A CONTEXT CLUSTERING TECHNIQUE FOR AVERAGE VOICE MODEL IN HMM-BASED SPEECH SYNTHESIS

Junichi Yamagishi, Masatsune Tamura, Takashi Masuko, Keiichi Tokuda, and Takao Kobayashi

Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology, Yokohama, 226-8502 Japan
Department of Computer Science, Nagoya Institute of Technology, Nagoya, 466-8555 Japan
Email: {Junichi.Yamagishi,mtamura,masuko,Takao.Kobayashi}@ip.titech.ac.jp, tokuda@ics.nitech.ac.jp

ABSTRACT

This paper describes a new technique for constructing a decision tree used for clustering average voice model, i.e., speaker independent speech units. In the technique, we first train speaker dependent models using multi-speaker speech database, and then construct a speaker independent decision tree for context clustering common to these speaker dependent models. When a node of the decision tree is split, only the context related questions which can split the node for all speaker dependent models is adopted. Consequently, all nodes of the decision tree have all speakers’ training data. From the result of the paired comparison test, we show that the average voice model trained using the proposed technique can synthesize more natural sounding speech than the conventional average voice model.

1. INTRODUCTION

Speech synthesis is one of the key component for realizing natural human-computer interaction. For this purpose, text-to-speech (TTS) synthesis systems are required to have an ability to generate speech with arbitrary speaker’s voice characteristics and various speaking styles. There have been proposed a number of TTS techniques, and state-of-the-art TTS systems based on unit selection and concatenation can generate natural sounding speech. However, it is still a difficult problem to synthesize speech with various voice characteristics and speaking styles.

We have proposed an HMM-based TTS system in which each speech synthesis unit is modeled by HMM [1][2]. A distinctive feature of the system is that speech parameters used in the synthesis stage are generated directly from HMMs by using a parameter generation algorithm [3][4]. Since the HMM-based TTS system uses HMMs as the speech units in both modeling and synthesis, we can easily change voice characteristics of synthetic speech by transforming HMM parameters appropriately. In fact, we have shown that the TTS system can generate synthetic speech which closely resembles an arbitrarily given speaker’s voice using a small amount of target speaker’s speech data by applying speaker adaptation techniques based on MLLR (Maximum Likelihood Linear Regression) algorithm [5]-[7]. In that system, speaker independent model trained using multi-speaker speech database is used as an initial model of speaker adaptation. Since the synthetic speech generated from the speaker independent model can be considered to have averaged voice characteristics and prosodic features of the speakers used for training, we refer to the speaker independent model as average voice model, and the synthetic speech generated from average voice model as average voice.

It is thought that the quality of the average voice crucially affects the quality of synthetic speech generated from adapted models. In order to obtain a good average voice model using speech data of a large number of speakers, it is desirable that the amount of data for each speaker is small to reduce the cost for constructing speech database and the computational cost for training. It is also desirable that sentence sets of speakers are different from each other to make database to be rich in phonetic and linguistic contexts. However, synthetic speech generated from the average voice model trained using different sentence set for each speaker sounds unnatural compared to the model trained using the same sentence set for all speakers, especially when the amount of training data of each speaker is limited. If the sentence sets for speakers are different from each other, the contexts contained in each speaker’s data are quite different. As a result, some nodes of the decision tree are split not by the context but by the speaker. That is, nodes of the tree do not always have training data from all speakers, and some nodes have data from only one speaker. This will cause degradation of quality of average voice, especially in prosody.

To overcome this problem, in this paper, we propose a new technique for constructing a decision tree used for clustering the average voice model. In the technique, we first train speaker dependent models using multi-speaker speech database, and construct a speaker independent decision tree for context clustering common to these speaker dependent models. All speaker dependent models are clustered using the common decision tree (as shown in Fig. 1). An average voice model is obtained by combining Gaussian pdfs of speaker dependent models at each leaf node of the decision tree. When a node of the decision tree is split, only the context related questions which can split the node for all speaker dependent models is adopted. As a result, all nodes of the decision tree have all speakers’ data.

2. TRAINING OF HMM-BASED SPEECH SYNTHESIS SYSTEM

In the training stage of the HMM-based TTS system [7], spectral parameters and F0 observations are combined into
one observation vector frame by frame, and speaker independent phoneme HMMs, which we refer to as the average voice HMMs, are trained using the observation vectors. Spectrum and F0 are modeled by multi-stream HMMs in which output distributions for spectral and F0 parts are modeled by continuous probability distribution and multi-space probability distribution (MSD) [8], respectively. To model variations of spectrum and F0, phonetic and linguistic contextual factors, such as phoneme identity factors and stress related factors, are taken into account [2]. Then, a decision tree based context clustering technique [9][10] is separately applied to the spectral and F0 parts of the context dependent phoneme HMMs. Finally, state durations are modeled by multi-dimensional Gaussian distributions, and the state clustering technique is applied to the duration models.

3. DECISION TREE FOR CONTEXT CLUSTERING OF AVERAGE VOICE MODEL

In the following, we will describe the case where the MDL (minimum description length) criterion [10] is used for selecting nodes to be split, however, it is also possible to use other criterion such as the ML (maximum likelihood) criterion [9].

Here a model represents a set of leaf nodes in a decision tree. Let $S_0$ be the root node of a decision tree and $U(S_1, S_2, \ldots, S_M)$ be a model defined for the leaf node set $\{S_1, S_2, \ldots, S_M\}$ (see Fig. 1). Different node sets correspond to different models. A Gaussian pdf $\mathcal{N}_{im}$ of speaker $i$ is assigned to each node $S_m$, and the set of Gaussian pdfs of each speaker $i$ for the node set $\{S_1, S_2, \ldots, S_M\}$ is defined as $\lambda_i(S_1, S_2, \ldots, S_M) = \{\mathcal{N}_{i1}, \mathcal{N}_{i2}, \ldots, \mathcal{N}_{iM}\}$.

The log-likelihood of $\lambda_i$ for the training data is given by

$$L(\lambda_i) = \sum_{m=1}^{M} L(\mathcal{N}_{im})$$

$$= -\frac{1}{2} \sum_{m=1}^{M} \Gamma_{im} (K + K \log(2\pi) + \log |\Sigma_{im}|)$$

(1)

where $\Gamma_{im}$ is the total state occupancy count at node $S_m$ for speaker $i$, $K$ is the dimensionality of the data vector, and $\Sigma_{im}$ is the covariance matrix of the Gaussian pdf of speaker $i$ at node $S_m$. Then, using (1), the description length of $\lambda_i$ is given by

$$D(\lambda_i) = -L(\lambda_i) + KM \log W_i + C$$

$$= \frac{1}{2} \sum_{m=1}^{M} \Gamma_{im} (K + K \log(2\pi) + \log |\Sigma_{im}|)$$

$$+ KM \log W_i + C$$

(2)

where $W_i = \sum_{m=1}^{M} \Gamma_{im}$, and $C$ is the code length required for choosing the model which is assumed here to be constant.

Finally, using (1) and (2), we can calculate a description length for model $U$ of average voice model as follows:

$$D(U) = \sum_{i=1}^{I} D(\lambda_i)$$

$$= \frac{1}{2} \sum_{i=1}^{I} \sum_{m=1}^{M} \Gamma_{im} (K + K \log(2\pi) + \log |\Sigma_{im}|)$$

$$+ \sum_{i=1}^{I} KM \log W_i + IC$$

(3)

where $I$ is the total number of speakers.

Suppose that node $S_m$ of model $U$ is split into two nodes $S_{mqy}$ and $S_{mqn}$ by applying question $q$. Let $D(U')$ be the model obtained by splitting $S_m$ of model $U$ by the question $q$. The description length of model $U'$ is calculated as follows:

$$D(U') = \frac{1}{2} \sum_{i=1}^{I} \sum_{m'=1}^{M} \Gamma_{im'} (K + K \log(2\pi) + \log |\Sigma_{im'}|)$$

$$+ \frac{1}{2} \sum_{i=1}^{I} \Gamma_{mqy} (K + K \log(2\pi) + \log |\Sigma_{mqy}|)$$

$$+ \frac{1}{2} \sum_{i=1}^{I} \Gamma_{mqn} (K + K \log(2\pi) + \log |\Sigma_{mqn}|)$$

$$+ \sum_{i=1}^{I} K(M+1) \log W_i + IC$$

(4)

where the number of nodes for $U'$ is $M+1$, $\Gamma_{mqy}$ and $\Gamma_{mqn}$ are the state occupancy counts and $\Sigma_{mqy}$ and $\Sigma_{mqn}$ are

![Figure 1: Speaker independent decision tree based context clustering.](image-url)
the covariance matrices of Gaussian pdfs of speaker $i$ for nodes $S_{iyq}$ and $S_{ijn}$, respectively.

The difference between the description lengths after and before the splitting, is given by the following equation:

$$
\delta_m(q) = D(U') - D(U)
$$

$$
= \frac{1}{2} \sum_{i=1}^{m} \left( \Gamma_{imqy} \log |\Sigma_{imqy}| + \Gamma_{imqn} \log |\Sigma_{imqn}| \right) - \Gamma_{im} \log |\Sigma_{im}| + \sum_{i=1}^{f} K \log W_i.
$$

In node splitting, we first determine the question $q'$ which would minimize $\delta_0(q')$ when used to split root node $S_0$. If $\delta_0(q') > 0$, then no splitting is conducted. If $\delta_0(q') < 0$, then node $S_0$ is split into two nodes, $S_{yq}'$ and $S_{yn}'$, and the same procedure is repeated for each of these two nodes. Note that only the questions which can split all the speaker dependent models are adopted. This node splitting is repeated until there remain no nodes to be split.

4. EXPERIMENTS

To evaluate the performance of the proposed technique, we conducted a paired comparison test for synthetic speech generated from the average voice models trained using the conventional [7] and proposed techniques.

4.1. Experimental Conditions

We used phonetically balanced sentences from ATR Japanese speech database for training HMMs. Based on phoneme labels and linguistic information included in the database, we made context dependent phoneme labels. We used 42 phonemes including silence and pause.

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 [11]. F0 values were obtained using ESPS get_F0 program [12]. 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 models. The average voice models were trained using 50, 100 or 150 sentences for each speaker from 3 male and 3 female speakers’ speech data. In the training of the average voice model, first, speaker and context dependent models were trained. Then, a speaker independent decision tree for context clustering was constructed using the proposed algorithm. A Gaussian pdf of average voice model was obtained by combining all speakers’ Gaussian pdfs at each node of the tree. After the reestimation of parameters of the average voice model using all speakers’ training data, state duration distributions was obtained for each speaker. Finally, state duration distributions of the average voice model was obtained by applying the same procedure.

Subjects were ten males. For each subject, eight test sentences were chosen at random from the 53 test sentences which were not contained in training data. Subjects were presented a pair of average voices synthesized from average voice models trained using conventional and proposed techniques in random order, and asked which synthetic speech sounded more natural.

4.2. Results of Context Clustering

Table 1 shows the number of leaf nodes of the decision trees constructed using conventional and proposed techniques. From Table 1, it can be seen that the number of leaf nodes of the decision tree constructed using the proposed technique is much smaller than conventional technique. This is caused by adopting only the questions which can split all the speaker dependent models, and calculating the term related to the number of training samples in the definition of description length separately for each speaker.

Table 2 shows the number of leaf nodes which did not have all speakers’ training data. (A) the number of leaf nodes lacking one or more speakers’ data and its percentage. (B) the number of leaf nodes which had only one speaker’s data and its percentage.

Table 1. The number of leaf nodes of decision tree.

<table>
<thead>
<tr>
<th>Sentences per Speaker</th>
<th>Conventional Spectrum</th>
<th>F0</th>
<th>Duration</th>
<th>Proposed Spectrum</th>
<th>F0</th>
<th>Duration</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>419</td>
<td>1011</td>
<td>911</td>
<td>132</td>
<td>204</td>
<td>115</td>
</tr>
<tr>
<td>100</td>
<td>670</td>
<td>1674</td>
<td>1416</td>
<td>237</td>
<td>337</td>
<td>174</td>
</tr>
<tr>
<td>150</td>
<td>834</td>
<td>2026</td>
<td>1820</td>
<td>321</td>
<td>475</td>
<td>228</td>
</tr>
</tbody>
</table>

Table 2. The number of leaf nodes which did not have all speakers’ training data. (A) the number of leaf nodes lacking one or more speakers’ data and its percentage. (B) the number of leaf nodes which had only one speaker’s data and its percentage.

<table>
<thead>
<tr>
<th></th>
<th>Conventional</th>
<th>Proposed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spectrum</td>
<td>(A) (B)</td>
<td>(A) (B)</td>
</tr>
</tbody>
</table>
| F0            | 505 (50%)    | 197 (19%)| 0 (0%)   | 0 (0%)
beginning of the sentence had only one female speakers’ training data. On the other hand, there is no clearly unnatural part observed in F0 contour generated from the proposed model.

4.3. Subjective Evaluations

Figure 3 shows the result of the paired comparison test. The horizontal axis indicates the preference score, and the bars indicate the results for the models trained using 50, 100 and 150 sentences for each speaker. From this figure, it can be seen that the average voices generated form proposed models sound more natural than the average voices from conventional models.

Since there are many leaf nodes which did not have all speakers’ training data in the conventional model, there are some parts biased to a speaker or a gender in a sentence to be synthesized, resulting in unnatural synthetic speech especially in prosody. Meanwhile, since there is no leaf node biased in the proposed model, quality of synthetic speech is improved.

5. CONCLUSION

In this paper, we have proposed a new technique for constructing a speaker independent decision tree for an HMM-based speech synthesis system. We have shown that the average voice models constructed using the proposed technique can synthesize more natural sounding speech than the conventional models.

Future work will focus on evaluation of synthetic speech generated using models adapted from average voice models based on the proposed technique. Application of the proposed technique to speech recognition is also our future work.

6. REFERENCES