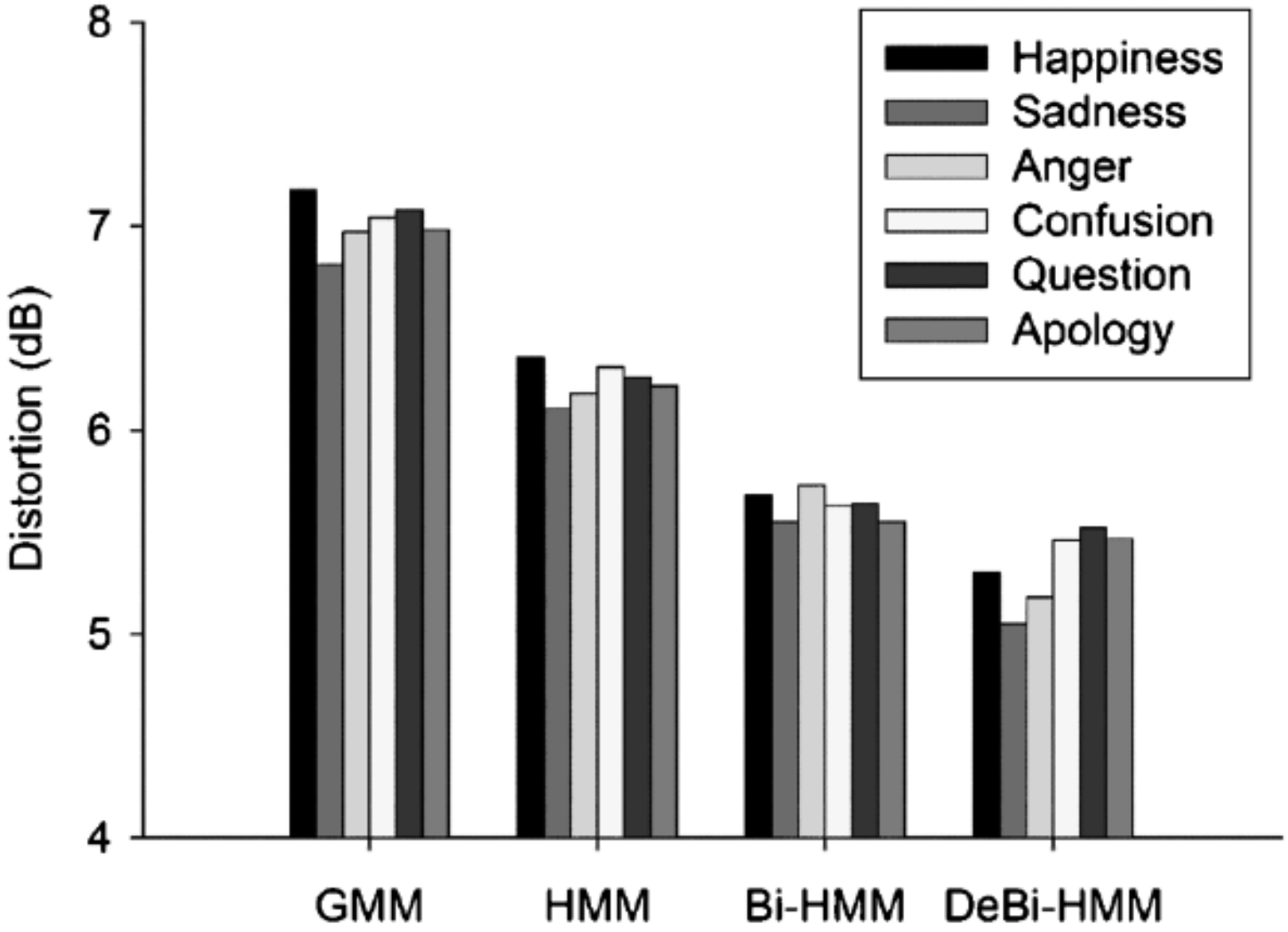
Objective test

- Distortion between the converted (from neutral) and the target (happy, sad and angry) speech is calculated using the following equation:

$$Avg\_Dis = \frac{20}{\ln 10}\left(\frac{1}{T}\sum_{t=1}^{T}\left\|2\left(y_t - \hat{y}_t\right)\right\|\right)$$

- where y_t and $\hat{y}_t$ denote the target and the spectrum-converted feature vectors at time.

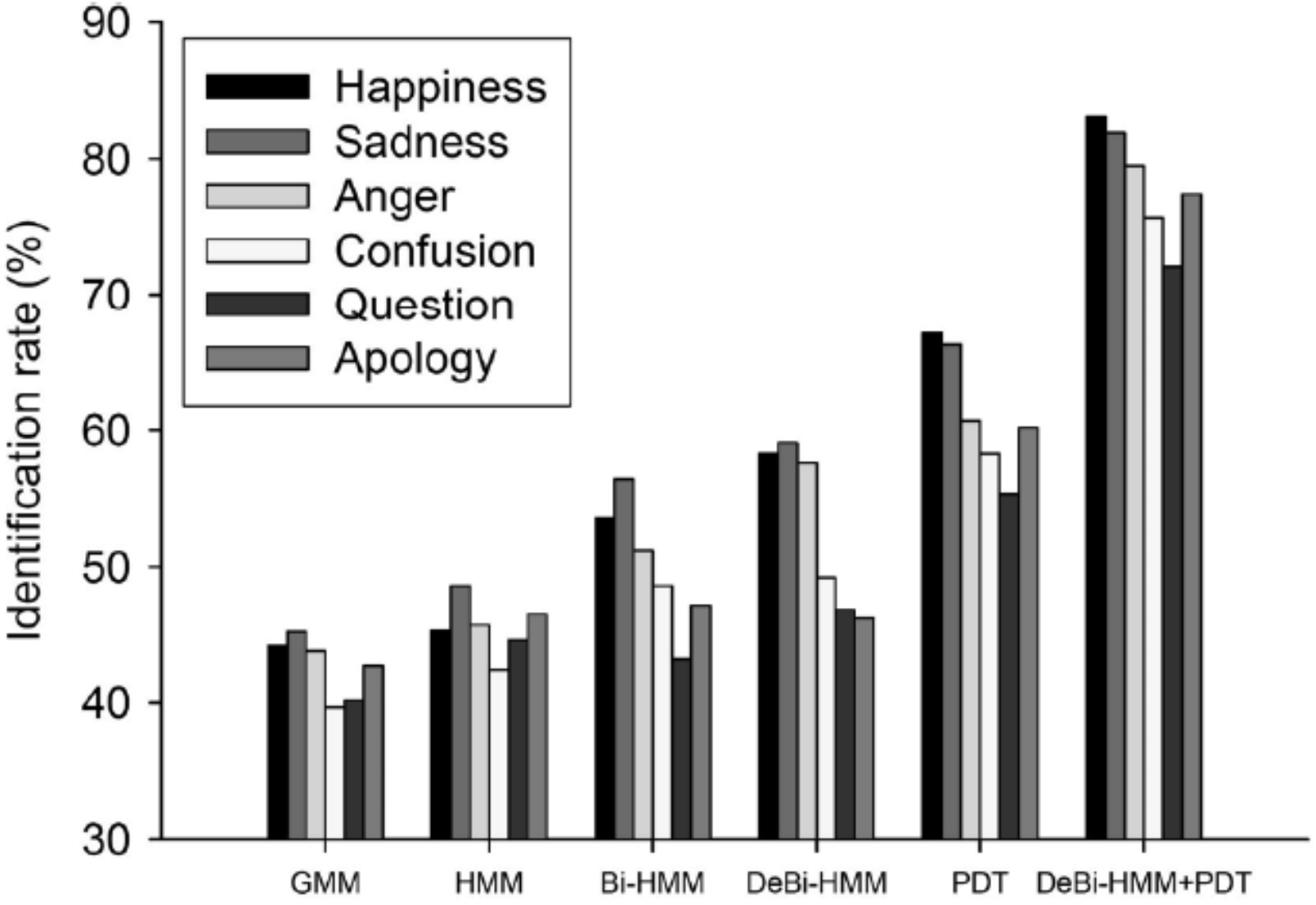

GMM
HMM
Bi-HMM
DeBi-HMM
Distortion (dB)
9
8
7
6
5
30
50
100
300
Number of training sentences


8
7
6
5
4
Distortion (dB)
Happiness
Sadness
Anger
Confusion
Question
Apology
GMM
HMM
Bi-HMM
DeBi-HMM

Subjective Test

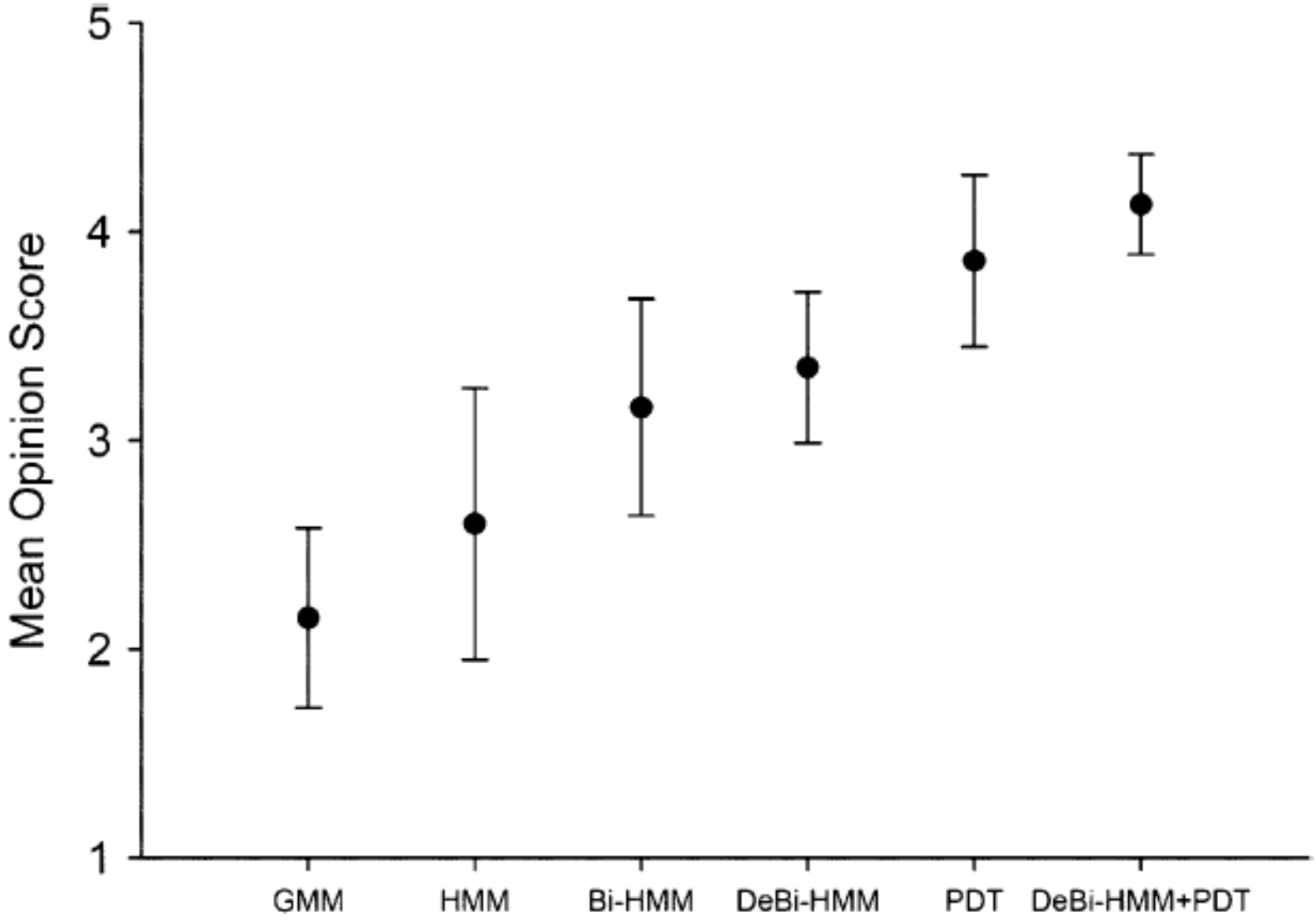
- For each test sentence randomly selected from the test set, 36 converted utterances processed by each conversion method to each expressive style were randomly output to the human subjects.

- Twenty adult subjects, around 22–31 years of age, were asked to classify each utterance as one of the six expressive styles.


Identification rate (%)
90
80
70
60
50
40
30
Happiness
Sadness
Anger
Confusion
Question
Apology
GMM
HMM
Bi-HMM
DeBi-HMM
PDT
DeBi-HMM+PDT

Subjective Test

- The naturalness of the converted utterance was also evaluated, according to a 5-scale scoring method excellent very poor.


Mean Opinion Score
5
4
3
2
1
GMM
HMM
Bi-HMM
DeBi-HMM
PDT
DeBi-HMM+PDT