Abstract

This paper describes an approach to controlling style of synthetic speech in HMM-based speech synthesis. The style is defined as one of speaking styles and emotional expressions in speech. We model each speech synthesis unit by using a context-dependent HMM whose mean vector of the output distribution function is given by a function of a parameter vector called style control vector. We assume that the mean vector is modeled by multiple regression with the style control vector. The multiple regression matrices are estimated by EM-algorithm as well as other model parameters of HMMs. In the synthesis stage, the mean vectors are modified by transforming an arbitrarily given control vector which is associated with a desired style. The results of subjective tests show that we can control styles by choosing the style control vector appropriately.

1. Introduction

Text-to-Speech synthesis is a key technology for bringing human-computer interaction (HCI) closer to human-human interaction. In HCI that includes agents or virtual humans, it would be required that TTS system can not only generate natural sounding speech but also control speaking style and/or emotional expression arbitrarily. Although many researchers have attempted to add emotional expression to synthetic speech (for example, see [1] and references therein), this still remains a difficult problem which has not been solved. The difficulty of the problem is that we have to control prosodic features as well as spectral features properly to realize emotional expressivity and speaking style variability. Moreover, since prosodic features are more or less related to spectral features, we have to avoid controlling them independently.

In this paper, we describe an approach to resolving this problem. So far we have proposed several approaches to realizing emotional expressivity and speaking style variability in the HMM-based speech synthesis, namely, style modeling [2], style interpolation [3], and style adaptation [4] approaches. In the style modeling approach, we showed that we can model and generate speech of a certain speaking style or emotional expression as that of all styles by incorporating a style control vector. Furthermore, in the style interpolation approach, we also showed that we can synthesize speech with an intermediate style between two or possibly more styles. However, these approaches can change the style of synthetic speech only in an implicit way. On the other hand, in this paper, we attempt to control the style explicitly by defining a set of style control parameters, which will be called the style control vector, and modeling styles as functions of the style control vector.

The proposing technique uses a framework of the HMM-based speech synthesis of [5]. Since spectral and prosodic features are simultaneously and fully statistically modeled and generated, we can incorporate the relationship between these features into speech synthesis process implicitly. Furthermore, we can realize any desired style of synthetic speech by specifying the style control vector representing a point in a style space where each coordinate represents a specific style, such as joyful or sad. In the following, we describe a modeling and control technique of styles and show results of preliminary experiments focusing on a problem of generating any of reading, rough, joyful, and sad style speech.

2. Style Modeling with Control Vector for HMM-Based Speech Synthesis

In the HMM-based speech synthesis, context dependent phoneme HMMs are used as the synthesis units, in which spectrum and F0 are modeled simultaneously [5]. To model variations of spectrum and F0, phonetic and linguistic contextual factors, such as phoneme identity factors, stress related factors and locational factors, are taken into account. Then, a decision tree based context clustering technique is applied to the spectral and F0 parts of the context dependent phoneme HMMs.

In [2], for the purpose of modeling styles, we proposed a style dependent modeling method, in which each style is individually modeled and its decision tree structure is different from those of other styles. In contrast, here we first train context-dependent HMMs without context clustering for respective styles independently (see Fig.1). Then we apply a shared decision tree context clustering (STC) technique [6] to these models to construct a common tree structure for all styles. Finally, we obtain a single model which has the same tree structure as that of all styles by incorporating a style control vector into reestimation procedure as follows.

Suppose that a leaf node $i$ of the model has a Gaussian PDF with mean vector $\mu_i$. We assume that the mean vector is modeled by a function of a set of parameters $\{v_1, v_2, \ldots, v_M\}$, specifically expressed by multiple regression as

$$\mu_i = H_i \xi, \quad \xi = [1, v^\top]^\top$$ \hspace{1cm} (1)

where $v = [v_1, v_2, \ldots, v_M]^\top$, which we will call the style control vector, $H_i = [h_{i1}, h_{i2}, \ldots, h_{iM}]$ is a $D \times (M+1)$-dimensional multiple regression matrix, $D$ is the dimensionality of $\mu_i$, and $^\top$ denotes the matrix transpose. Here the multiple regression matrix $H_i$ is estimated as well as other HMM parameters in an ML sense. It is noted that although the training procedure is similar to that of the STC technique with speaker adaptive training (SAT) proposed in [7], the proposing approach differs from [7] in that the model retains multiple regression matrices and does not have mean vectors as the model parameters.
In each style model after applying the STC of [6]. We assume \( \mu_i \) as an initial estimate.

Suppose that speech database contains \( S \) styles. Let \( \mu_i^{(s)} \), \( 1 \leq s \leq S \) be the mean vector of the Gaussian PDF at a leaf node \( i \) in each style model after applying the STC of [6]. We assume that the style control vector for each style is given by \( \nu^{(s)} = [v_1^{(s)}, v_2^{(s)}, \ldots, v_M^{(s)}] \). We choose \( H_i \) that minimizes

\[
E = \sum_{s=1}^{S} \left\| \mu_i^{(s)} - \left( h_{i0} + \sum_{m=1}^{M} v_{im}^{(s)} h_{im} \right) \right\|^2
\]

as an initial estimate.

By differentiating \( E \) with respect to \( h_{i, j} \) and setting the result to zero, we have

\[
S h_{i0} + \sum_{m=1}^{M} \sum_{s=1}^{S} v_{im}^{(s)} h_{im} = \sum_{s=1}^{S} \mu_i^{(s)} \quad (j = 1, 2, \ldots, M).
\]

Hence we can obtain the initial estimate of \( H_i \) by solving a set of linear equations.

### 3.2. ML estimation of \( H_i \)

Suppose that the training data contain \( N \) feature vector sequences and \( O^{(n)} = (o_1^{(n)}, o_2^{(n)}, \ldots, o_T^{(n)}) \), the \( n \)-th observation sequence, is observed as the output from an HMM \( \lambda \). Here we use EM algorithm to obtain an ML-estimates of the HMM parameters and \( H_i \). In the EM algorithm, the auxiliary function (Q-function) for the parameter vector \( b_i \) that defines the PDF \( b_i(\cdot) \) is given by

\[
Q_{b_i}(\lambda, b_i) = \sum_{n=1}^{N} \frac{1}{P(O^{(n)}|\lambda, v^{(n)})}.
\]

where \( q_t^{(n)} = i \) means being in state whose PDF belongs to leaf node \( i \) at time \( t \). Since we assumed a Gaussian PDF, we can express the log likelihood \( \log b_i(o_t^{(n)} | v^{(n)}) \) as

\[
\log b_i(o_t^{(n)} | v^{(n)}) = -\frac{1}{2} \sum_{n=1}^{N} \log \left| U_i \right| - \frac{1}{2} (o_t^{(n)} - H_i \xi^{(n)})^\top U_i^{-1} (o_t^{(n)} - H_i \xi^{(n)})
\]

where \( U_i \) is the covariance matrix and \( v^{(n)} \) is the mean vector given by (1).

Differentiating (5) with respect to \( H_i \) and equating the result to zero, we obtain a reestimation formula for \( H_i \) given by

\[
H_i = \sum_{n=1}^{N} \sum_{t=1}^{T_n} \gamma_t^{(n)}(i) o_t^{(n)} \xi^{(n)} \xi^{(n)} \gamma_t^{(n)} \quad (j = 1, 2, \ldots, M).
\]

where

\[
\gamma_t^{(n)}(i) = P(q_t^{(n)} = i | O^{(n)}, \lambda, v^{(n)}).
\]

The other parameters of HMM \( \lambda \), state-transition probabilities and the covariance matrices can be obtained by using reestimation formulas for the conventional HMM training.

It is noted that the above formulation is similar to the result of [8].

### 4. Experiments

#### 4.1. Choice of style control vector

We carried out preliminary experiments for controlling style of synthetic speech. We used a speech database which is the same as that used in the style modeling experiment of [2] and contains four styles, namely reading, rough, joyful, and sad styles.
sentences uttered by a male speaker in each style. The speech database is composed of phonetically balanced 503 sentences which consists of the four styles, we chose a two-dimensional space as shown in Fig.2 and set the style control vectors as

\[
\begin{align*}
\text{Reading} & : (p, q) = (0, 0) \\
\text{Rough} & : (p, q) = (0, 1) \\
\text{Joyful} & : (p, q) = (1, 0) \\
\text{Sad} & : (p, q) = (-1, 0)
\end{align*}
\]

where the style control vector \( \mathbf{u}^{(s)} \) is expressed by a point on \((p, q)\) plane. From Fig.2, it is seen that the reading style is chosen as the origin of the \((p, q)\) plane and the joyful and sad styles are represented as opposite points on the \(p\)-axis. This is from the fact that the reading style is close to the interpolated style between joyful and sad styles with interpolation ratio of 1:1 [3]. Although it is thought that polite style is the opposite to rough style, we found that the polite style speech that we recorded was very close to reading style speech from a subjective evaluation test. Hence we did not use the polite style and defined the \(q\)-axis only in the positive region.

### 4.2. Experimental conditions

Experimental conditions for model training and waveform generation are basically the same as those of [2]. Speech signals were sampled at a rate of 16kHz and windowed by a 25ms Blackman window with a 5ms shift. Then mel-cepstral coefficients were obtained by mel-cepstral analysis. The feature vectors consisted of 25 mel-cepstral coefficients including the zeroth coefficient, logarithm of fundamental frequency, and their delta and delta-delta coefficients. We used 5-state left-to-right HMMs. Models were trained using 450 sentences for each style. It is noted that only the mean vectors of PDFs are reestimated and the covariance matrices and state duration distributions of the reading style were used in the following experiments. Subjects were 10 males, and for each subject, eight test sentences for each style were chosen at random from 53 test sentences which were not contained in training data.

### 4.3. Style reproduction test

In the first experiment, we synthesized speech with setting the values of the control vector to the same values used in the model training shown in Fig.2. Then we evaluated the synthetic speech by a style classification test. Subjects were asked which style, namely reading, rough, joyful, or sad, the test speech sounded. Test speech was classified into other style when it was thought to be classified into none of the above four styles.

The result is shown in Fig.3. In the figure, each score gives the rates that the synthetic speech was classified into the same style as the specified style. Although the scores are slightly lower than those of the style dependent model in [2], it can be seen that the original style is almost reproduced in the synthetic speech.

### 4.4. Style control with changing single control parameter

In the second experiment, we synthesized speech with changing the values of the control parameters as shown in Fig.4. Here we evaluated the synthetic speech by comparing it with that of the style dependent models of [2]. Subjects were presented pairs of a test sample and the reference sample, where the reference sample is the synthetic speech generated using the corresponding style dependent model. Then subjects were asked to evaluate the style of the test sample in a five point scale, where 5 for sounding that the specified style is very emphasized, 3 for sounding with almost the same style as the reference, and 1 for sounding with almost the same style as the reading style.

The result is shown in Fig.5. We calculate each score by

\[
\text{Score} = \frac{\sum_{x=1}^{5} x \times \text{(number of times that } x \text{ was chosen)}}{\text{(total number of evaluation pairs)}}
\]

where \(x\) is the evaluated score by subjects. It is seen from the result that we can control the degree of the style by changing corresponding style parameter value.

### 4.5. Style control with changing multiple control parameters

In the third experiment, we synthesized speech with changing the values of the control parameters as shown in Fig.6. The evaluation method is almost the same as the second experiment except that each test sample was compared with the reference twice by changing the style dependent model. Figure 7 shows...
the result. It is seen again that we can control the degree of the style by changing corresponding style parameter values.

5. Conclusion

We have presented a new approach to realizing emotional expressivity and speaking style variability in HMM-based speech synthesis. We have attempted to control the style in an explicit way by defining a set of style control parameters called the style control vector and modeling styles as functions of the style control vector. From the results of subjective tests, we have shown that we can control styles by choosing the style control vector appropriately. Moreover we can generate any desired style of synthetic speech by specifying the style control vector which represents a point in a style space. Our future work will focus on the modeling of the covariance matrices and durations as well as the mean vectors and detailed subjective evaluation.

6. References


