A Hidden Markov Model-Based Approach for Emotional Speech Synthesis

Chih-Yung Yang and Chia-Ping Chen
Department of Computer Science and Engineering
National Sun Yat-Sen University
Kaohsiung, Taiwan
ycyg97m@cse.nsysu.edu.tw, cpcchen@cse.nsysu.edu.tw

Abstract
In this paper, we describe an approach to automatically syn-
thesize the emotional speech of a target speaker based on the
hidden Markov model for his/her neutral speech. The basic idea
is the model interpolation between the neutral model of the tar-
get speaker and an emotional model selected from a candidate
pool. Both the interpolation model selection and the interpo-
lation weight computation are determined based on a model-
distance measure. In this paper, we propose a monophone-
based Mahalanobis distance (MBMD). We evaluate our ap-
proach on the synthesized emotional speech of anger, happiness, and sadness with several subjective tests. Experimental
results show that the implemented system is able to synthesize
speech with emotional expressiveness of the target speaker.

Index Terms: speech synthesis, HMM, emotional expressiveness, Mahalanobis distance, model interpolation

1. Introduction
The state-of-the-art text-to-speech (TTS) systems are com-
monly based on unit selection [1] or statistical parametric mod-
els [2]. While unit selection is still the dominant methodol-
ogy, synthesis based on statistical parametric models has been
gaining popularity. Compared to the unit-selection approach,
model-based technique is flexible for controlling the synthet-
ized speech via model parameter transformation. Hence, in
this paper, we use the model-based method to synthesize em-
otional speech.

Many approaches have been proposed to generate speech
with emotional expressiveness. In [3], emotional expressiveness are deemed as different styles, which are modeled by style-
dependent and style-mixed modeling. In [4], the maximum like-
lihood linear regression (MLLR) adaptation and the constrained
structural maximum a posteriori linear regression (CSMAPLR)
adaptation are used to transform the neutral model to the target
style model, requiring only a small amount of adaptation data. In [5], model interpolation interpolates two speaking-style mod-
els to generate a spectrum of speaking styles. In [6], multiple-
regression HMMs for an average voice model is used, and sim-
taneous adaptation of speaker and style is applied to syn-
these models with style expressiveness of the target speaker.

One major difference between our proposed approach for
emotionally expressive and the previous approaches is that we
do not require the emotional speech data from the target
speaker. It is not difficult to see that collecting such speech data
could be infeasible. Therefore, we hope to develop a system
capable of synthesizing emotional speech with only the neutral
speech data of the target speaker.

The basic idea in the proposed approach is model interpo-
lation [7]. First, we train the emotional speech models of a pool
of speakers with their emotional speech data. Given a target
speaker and his/her neutral speech data, we first train the neu-
tral model of the target speaker. Then we interpolate this neutral
model with an emotional model of a speaker selected from the
pool of speakers.

This paper is organized as follow. First, in order to put
the discussion in a proper framework, we describe the hidden
Markov model-based speech synthesis with sufficient and rele-
vant details in Section 2. The proposed method for emotional
expressiveness based on model interpolation is deduced in Sec-
3. Evaluation method and results of the proposed approach
are presented in Section 4. The conclusion remarks are given in
Section 5.

2. HMM-Based Speech Synthesis
A hidden Markov model (HMM)-based speech synthesis sys-
tem [2] consists of the training part and the synthesis part. The
parameters of the statistical models used in the system are esti-

dated in the training part. First, the speech parameters related
to the spectrum and the fundamental frequency are extracted.
The spectral parameters are the mel cepstral coefficients and the
delta and delta-delta coefficients, while the excitation param-
ters are log F0, where F0 is the fundamental frequency, and the
delta and delta-delta. Note that F0 is a one-dimensional continu-
ous value for voiced regions, and is discrete for the unvoiced
regions. In this system, the spectral and excitation parameter is
modeled by multi-space probability distribution HMM (MSD-}
HMM),

$$b(o) = \sum_g w_g p_g(V(o)),$$

where o is the observation vector, g is the space index, w_g is
the weight, and p_g(·) is the probability density in space g. The
state durations are modeled by a multi-variate Gaussian dis-


tribution, whose mean vector and covariance matrix are esti-

dated by the HMM state occupancies of training data. Note
that the context-dependent modeling scheme is used for the
phonetic and linguistic contextual factors, where the decision
trees are constructed based on the minimum-description-length
(MDL) criterion. In the synthesis part, a text string is mapped
to a context-dependent phoneme label sequence. This sequence
is used to concatenate the corresponding context-dependent
phoneme HMMs to construct a sentence HMM. The state se-
sequence \( q = \{ q_1, q_2, \ldots, q_T \} \) of the sentence HMM \( \lambda \) is determined by maximizing the probability

\[
\log P (q|\lambda, T) = \sum_{i=1}^{T} \log p_i (d_i)
\]

(2)

where \( p_i (d_i) \) is the probability of duration \( d_i \), in state \( i \). From \( q \), we decide the static mel cepstral coefficients and \( \log F_0 \) streams by maximizing the joint data-likelihood of static and dynamic feature streams. Finally, the synthesized speech is generated through a mel-log-spectral approximation (MLSA) filter [8], which is approximated by a rational function

\[
H(z) \approx \exp \left( \sum_{m=0}^{M} c(m) z^{-m} \right)
\]

(3)

where \( c(m) \) is the cepstral coefficient, and

\[
\hat{z}^{-1} = \frac{z^{-1} - \alpha}{1 - \alpha z^{-1}},
\]

(4)

where \( \alpha \), the parameter for frequency warping, is dependent on the speech sampling rate. Table 1 lists a few instances of the sampling rate and \( \alpha \).

<table>
<thead>
<tr>
<th>Sampling Rate</th>
<th>8kHz</th>
<th>10kHz</th>
<th>16kHz</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \alpha )</td>
<td>0.31</td>
<td>0.35</td>
<td>0.42</td>
</tr>
</tbody>
</table>

The block diagram of the HMM-based speech synthesis system is shown in Figure 1.

Figure 1: Overview of HMM-based speech synthesis system

3. Model Interpolation

Interpolation methods are used to combine several speech synthesis models to generate a new model. The interpolation weights (ratios) can be adjusted for the intended target speaker and emotion. Various applications of interpolation have been experimented. In [7], model interpolation gradually changes the synthesized voice from a male speaker to a female speaker. In [5], model interpolation generates a spectrum of speaking styles, such as the “slightly joyful” speech. In [9], model interpolation is used to bridge the gap between the Austrian German and the Viennese dialect.

Let the models be denoted by \( \lambda_1, \lambda_2, \ldots, \lambda_K \), from which we want to create an interpolated model \( \lambda \). Let the interpolation ratios be \( a_k(i), a_k(i), \ldots, a_K(i) \), with the constraint that

\[
\sum_{k=1}^{K} a_k(i) = 1, \quad i = 1, \ldots, I.
\]

(5)

\( a_k(i) \) represents the contribution of \( \lambda_k \) in \( \lambda \) for state \( i \), and is often given as

\[
a_k(i) = \frac{\gamma_k(i)}{\sum_{k'=1}^{K} \gamma_{k'}(i)}, \quad k = 1, \ldots, K,
\]

(6)

where \( \gamma_k(i) \) is the occupancy for state \( i \) from model \( \lambda_k \). In general, the interpolation ratios can vary with state \( i \). However, one can tie the parameters \( a_k(i) \) of all states together, leading to

\[
a_k = \frac{\gamma_k}{\sum_{k'=1}^{K} \gamma_{k'}}, \quad k = 1, \ldots, K,
\]

(7)

where \( \gamma_k \) is the total data amount for model \( \lambda_k \). The simplified version is used in our implementation.

In the method of interpolation of observation [5], the parameters of \( \lambda \) is given by

\[
\bar{\mu}(i) = \frac{\sum_{k=1}^{K} a_k \mu_k(i)}{\sum_{k=1}^{K} a_k}, \quad \bar{U}(i) = \frac{\sum_{k=1}^{K} a_k^2 U_k(i)}{\sum_{k=1}^{K} a_k} - \bar{\mu}(i) \bar{\mu}^T(i), \quad i = 1, \ldots, I,
\]

(8)

where \( \mu_k(i) \) and \( U_k(i) \) are the mean vector and the covariance matrix for the state \( i \) in model \( \lambda_k \), and \( \bar{\mu}(i) \) and \( \bar{U}(i) \) be the mean vector and the covariance matrix for the output model \( \lambda \).

Alternatively, one can also minimize the weighted sum of the Kullback-Leibler distance between \( \lambda_k \) and \( \lambda \) [7]

\[
\epsilon = \sum_{k=1}^{K} a_k I(\lambda_\k, \lambda_k).
\]

(9)

As a result, \( \bar{\mu}(i) \) and \( \bar{U}(i) \) are determined by

\[
\bar{\mu}(i) = \left( \sum_{k=1}^{K} a_k U_k(i)^{-1} \right)^{-1} \sum_{k=1}^{K} a_k U_k(i)^{-1} \mu_k(i),
\]

(10)

\[
\bar{U}(i) = \left( \sum_{k=1}^{K} a_k U_k(i)^{-1} \right)^{-1} \sum_{k=1}^{K} a_k U_k(i).
\]

In this paper, we use the interpolation of observation (8). We apply the model interpolation of two models \( K = 2 \). One
is the neutral model, denoted by \( N \), for the target speaker. The other is the emotional model, denoted by \( E \), of a speaker selected from a pool. We interpolate the spectral, \( F_0 \), and the state duration model parameters. From models \( N \) and \( E \), the interpolation model with ratio \( r \) is denoted by \( I(N, E, r) \). The interpolated mean vector \( \hat{\mu}(i) \) and covariance matrix \( \hat{\Sigma}(i) \) for state \( i \) are given by

\[
\begin{align*}
\hat{\mu}(i) &= r\mu_N(i) + (1-r)\mu_E(i), \\
\hat{\Sigma}(i) &= r^2\Sigma_N(i) + (1-r)^2\Sigma_E(i).
\end{align*}
\]

(11)

The block diagram of the interpolation procedure is shown in Figure 2.

Figure 2: Block diagram of the interpolation procedure

### 3.1. Methods

Let \( N_1, N_2, \ldots, N_K \) be the neutral models of \( M \) candidate speakers. Likewise, let \( E_1, E_2, \ldots, E_K \) be the emotional models. Let the target speaker neutral model be denoted by \( N_T \). We need a distance measure between models which is critical in our methods. Since the logical context-dependent HMMs in different synthesis model sets have different structures, the distance measure based on context-dependent models would be difficult. Therefore, we propose to use the monophone-based Mahalanobis distance (MBMD). The MBMD between two synthesis models, say \( \lambda_1 \) and \( \lambda_2 \), are defined as follows

\[
\text{MBMD}(\lambda_1, \lambda_2) = \sum_{p=1}^{P} (\mu_1(p) - \mu_2(p))^T \left( \frac{\Sigma_1(p) + \Sigma_2(p)}{2} \right)^{-1} (\mu_1(p) - \mu_2(p)),
\]

(12)

where \( P \) is the total number of states in the monophone models, \( \mu_1(p) \) and \( \Sigma_1(p) \) are mean vector and covariance matrix of monophone-state \( p \) of model \( \lambda_1 \). Note that the MBMD as defined in (12) is a proper distance measure since

\[
\begin{align*}
\text{MBMD}(\lambda_1, \lambda_2) &\geq 0 \\
\text{MBMD}(\lambda_1, \lambda_2) &\geq 0 \iff \lambda_1 = \lambda_2.
\end{align*}
\]

(13)

With MBMD, our methods for model selection and interpolation ratio determination can be described as follows.

- From the neutral models \( N_1, \ldots, N_K \), we decide the one that is closest to \( N_T \). This model is denoted by \( N_X \), i.e.,

\[
N_X = \arg \min_{N_k} \text{MBMD}(N_T, N_k);
\]

(14)

- Find the model in \( E_1, \ldots, E_K \) which is closest to \( N_T \). This model is denoted by \( E_Z \), i.e.,

\[
E_Z = \arg \min_{E_k} \text{MBMD}(N_T, E_k);
\]

(15)

- Let \( E_X \) be the emotional model corresponding to \( N_X \). We can find the interpolation ratio \( R \) by using the monophone model interpolation such that

\[
R^* = \arg \min_{R} \text{MBMD}(I_{\text{mono}}(N_T, E_Z, R), E_X);
\]

(16)

where \( I_{\text{mono}}(N_T, E_Z, R) \) is the monophone model interpolation between \( N_T \) and \( E_Z \) with ratio \( R \). A grid search in \([0, 1]\) with increment of 0.1 is used to find \( R^* \).

- The final interpolation model is

\[
E_T = I(N_T, E_Z, R^*).
\]

(17)

In summary, we choose a speaker \( X \) whose neutral model \( N_X \) is closest to \( N_T \). We also choose a speaker \( Z \) whose emotional model \( E_Z \) is closest to \( N_T \). We find the interpolation ratio by minimizing the distance between \( E_X \) (which is assumed close to \( E_T \)) and interpolated model from \( N_T \) and \( E_Z \). The closest interpolated model is our proposed \( E_T \). The block diagram of the proposed method is shown in Figure 3. Note that the interpolation scheme is applied to the spectral, \( F_0 \), and state-duration models.

Figure 3: The proposed method

### 4. Experiments

#### 4.1. Test Data

We collect a Mandarin speech database for emotional speech synthesis. In this database, there are three target emotions, i.e.,
“angry”, “sad”, and “happy”, each corresponds to a phonetically balanced set of 300 sentences. Three male speakers record each set of sentences with emotional expressiveness, and neutrally. Thus, each speaker records 300 emotional utterances and 900 neutral utterances. Each neutral model is trained by 900 utterances, and each emotional model is trained by 300 utterances. These models constitute the model pool.

For each target speaker, we collect 100 neutral utterances to train his/her neutral model, as well as 30 utterances for each target emotion. These emotional speech are used in the evaluation as the reference samples. In the subjective tests, we synthesize 3 sentences randomly chosen from the 30 reference samples for each target emotion. Ten subjects are invited to perform the subjective tests.

4.2. Emotional Expressiveness Test

We perform the “ABX” test to evaluate the emotional expressiveness of the synthesized speech. “A” represents the reference neutral speech, “B” represents the reference emotional speech, and “X” represents the synthesized speech by the proposed method. Tested subjects listen to instances of “A”, “B”, and “X”, and select either “A” or “B” as being closer to “X”. Table 2 shows the percentage of decisions in which “B” is the chosen answer. Since “X” is the synthesized emotional speech, the numbers can also be interpreted as the accuracy rate. For each target emotion, the accuracy rate is higher than 70%.

<table>
<thead>
<tr>
<th></th>
<th>Angry</th>
<th>Happy</th>
<th>Sad</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>77%</td>
<td>77%</td>
<td>73%</td>
</tr>
</tbody>
</table>

4.3. Naturalness Test

We also conduct a subjective test for the naturalness of the synthesized speech, using the mean opinion score (MOS). A 5-point scale is used with 5 for excellent, 4 for good, 3 for fair, 2 for poor, and 1 for bad. Table 3 lists the MOS averages along with the confidence intervals for the naturalness test. From Table 3, we can see that naturalness test MOS of the synthesized emotional speech by the proposed method is close to the synthesized neutral speech. These results show that our proposed method maintains naturalness. This is likely due to our choice of the interpolating emotional model $E_x$, which is closest (in the MBMD sense) to the target-speaker neutral model $N_T$.

<table>
<thead>
<tr>
<th></th>
<th>Neutral</th>
<th>Angry</th>
<th>Happy</th>
<th>Sad</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ave. MOS</td>
<td>3.7 (0.2)</td>
<td>3.3 (0.3)</td>
<td>3.4 (0.4)</td>
<td>3.6 (0.3)</td>
</tr>
</tbody>
</table>

5. Conclusion

We propose an interpolation approach for HMM-based emotional speech synthesis which does not require the emotional speech data from the target speaker. We interpolate the target speaker neutral model and an emotional model selected from a model pool. We use a monophone-based Mahalanobis distance for model selection and weight determination. Subjective evaluation test results show that the synthesized speech by our method maintains emotional expressiveness, naturalness, and speaker identity.

6. References


Table 2: Result of the ABX test.

<table>
<thead>
<tr>
<th></th>
<th>Angry</th>
<th>Happy</th>
<th>Sad</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>77%</td>
<td>77%</td>
<td>73%</td>
</tr>
</tbody>
</table>

Table 4: Mean opinion scores (MOS) of the similarity test. The number in the parentheses is the confidence interval.

<table>
<thead>
<tr>
<th></th>
<th>Angry</th>
<th>Happy</th>
<th>Sad</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ave. MOS</td>
<td>3.1 (0.4)</td>
<td>3.4 (0.4)</td>
<td>3.4 (0.4)</td>
</tr>
</tbody>
</table>