EURASIP Journal on Audio, Speech, and Music Processing
a SpringerOpen Journal

RESEARCH Open Access

Speaker-dependent model interpolation for statistical emotional speech synthesis

Chih-Yu Hsu[1] and Chia-Ping Chen[1*]

Abstract

In this article, we propose a speaker-dependent model interpolation method for statistical emotional speech synthesis. The basic idea is to combine the neutral model set of the target speaker and an emotional model set selected from a pool of speakers. For model selection and interpolation weight determination, we propose to use a novel monophone-based Mahalanobis distance, which is a proper distance measure between two Hidden Markov Model sets. We design Latin-square evaluation to reduce the systematic bias in the subjective listening tests. The proposed interpolation method achieves sound performance on the emotional expressiveness, the naturalness, and the target speaker similarity. Moreover, such performance is achieved without the need to collect the emotional speech of the target speaker, saving the cost of data collection and labeling.

Introduction

Statistical speech synthesis (SSS) is a fast-growing research area for text-to-speech (TTS) systems. While a state-of-the-art concatenative method [1,2] for TTS is capable of synthesizing natural and smooth speech for a specific voice, an SSS-based approach [3,4] has the strength to produce a diverse spectrum of voices without requiring significant amount of new data. This is an important feature for building next-generation applications such as a story-telling robot capable of synthesizing the speech of multiple characters with different emotions, personalized speech synthesis such as in speech-to-speech translation [5,6], and clinical applications such as voice reconstruction of patients with speech disorders [7]. In this article, we study the problem of generating new models of SSS from existing models. The model parameters of SSS can be systematically modified to express different emotions. Many instances of this problem have been investigated in the literature. In [8], the prosody is mapped from neutral to emotional using Gaussian mixture models and classification and regression trees. In [9], the spectrum and duration are converted in a voice conversion system with duration-embedded hidden Markov models (HMMs). In [10,11], style-dependent and style-mixed modeling methods for emotional expressiveness are investigated. In [12], adaptation methods are used to transform the neutral model to the target model, requiring only small amounts of adaptation data. In [13-15], simultaneous adaptation of speaker and style is applied to an average voice model of multiple-regression HMMs to synthesize speaker-dependent styled speech. A few methods without the requirement of target speaker's emotional speech have been studied. In [16], neutral speech are adapted based on analysis of emotional speech from the prosodic point of view. In [17,18], speech with emotions or mixed styles are generated by interpolating styled speech models trained independently. In [19], prosodic parameters including pitch, duration, and strength factors are adjusted to generate emotional speech from neutral voice.

The method that we propose for emotional SSS models is called the speaker-dependent model interpolation. By being speaker-dependent, we mean that the interpolating model sets and weights are dependent on the speaker identity. By model interpolation, we mean that the target synthesis model set is a convex combination of multiple synthesis model sets. One major difference between our approach and the previous approaches for emotional expressiveness is that the emotional speech directly from the target speaker is not required by our design. This feature is particularly attractive when the collection of target emotional speech is difficult or even infeasible.

This article is organized as follows. First, we introduce our HMM-based speech synthesis system in

*Correspondence: cpchen@cse.nsysu.edu.tw
[1] Department of Computer Science and Engineering, National Sun Yat-Sen University, 70 Lien-Hai Road, Kaohsiung 800, Taiwan

Section “HMM-based speech synthesis”. The proposed method for emotional expressiveness based on speaker-dependent model interpolation is described in Section “Interpolation methods”. The evaluation methods and the results of the proposed approach are presented in Section “Experiments”. Lastly, the concluding remarks are given in Section “Conclusion and future work”.

HMM-based speech synthesis

An HMM-based speech synthesis system (also known as HTS) models speech units as HMMs [20]. An HTS system uses parameters of the multi-stream HMMs structure which combine the spectrum and excitation to generate the speech feature sequence, and uses a vocoder to convert the feature sequence to speech waveforms [21]. The parameters of the HMMs are learned in the training stage with labeled speech data via expectation-maximization algorithm [22,23], as is well known and commonly used in machine learning and automatic speech recognition.

The block diagram of an HTS system is shown in Figure 1. The spectral features are modeled by HMM with a single Gaussian per state, while the excitation features are modeled by the multi-space probability distribution HMM (MSD-HMM) [24] to deal with the off-and-on property of periodic excitation. The duration of an HMM state is modeled by a Gaussian random variable, whose parameters are estimated by the state occupancies estimated on the training data. In the synthesis phase, given the input text, the corresponding state sequence is decided by maximizing the overall duration probability. With the state sequence, the static spectral and excitation features are determined by maximizing the joint data-likelihood of the combined static and dynamic feature streams. Finally, a synthesis filter is used to synthesize the speech samples.

Our system is based on HTS version 2.1. We use the mel-generalized cepstral coefficients [25] ($\alpha = 0.42, \gamma = 0$) as the spectral features, and use the logarithm of the fundamental frequency ($\log F_0$) as the excitation feature. The hts_engine API version 1.02 is used to synthesize speech waveforms from trained HMMs via a mel-generalized log-spectral approximation filter [26].

An HTS system for continuous Mandarin speech synthesis is constructed for this research. In this system, the basic HMM units are the tonal phone models (TPMs) [27]. The TPMs are based on the well-known initial/final models. In order to model the *tones* in Mandarin, we include two or three variants of tonal-final models based on the pitch positions (High, Medium, Low). In order to model the transition of pitch position during a syllable final, we concatenate an initial model and two tonal-final models for a tonal syllable. In total, there are 53 initial models and 52 tonal-final models. Thus, there are 105 “monophones” in our acoustic model. The initial models and the tonal-final models are listed in Table 1. The context-dependent phones (called the full-context phones) are based on these monophones. The question set for training the decision trees for state-tying consists of questions on the tonal context, the phonetic context, the syllabic context, the word context, the phrasal context, and the utterance context [28].

Interpolation methods

Background

In this article, the target model set[a] for SSS is obtained through interpolation. Model interpolation offers two distinctive advantages. First, the data collection cost is

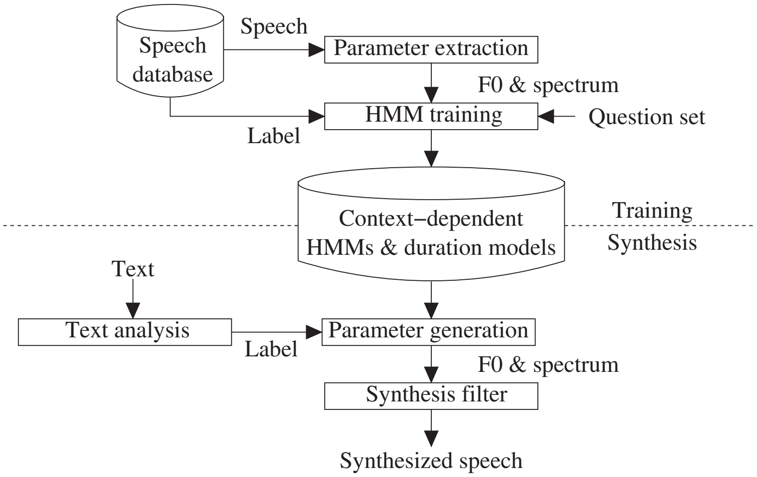


Figure 1 The block diagram of a general HMM-based speech synthesis system. In the training phase we train the parameters in the HMMs given a labeled speech data set. In the synthesis phase, we use the model parameters to generate speech given text.

Table 1 The TPMs for Mandarin with the example 太棒了 (wonderful)

initial models	ch chi chiu chu d di du f g gu h hu j ji jiu ju k ku l li liu lu m mi n nu ni p pi ru sh shi shiu shu su t ti tsu tu tzu wu niu b bi jr chr shr r sz tsz tz yi yu			
tonal-final models	a{H,L,M} ai{H,L} ang{H,L} an{H,L} au{H,L} chr{H,L} e{H,L,M} eh{H,L} ei{H,L} en{H,L,M} er{H,L} jr{H,L} ng{H,L} o{H,L,M} ou{H,L,M} r{H,L} shr{H,L} sz{H,L} tsz{H,L} tz{H,L,M} wu{H,L} yi{H,L} yu{H,L}			
example		太	棒	了
	tonal syllable model	tai(4)	bang(4)	le(0)
	tonal phone model	t aiH aiL	b angH angL	l eM eM

reduced compared to retraining or model adaptation. Second, the properties of the synthesized speech can be refined by incrementally adjusting the interpolation ratios. These features are analyzed in the following sections.

Data collection costs

Suppose we need to synthesize the voices of S different speakers, and each speaker has E non-neutral emotions. Let the data required for training an SSS model set be D_t, and for adapting a model set be D_a, where we often have $D_a < D_t$. To synthesize all emotional voices of all speakers, retraining (training each model set from scratch) requires collecting and labeling a dataset of the size of

$$S \times D_t + SE \times D_t, \tag{1}$$

model adaptation requires the amount of

$$S \times D_t + SE \times D_a, \tag{2}$$

and our interpolation-based approach requires the amount of

$$S \times D_t + LE \times D_t, \tag{3}$$

where L is the number of model sets in the pool for each emotion (cf. Section "Speaker-dependent model interpolation"). The difference among (1), (2), and (3) is the marginal amount of data required per new speaker. For comparison, Equation (1) requires the amount of $D_t + E \times D_t$ to train a neutral model set and E different emotional model sets, Equation (2) requires the amount of $D_t + E \times D_a$ to train a neutral model set for the new speaker and then adapt for E different emotions, while Equation (3) requires only the amount of D_t to train a new neutral model set. Clearly, the unpleasant efforts of data collection and labeling are significantly reduced with the proposed interpolation approach.

Applications

Model interpolation has been applied to continuously display a full spectrum of specific attributes for the synthesized speech, such as gender [29], style [17], and accent [30]. In this article, we study three emotional attributes including angry, happy, and sad.

HMM interpolation

Given model sets $M_1, \ldots, M_K$, where K is the number of available model sets, a model set $\tilde{M}$ can be created by interpolation as follows. Let $\mu_k(i)$ be the mean vector of state i in M_k, $\Sigma_k(i)$ be the covariance matrix, and $a_k(i)$ be the interpolation weights of M_k for state i. Then, the interpolated model is

$$\tilde{\mu}(i) = \sum_{k=1}^{K} a_k(i)\mu_k(i), \quad \tilde{\Sigma}(i) = \sum_{k=1}^{K} a_k^2(i)\Sigma_k(i), \tag{4}$$

where $\tilde{\mu}(i)$ and $\tilde{\Sigma}(i)$ are the mean and covariance of state i of $\tilde{M}$.

There are a few methods to decide the interpolation weights $a_k(i)$'s in (4). In [29], they are tied across states and then be decided by minimizing the weighted sum of the Kullback–Leibler distance. In [17], $a_k(i)$ is assumed to be proportional to $\gamma_k(i)$, the occupancy counts of state i in M_k (with respect to a training set)

$$a_k(i) = \frac{1}{c}\gamma_k(i), \quad c = \sum_{k'=1}^{K} \gamma_{k'}(i), \quad k = 1, \ldots, K. \tag{5}$$

The block diagram of model interpolation is shown in Figure 2.

Monophone-based Mahalanobis distance measure

In our approach, we need a distance measure to find the closest model set and the interpolation weight. Since two context-dependent phone model sets generally have different state-tying structures (decision trees), we propose a measure based on the context-independent models. It is called the monophone-based Mahalanobis distance (MBMD) defined by

$$d_M(\alpha, \beta) = \sum_{i \in Z} \left(\mu_\alpha(i) - \mu_\beta(i)\right)^T \left(\frac{\Sigma_\alpha(i) + \Sigma_\beta(i)}{2}\right)^{-1} \times \left(\mu_\alpha(i) - \mu_\beta(i)\right). \tag{6}$$

In (6), α and β are HMM model sets, Z is the set of monophone HMM states, and $\mu_\alpha(i)$ and $\Sigma_\alpha(i)$ are the mean vector and the covariance matrix of the Gaussian of state i in model set α (similarly for β).

Note that (6) defines a proper distance measure, since

$$d_M(\alpha, \beta) \geq 0, \quad d_M(\alpha, \beta) = 0 \quad \Leftrightarrow \quad \alpha = \beta. \tag{7}$$

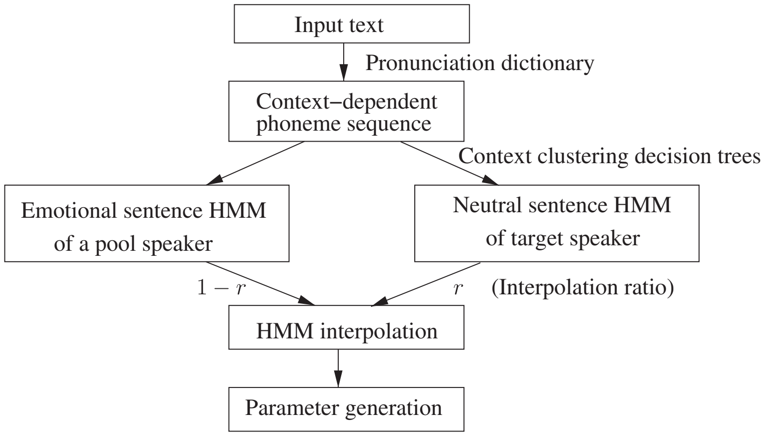

Input text
Pronunciation dictionary
Context–dependent phoneme sequence
Context clustering decision trees
Emotional sentence HMM of a pool speaker
Neutral sentence HMM of target speaker
1 − r
r
(Interpolation ratio)
HMM interpolation
Parameter generation

Figure 2 The block diagram of the procedure for model interpolation. Note that the interpolation is carried out with the model parameters.

Furthermore, the difference in the temporal structures of model sets α, β is *not* ignored in our method. That is, we include the Mahalanobis distance between the Gaussians modeling the duration of HMM states in (6). Thus, the difference in the state transition distributions of two model sets is taken into account.

Speaker-dependent model interpolation

We can now describe the proposed speaker-dependent model interpolation for SSS. Let the number of pool speakers be L. Let $\phi_1, \ldots, \phi_L$ denote the neutral model sets, and $\psi_1, \ldots, \psi_L$ denote the emotional model sets. These model sets are trained by HTS. Note that as mentioned in Section "HMM-based speech synthesis", certain features of HTS have to be customized for Mandarin, such as the phone set and the question set for full-context phone models. Given a target speaker $\mathcal{T}$ with neutral model set $\phi_{\mathcal{T}}$, an emotional model set $\psi_{\mathcal{T}}$ for $\mathcal{T}$ is created by interpolation with the following proposed Method of Model Interpolation.

Method of model interpolation

- Find the speaker whose neutral model set is closest to $\phi_{\mathcal{T}}$,

$$l^* = \arg\min_{l\in\{1,\ldots,L\}} d_M(\phi_{\mathcal{T}}, \phi_l). \tag{8}$$

- Find the emotional model set closest to $\phi_{\mathcal{T}}$,

$$\psi^* = \arg\min_{\psi_j\in\{\psi_1,\ldots,\psi_L\}} d_M(\phi_{\mathcal{T}}, \psi_j). \tag{9}$$

- Find the interpolation weight r^* which results in the model set closest to ψ_{l^*},

$$r^* = \arg\min_{r\in[0,1]} d_M(I_{\text{mono}}\left(\phi_{\mathcal{T}}, \psi^*, r\right), \psi_{l^*}). \tag{10}$$

  Note $I_{\text{mono}}\,(\phi_{\mathcal{T}}, \psi^*, r)$ is the monophone model set interpolation by $M_1 = \phi_{\mathcal{T}}$ and $M_2 = \psi^*$ with weights

$$a_1 = r, \quad a_2 = 1 - r. \tag{11}$$

  The interpolation weight is found by a grid search. That is, r is varied from 0 to 1 with increment 0.1 to find r^*.
- Given a text, the speech is synthesized by interpolating the parameters computed by ψ^* and $\phi_{\mathcal{T}}$ with weights r^* and $1 - r^*$. This is denoted by

$$\psi_{\mathcal{T}} = I(\phi_{\mathcal{T}}, \psi^*, r^*). \tag{12}$$

The basic idea behind using ψ_{l^*} as the target in determining the interpolation weight (10) is that similar neutral speech implies similar emotional speech. The motivation of making interpolation between ϕ_T and ψ^* is twofold. First, if ψ_{l*} were used directly to represent ψ_T, the target speaker identity would have been lost. Therefore, we interpolate the target speaker's speech model ϕ_T. Secondly, to achieve better speech quality, we find an emotional model ψ^* that is closest to the target speaker's model to interpolate, reducing potential artifacts resulting from interpolating different models. The block diagram of the proposed Method of Model Interpolation is shown in Figure 3. The interpolation scheme is applied separately to the spectrum, the excitation, and the duration models. We note that

- The interpolation ratio (12) used for the context-dependent models is estimated via MBMD (10), which is based on context-independent models. This approximation works for our system.
- For a given target speaker, if $\psi^* = \psi_{l^*}$, the optimal weight r^* in (10) would be 0, and we would have $\psi_{\mathcal{T}} = \psi_{l^*}$. That is, the model set of the target speaker would not have contributed to the synthesized voice. In our system, if the closest neutral model set and the

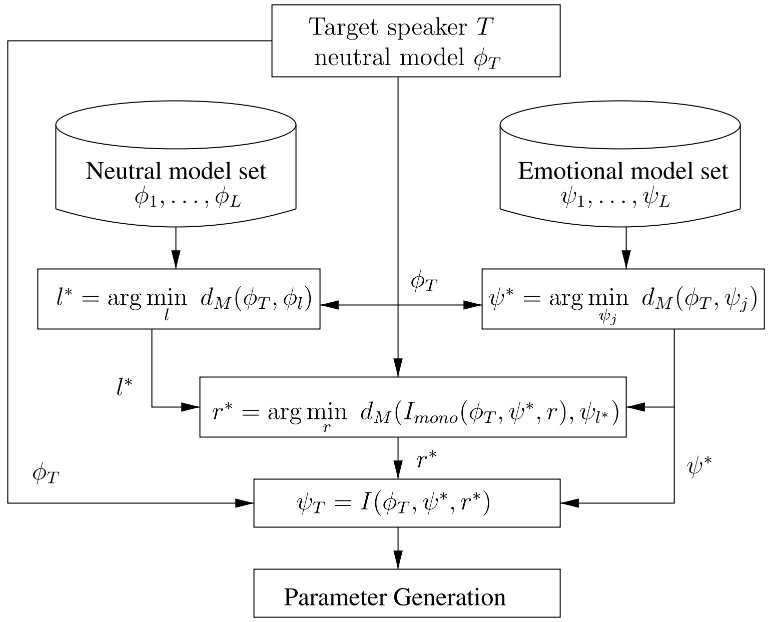

Target speaker T neutral model ϕ_T
Neutral model set $\phi_1, \ldots, \phi_L$
Emotional model set $\psi_1, \ldots, \psi_L$
$l^* = \arg\min_{l} \; d_M(\phi_T, \phi_l)$
ϕ_T
$\psi^* = \arg\min_{\psi_j} \; d_M(\phi_T, \psi_j)$
l^*
$r^* = \arg\min_{r} \; d_M(I_{mono}(\phi_T, \psi^*, r), \psi_{l^*})$
r^*
ϕ_T
ψ^*
$\psi_T = I(\phi_T, \psi^*, r^*)$
Parameter Generation

Figure 3 The block diagram of the proposed scheme for model interpolation. The interpolation is speaker-dependent.

closest emotional model set are from the same pool speaker, the system simply uses the second closest emotional model set.

Experiments

Speech data collection

We collect the speech of five pool speakers, including four male speakers and one female speaker. The speech samples of three emotions from each pool speaker, angry, happy, and sad, are collected, as well as the neutral (non-emotional) speech samples. There are three disjoint sets of sentences corresponding to the three different emotions. Each set consists of 300 sentences which are phonetically balanced by design. The set of neutral prompts is a subset of the union of the three disjoint sets. As a result, each pool speaker contributes 900 emotional utterances and 300 neutral utterances. The neutral model set of a pool speaker is trained by 300 neutral utterances, and each emotional model set is also trained by 300 emotional utterances. These model sets constitute the pools for $L = 5$ speakers and $E = 3$ emotions (cf. Section "Speaker-dependent model interpolation"). The MBMDs between the neutral models of the pool speakers are summarized in Table 2. It summarizes the distance (similarity) between pairs of neutral models of the pool speakers. Note that to achieve good speaker-independent interpolation performance, the set of pool speakers should be representative of the voice space, so the voices in the pool are better to be diverse.

In addition to the pool speakers, we also collect the speech of five target speakers, including one female speaker and four male speakers, for the evaluation of the proposed approach. The neutral speech samples of 300 utterances from each target speaker are collected and used to train the neutral speech model set of that target speaker. The MBMDs between the neutral models of the pool speakers and neutral models of the target speakers are summarized in Table 3. It is used to decide l^* in Equation (8), as shown in boldface in the table.

Evaluation on emotional expressiveness

For the emotional expressiveness, we evaluate whether the synthesized speech conveys the target emotion by subjective listening tests. In one selection, a listener listens to five waveforms (from five different target speakers in five different emotions in randomized pairs), and chooses the speech with the nominated emotion (1-out-of-5 choice of emotion identification). We note that this setting is significantly harder than a binary choice between a synthesized neutral speech and a synthesized emotional speech, which

Table 2 The MBMDs between the neutral model sets of the pool speakers, P1m to P5f (the gender is indicated in the speaker ID)

	P1m	P2m	P3m	P4m	P5f
P1m	0	15856	11127	10755	16776
P2m	15856	0	16936	22019	30851
P3m	11127	16936	0	19670	28008
P4m	10755	22019	19670	0	19150
P5f	16776	30851	28008	19150	0

Table 3 The MBMDs between the neutral model sets of the target speakers, T1m to T5f, and the pool speakers

	T1m	T2m	T3m	T4m	T5f
P1m	**10085**	**14047**	14101	13145	17590
P2m	19403	25005	22548	19129	27268
P3m	18496	19660	19841	14379	24677
P4m	12330	16427	**13234**	16549	17762
P5f	14550	26034	18835	22080	**10606**

could lead to a systematic bias toward the emotional speech to be selected.

The pairing of the target speaker and the emotion are randomized via a *Graeco-Latin square* commonly used in evaluation [31], as shown in Figure 4, for each emotion. With five listener groups and five utterance sets, each column of the 5 × 5 Graeco-Latin square corresponds to an utterance set and each row corresponds to a listener group. The entity in the intersection of row i and column j, denoted by (s_{ij}, e_{ij}), means that a listener in group i listens to utterance set j synthesized via the model set of target speaker s_{ij} for the emotion e_{ij}. We make sure that

$$\begin{aligned} s_{ij} &\neq s_{ij'} \quad \text{for all } j \neq j', \\ s_{ij} &\neq s_{i'j} \quad \text{for all } i \neq i', \end{aligned} \tag{13}$$

and

$$\begin{aligned} e_{ij} &\neq e_{ij'} \quad \text{for all } j \neq j', \\ e_{ij} &\neq e_{i'j} \quad \text{for all } i \neq i'. \end{aligned} \tag{14}$$

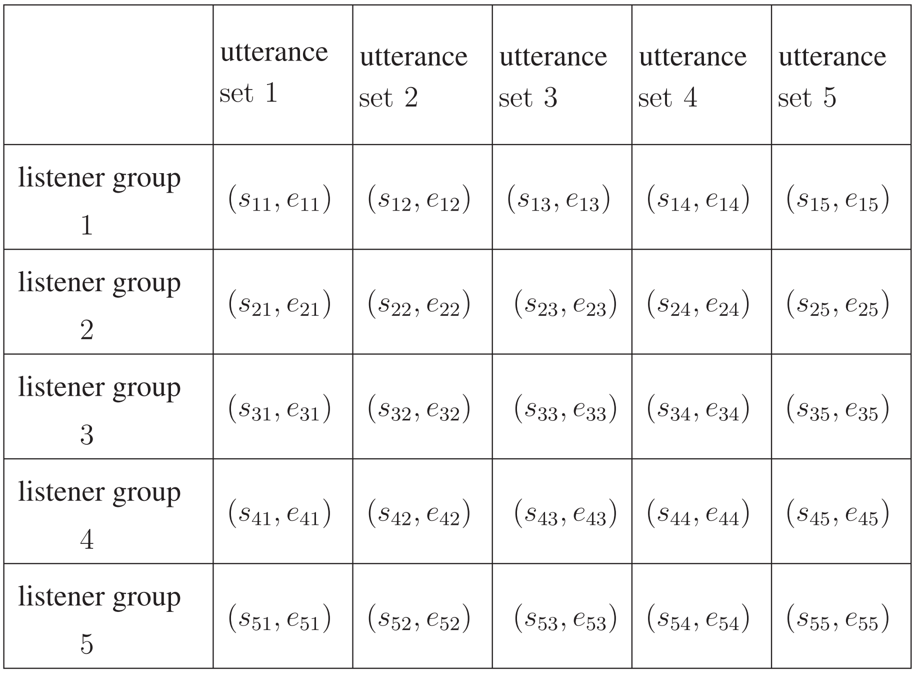
In reality, we only have four model sets per target speaker, angry, happy, sad, and neutral. To complete the Graeco-Latin square, we add the non-synthetic flavor. The score of this flavor provides us with a performance reference.

The test results are summarized in Table 4. For the reference case of the non-synthetic flavor, the correct answer is almost always chosen (40/45). However, in five cases the synthesized speech are chosen as non-synthetic, meaning that they sound more natural than the natural (non-synthetic) speech! For the synthesized emotional speech, we can see that the proposed method works remarkably well for sad and reasonably well for the neutral, given that the choice is 1-out-of-5. The happy and angry appears to form a confusable set, which can be attributed to the similar patterns in their respective excitation (pitch) models and duration (speaking rate) models.

To decide if the results are statistically significant, we conduct the following test of significance. For the results of the angry case, the p-value is

$$\begin{aligned} p_m &\leq 2 \cdot \left(1 - \Phi\left(\frac{|n_c - \mu|}{\sqrt{n}\sigma}\right)\right) \leq 2 \cdot \left(1 - \Phi\left(\frac{|21 - 9|}{\sqrt{45}/2}\right)\right) \\ &\approx 2 \cdot (1 - 0.9999) \approx 0, \end{aligned} \tag{15}$$

where n_c is the number of correctly answered questions, μ is the mean under the null hypothesis of random guess, n is the number of tested questions, σ is the variance



	utterance set 1	utterance set 2	utterance set 3	utterance set 4	utterance set 5
listener group 1	(s_{11}, e_{11})	(s_{12}, e_{12})	(s_{13}, e_{13})	(s_{14}, e_{14})	(s_{15}, e_{15})
listener group 2	(s_{21}, e_{21})	(s_{22}, e_{22})	(s_{23}, e_{23})	(s_{24}, e_{24})	(s_{25}, e_{25})
listener group 3	(s_{31}, e_{31})	(s_{32}, e_{32})	(s_{33}, e_{33})	(s_{34}, e_{34})	(s_{35}, e_{35})
listener group 4	(s_{41}, e_{41})	(s_{42}, e_{42})	(s_{43}, e_{43})	(s_{44}, e_{44})	(s_{45}, e_{45})
listener group 5	(s_{51}, e_{51})	(s_{52}, e_{52})	(s_{53}, e_{53})	(s_{54}, e_{54})	(s_{55}, e_{55})

Figure 4 The 5 × 5 Graeco-Latin square used in the evaluation of emotional expressiveness of the synthesized speech.

Table 4 The confusion matrix for the evaluation of the emotional expressiveness of synthesized speech using the proposed interpolation method

	Happy	Angry	Sad	Neutral	Non-synthetic	Accuracy
Happy	34	6	1	3	1	34/45
Angry	13	21	1	7	3	21/45
Sad	1	0	38	2	5	38/45
Neutral	0	0	9	20	15	20/45
Non-synthetic	1	0	0	4	40	40/45

There are five listener groups, three listeners in a listener group, and three utterances in an utterance set. Thus, the result of emotional expressiveness for each emotion is based on answers to 45 = 5 × 3 × 3 test samples

of a Bernoulli random variable, and $\Phi(x)$ is the cumulative distribution function of a standard normal random variable $X \sim \mathcal{N}(0,1)$. For other emotions (happy and sad), the p-values of the test statistics are even smaller than p_m. Therefore, the results in Table 4 are statistically significant.

Evaluation on naturalness

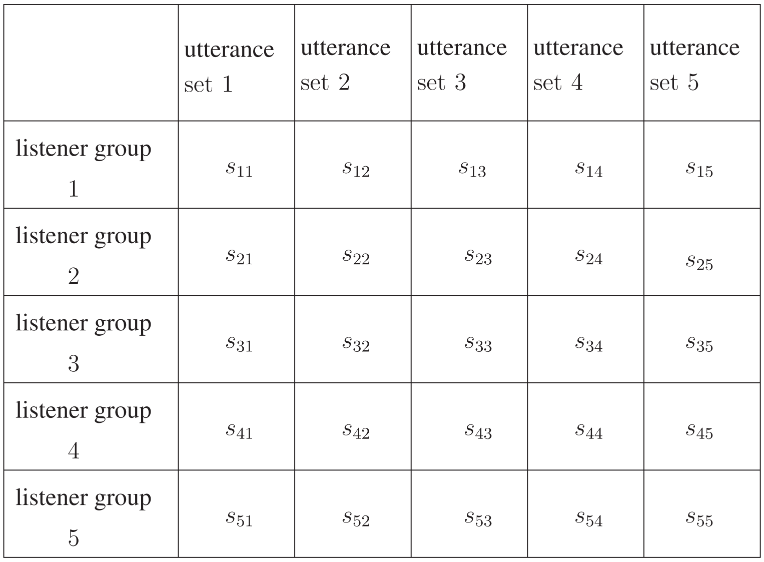
We use an MOS-based method for the evaluation on the naturalness of the synthesized speech. Basically, a listener listens to an utterance and rate the quality between 1 and 5 with the following scale: 5 for excellent, 4 for good, 3 for fair, 2 for poor, and 1 for bad. Note that the emotion of the speech is revealed to the listener, so a synthesized speech needs to be both natural and emotionally expressive to score high.

A Latin square as shown in Figure 5 is used in this evaluation. For each emotion, a listener listens to five synthesized speeches, each from a different target speaker and with a different prompt (text). In addition, a listener also rates a non-synthetic utterance. This non-synthetic utterance is randomly inserted during a test session. These non-synthetic utterances serve the purpose of calibrating the MOS scale to avoid MOS scores that are systematically biased upwards.

The test results are summarized in Table 5. Note that the score is based on the average of $75 = 5 \times 3 \times 5$ MOS scores per emotion, as there are five listener groups, three listeners per group, and five synthesized speeches per listener. On average, the synthesized speech achieves fair on angry and happy, slightly worse than fair for sad (2.9). Given that non-synthetic speech only achieves 4.6, the performance of the emotional synthesized speech is arguably acceptable.

Evaluation on similarity

The evaluation on the similarity is based on to what degree the synthesized emotional speech conveys the identity to the target speaker. The method of ABX test, where X is the synthesized emotional speech, A is a neutral



	utterance set 1	utterance set 2	utterance set 3	utterance set 4	utterance set 5
listener group 1	s_{11}	s_{12}	s_{13}	s_{14}	s_{15}
listener group 2	s_{21}	s_{22}	s_{23}	s_{24}	s_{25}
listener group 3	s_{31}	s_{32}	s_{33}	s_{34}	s_{35}
listener group 4	s_{41}	s_{42}	s_{43}	s_{44}	s_{45}
listener group 5	s_{51}	s_{52}	s_{53}	s_{54}	s_{55}

Figure 5 **The** 5×5 **Latin square used in the evaluation of naturalness of the synthesized speech.**

Table 5 Results for the evaluation of the naturalness of synthesized speech using the proposed method

Angry	Happy	Sad	Non-synthetic
3.0	3.1	2.9	4.6

non-synthetic speech from the target speaker, and B is a neutral non-synthetic speech from the pool speaker used in the interpolation (12), is used in this evaluation. Basically, a listener is asked to decide whether X sounds like the speaker of A or the speaker of B, without knowing the emotion label. For a fair comparison, the order of A and B for any given X is randomized, so there is no systematic bias towards the first choice or the second choice.

The test results are summarized in Table 6. They are based on $150 = 15 \times 10$ test samples as there are 15 listeners and each listener makes 10 choices. In all emotions, the correct choice is made by the test subjects in the majority of cases, with 73% for happy, 60% for angry, and 55% for sad. The overall accuracy is 62%. Therefore, the synthesized speech, even with emotions, still conveys the identity of the target speaker.

We use the Chi-squared test for testing statistical significance. The χ^2-value is

$$\chi^2 = 2 \cdot \frac{(93-75)^2}{75} = 8.64 > \chi^2_{0.01(1)}. \tag{16}$$

Therefore, the null hypothesis is rejected at the 0.01 significance level. Therefore, the synthesized speech does convey the speaker identity information. Alternatively, the same conclusion can be drawn by computing the p-value via the central-limit theorem, which yields

$$p \approx 2 \cdot \left(1 - \Phi\left(\frac{|93-75|}{\sqrt{150}/2}\right)\right) \approx 2 \cdot (1 - 0.9984) = 0.0032. \tag{17}$$

Comparison with the adaptation method

In this section, the proposed model interpolation method is compared to the model adaptation method. For fair comparison, the same amount of speech data used in interpolation-based system is used in adaptation-based system. For each emotion, 1,500 emotional utterances from 5 pool speakers are used to train an average-voice emotional model. For a target speaker, 300 neutral utterances are used to adapt the average-voice emotional model. Fifteen test subjects participate in each of the following sets of evaluation. The results are summarized as follows.

Table 6 Results for the evaluation of the similarity of synthesized speech using the proposed method

Angry	Happy	Sad	Overall
33/45(60%)	27/45(73%)	33/60(55%)	93/150(62%)

Table 7 Results of comparison between interpolation and adaptation regarding emotional expressiveness

Method	Angry	Happy	Sad	Overall
Interpolation	28/45	41/45	25/45	94/135
Adaptation	17/45	4/45	20/45	41/135

- *Expressiveness*
  In the evaluation of emotional expressiveness, a test subject is asked to listen to two synthesized speech and indicate the one with better expressiveness. One speech is synthesized by the interpolation system, while the other is synthesized by the adaptation system. The order of the synthesized speech is randomized. For each emotion, a test subject listens to three pairs. The results are summarized in Table 7. In 70% (94 out of 135) of the tests, the interpolation-based synthesized speech is chosen. Thus, interpolation outperforms adaptation in emotional expressiveness. The p-value for the null hypothesis of random guess based on evaluation results in Table 7 is

  $$p \approx 2 \cdot \left(1 - \Phi\left(\frac{|94-67.5|}{\sqrt{135}/2}\right)\right) \approx 2 \cdot (1 - 0.9999) \approx 0, \tag{18}$$

  where $\Phi_X(x)$ is the cumulative distribution function of a standard normal random variable $X \sim \mathcal{N}(0,1)$. Therefore, the null hypothesis is rejected and the difference in performance is statistically significant. Note that in the case of happy emotion, over 90% (41 out of 45) cases the interpolation-based synthesized speech is chosen.
- *Similarity*
  For the comparison of interpolation and adaptation, we adopt the ABX scheme, where X is the target speaker's non-synthetic speech, A is the synthetic emotional speech using adaptation, and B is the synthetic emotional speech using interpolation. Again, the order of A, B is randomized. The results are summarized in Table 8. Contrary to the emotional expressiveness, adaptation outperforms interpolation in 64% of the tests. The p-value for the

Table 8 Results of comparison between interpolation and adaptation regarding similarity

Method	Angry	Happy	Sad	Overall
Interpolation	9/45	14/45	31/60	54/150
Adaptation	36/45	31/45	29/60	96/150

Table 9 Results for the evaluation of the naturalness of synthesized speech using model adaptation

Angry	Happy	Sad	Non-synthetic
3.0	2.8	2.9	4.9

null hypothesis of random guess based on evaluation results in Table 8 is

$$p \approx 2\cdot\left(1-\Phi\left(\frac{|96-75|}{\sqrt{150}/2}\right)\right) \approx 2\cdot(1-0.9997) \approx 0.0006. \quad (19)$$

Therefore, the performance of adaptation is significantly better than interpolation regarding similarity.

- *Naturalness*
  The same Latin-square style MOS evaluation on naturalness as described in Section "Evaluation on naturalness" is used with synthesized speech by the adaptation-based system. The results are shown in Table 9. Comparing Table 5 with Table 9, we can see that interpolation-based system is slightly better. Furthermore, the "reference level" of non-synthetic speech is higher (by 0.3) in Table 9, which means that the difference between non-synthetic (natural) speech and synthesized speech is smaller with interpolation.

In summary, interpolation outperforms adaptation regarding emotional expressiveness, underperforms adaptation regarding similarity, and is slightly better than adaptation regarding naturalness. Note that the adaptation method is computationally more expensive than the interpolation method due to the training of average voice models.

Conclusion and future work

In this article, we propose and implement a speaker-dependent model interpolation method for HMM-based emotional speech synthesis. We use a novel MBMD measure to decide the interpolation model sets and weights. Comprehensive evaluation with subjective listening tests randomized by the Latin and Graeco-Latin squares to avoid systematic biases are carried out. The proposed model interpolation method has an intrinsic tradeoff between emotional expressiveness and similarity. Therefore, it is quite difficult to achieve both goals within the framework of interpolation. This can be seen as a drawback resulting from not using any emotional speech of the target speaker. Experiment results show that our method strikes a good balance between the emotional expressiveness, the naturalness, and the speaker identity. Additionally, our method does not require the emotional speech of new speakers, and can save enormous data collection and labeling costs. In the future, it will be interesting to compare the synthesized emotional speech with the non-synthetic emotional speech, with data collected from the target speaker, to further improve the performance.

Endnote

[a] Note that a *model set* refers to the entire set of HMMs for a given voice, while a *model* refers to only one basic linguistic unit such as a phone or a word, in this article.

Competing interests

The authors declare that they have no competing interests.

Received: 6 January 2012 Accepted: 8 June 2012
Published: 16 August 2012

References

1. AJ Hunt, AW Black, in *Proceedings of International Conference on Acoustics, Speech and Signal Processing,* Unit selection in a concatenative speech synthesis system using a large speech database. 1996, pp. 373–376
2. JZ Gros, M Zganec, An efficient unit-selection method for concatenative text-to-speech synthesis systems. J. Comput. Inf. Technol. **16**, 69–78 (2008)
3. T Yoshimura, K Tokuda, T Masuko, T Kobayashi, T Kitamura, in *Proceedings of Eurospeech,* Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis. 1999, pp. 2347–2350
4. H Zen, K Tokuda, AW Black, in *Proceedings of International Conference on Acoustics, Speech and Signal Processing,* Statistical parametric speech synthesis. 2007, pp. 1229–1232
5. M Kurimo, W Byrne, J Dines, PN Garner, M Gibson, Y Guan, T Hirsimäki, R Karhila, S King, H Liang, K Oura, L Saheer, M Shannon, S Shiota, J Tian, K Tokuda, M Wester, YJ Wu, J Yamagishi, in *Proceedings of the ACL 2010 System Demonstrations,* Personalising speech-to-speech translation in the EMIME project. Uppsala, Sweden, 2010
6. M Wester, J Dines, M Gibson, H Liang, YJ Wu, L Saheer, S King, K Oura, PN Garner, W Byrne, Y Guan, T Hirsimäki, R Karhila, M Shannon, S Shiota, J Tian, K Tokuda, J Yamagishi, in *Proceedings of 7th ISCA Speech Synthesis Workshop,* Speaker adaptation and the evaluation of speaker similarity in the EMIME speech-to-speech translation project. Kyoto, Japan, 2010
7. S Creer, P Green, S Cunningham, J Yamagishi, in *Computer Synthesized Speech Technologies: Tools for Aiding Impairment,* Building personalised synthesised voices for individuals with dysarthria using the HTS toolkit. January 2010
8. J Tao, Y Kang, A Li, Prosody conversion from neutral speech to emotional speech. IEEE Trans. Audio Speech Lang. Process. **14**(4), 1145–1154 (2006)
9. CH Wu, CC Hsia, TH Liu, JF Wang, Voice conversion using duration-embedded bi-HMMs for expressive speech synthesis. IEEE Trans. Audio Speech Lang. Process. **14**(4), 1109–1116 (2006)
10. J Yamagishi, K Onishi, T Masuko, T Kobayashi, in *Proceedings of Eurospeech,* Modeling of various speaking styles and emotions for HMM-based speech synthesis. 2003, pp. 2461–2464
11. J Yamagishi, K Onishi, T Masuko, T Kobayashi, Acoustic modeling of speaking styles and emotional expressions in HMM-based speech synthesis. IEICE Trans. Inf. Syst. **E88-D**(3), 502–509 (2005)
12. J Yamagishi, T Kobayashi, Y Nakano, K Ogata, J Isogai, Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm. IEEE Trans. Audio Speech Lang. Process. **17**, 66–83 (2009)
13. T Nose, J Yamagishi, T Masuko, T Kobayashi, A style control technique for HMM-based expressive speech synthesis. IEICE Trans. Inf. Syst. **E90-D**(9), 1406–1413 (2007)
14. M Tachibana, S Izawa, T Nose, T Kobayashi, in *ICASSP '08,* Speaker and style adaptation using average voice model for style control in HMM-based speech synthesis. 2008, pp. 4633–4636
15. T Nose, M Tachibana, T Kobayashi, HMM-based style control for expressive speech synthesis with arbitrary speaker's voice using model adaptation. IEICE Trans. Inf. Syst. **E92-D**(3), 489–497 (2009)
16. DN Jiang, W Zhang, LQ Shen, LH Cai, in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing 2005 (ICASSP '05),* Prosody analysis and modeling for emotional speech synthesis. 2005, pp. 281–284

17. M Tachibana, J Yamagishi, T Masuko, T Kobayashi, Speech synthesis with various emotional expressions and speaking styles by style interpolation and morphing. IEICE Trans. Inf. Syst. **E88-D**(11), 2484–2491 (2005)
18. T Yoshimura, T Masuko, K Tokuda, T Kobayashi, T Kitamura, in *Fifth European Conference on Speech Communication and Technology, EUROSPEECH 1997,* Speaker interpolation in HMM-based speech synthesis system. Rhodes, Greece, September 22–25, 1997
19. D Govind, SRM Prasanna, B Yegnanarayana, in *INTERSPEECH,* Neutral to target emotion conversion using source and suprasegmental information. ISCA, 2011, pp. 2969–2972
20. LR Rabiner, A tutorial on hidden Markov models and selected applications in speech recognition. Proc. IEEE. **77**(2), 257–286 (1989)
21. T Tokuda, T Masuko, N Miyazaki, T Kobayashi, Multi-space probability distribution HMM. IEICE Trans. Inf. Syst. **E85-D**(3), 455–464 (2002)
22. AP Dempster, NM Laird, DB Rubin, Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. Ser. B. **39**, 1–38 (1977)
23. JA Bilmes, A gentle tutorial of the EM algorithm and its application to parameter estimation for gaussian mixture and hidden Markov models. Technical report, International Computer Science Institute, University of California at Berkeley TR-97-021, April 1998
24. K Tokuda, T Masuko, N Miyazaki, T Kobayashi, in *Proceedings of International Conference on Acoustics, Speech and Signal Processing,* Hidden Markov models based on multi-space probability distribution for pitch pattern modeling. 1999, pp. 229–232
25. K Tokuda, T Kobayashi, T Masuko, S Imai, in *Proceedings of International Conference on Spoken Language Processing,* Mel-generalized cepstral analysis. 1994, pp. 1043–1046
26. T Kobayashi, S Imai, T Fukuda, Mel-generalized log spectral approximation filter. IEICE Trans. Fund. **J68-A**(6), 610–611 (1985)
27. C Huang, Y Shi, J Zhou, M Chu, T Wang, E Chang, in *Proceedings of International Conference on Acoustics, Speech and Signal Processing,* Segmental tonal modeling for phone set design in, Mandarin LVCSR. 2004, pp. 901–904
28. H Zen, An Example of Context-dependent label format for HMM-based Speech synthesis in English (2006). [https://wiki.inf.ed.ac.uk/twiki/pub/CSTR/F0parametrisation/hts_lab_format.pdf]
29. T Yoshimura, K Tokuda, T Masuko, T Kobayashi, T Kitamura, Speaker interpolation for HMM-based speech synthesis system. J. Acoust. Soc. Jpn. **21**(4), 199–206 (2000)
30. M Pucher, D Schabus, J Yamagishi, F Neubarth, V Strom, Modeling and interpolation of Austrian German and Viennese dialect in HMM-based speech synthesis. Speech Commun. **52**(2), 164–179 (2010)
31. M Fraser, S King, in *Proceedings of the Sixth ISCA Workshop on Speech Synthesis (SSW6),* The Blizzard challenge 2007, (2007)

doi:10.1186/1687-4722-2012-21
Cite this article as: Hsu and Chen: **Speaker-dependent model interpolation for statistical emotional speech synthesis.** *EURASIP Journal on Audio, Speech, and Music Processing* 2012 **2012**:21.