IEICE TRANS. INF. & SYST., VOL.E86–D, NO.3 MARCH 2003

PAPER *Special Issue on Speech Information Processing*

A Context Clustering Technique for Average Voice Models

Junichi YAMAGISHI[†], ***Student Member***, **Masatsune TAMURA**[†*], **Takashi MASUKO**[†], **Keiichi TOKUDA**[††], ***and*** **Takao KOBAYASHI**[†], ***Regular Members***

SUMMARY This paper describes a new context clustering technique for average voice model, which is a set of speaker independent speech synthesis units. In the technique, we first train speaker dependent models using multi-speaker speech database, and then construct a decision tree common to these speaker dependent models for context clustering. When a node of the decision tree is split, only the context related questions which are applicable to all speaker dependent models are adopted. As a result, every node of the decision tree always has training data of all speakers. After construction of the decision tree, all speaker dependent models are clustered using the common decision tree and a speaker independent model, i.e., an average voice model is obtained by combining speaker dependent models. From the results of subjective tests, we show that the average voice models trained using the proposed technique can generate more natural sounding speech than the conventional average voice models.

key words: *decision tree, context clustering, average voice model, HMM-based speech synthesis, speaker independent 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 in [5]–[7] 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 [8]. In that system, a 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 speakers used for training, we refer to the speaker independent model as the average voice model, and the synthetic speech generated from the average voice model as average voice.

It is thought that quality of the average voice crucially affects quality of synthetic speech generated from adapted models, and that training data of the average voice model affects quality of the average voice. Although it is shown empirically that we can synthesize average voice of good quality from the average voice model trained using a large amount of speech data, recording of a large number of sentences is not an easy task for speakers and the cost for constructing database will be very expensive. To reduce the cost for constructing speech database and the computational cost for training, it is desirable that the amount of data for each speaker is as small as possible. It is also desirable that the individual sentence sets are used for respective speakers to make database rich in phonetic and linguistic contexts. However, synthetic speech generated from the average voice model trained using the individual sentence sets for respective speakers would sound 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 individual sentence sets are used for respective speakers, the contexts contained in each speaker's data are quite different. As a result, after the decision tree based context clustering, the nodes of the tree do not always have training data of 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 context clustering technique for the average

Manuscript received June 30, 2002.
Manuscript revised November 13, 2002.

[†]The authors are with Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology, Yokohama-shi, 226–8502 Japan.
[††]The author is with the Department of Computer Science, Nagoya Institute of Technology, Nagoya-shi, 466–8555 Japan.
[*]Presently, with Toshiba Corporation.

voice model, which will be referred to as shared decision tree context clustering (STC). In the technique, we first train speaker dependent models using multi-speaker speech database, and construct a 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 are applicable to all speaker dependent models are adopted. As a result, every node of the decision tree always has the data of all speakers. Using the common decision tree, all speaker dependent models are clustered and an average voice model is obtained by combining Gaussian pdfs of speaker dependent models at each leaf node of the decision tree.

A similar approach to the proposing technique has been reported for SSS (successive state splitting) in [9], named SP-SSS (speaker parallel SSS), in which the initial state of SSS is split in speaker domain before split in temporal and contextual domains. However, there are some differences between the proposing technique and SP-SSS. In the proposing technique, models for all speakers are always trained separately. Furthermore, it is aimed to construct speaker independent models for speech synthesis.

2. Speech Synthesis System Using Average Voice Model

A block diagram of the HMM-based TTS system [7] is shown in Fig. 1. The system consists of two stages, the training stage and the synthesis stage.

In the training stage, spectral parameters and F0 observations are obtained from multi-speaker speech database, and combined into one observation vector frame by frame. Speaker independent phoneme 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) [10], respectively. 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 [11], [12] 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.

In the synthesis stage, first, an arbitrarily given text to be synthesized is transformed into a context dependent phoneme label sequence. According to the label sequence, a sentence HMM, which represents the whole text to be synthesized, is constructed by concatenating context dependent HMMs. From the sentence HMM, spectral and F0 parameter sequences are obtained using the algorithm for speech parameter generation from HMMs with dynamic features [3], where phoneme durations are determined based on state duration distributions [13]. Finally, by using the MLSA filter [14], speech is synthesized from the generated mel-cepstral and F0 parameter sequences.

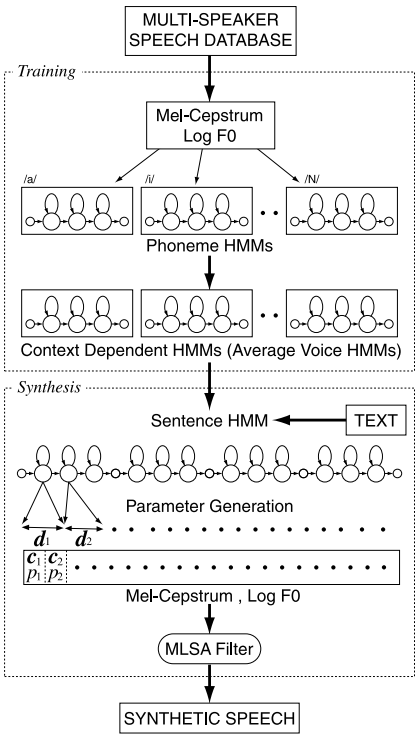


Fig. 1 A block diagram of an HMM-based speech synthesis system using the average voice model.

Since the synthetic speech generated from the speaker independent model can be considered to have averaged voice characteristics and prosodic features of the speakers, we refer to the speaker independent model as the average voice model. Using an appropriate model adaptation technique, it is shown that arbitrary speakers' voice can be generated from average voice model [5]–[7].

In the average voice model of [6], [7] a single Gaussian density was used. In speech recognition, a mixture Gaussian density is usually used to model variations of the parameters caused by variations of speakers in detail for speaker independent models. On the other hand, in speech synthesis using the average voice model, since the average parameters of speakers in the training data should be obtained for the average voice model, it is thought that the use of a single Gaussian density is an appropriate choice. In this paper, therefore, we consider the case where a single Gaussian model is used, as in [6], [7]. However, the extension to the mixture Gaussian density can be done easily [4], [15].

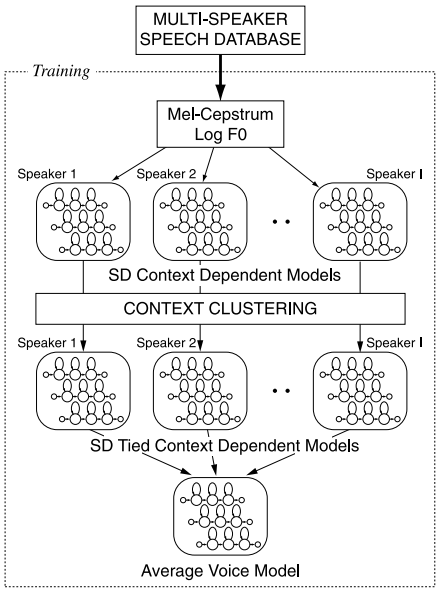


Fig. 2 A block diagram of training stage of the average voice model.

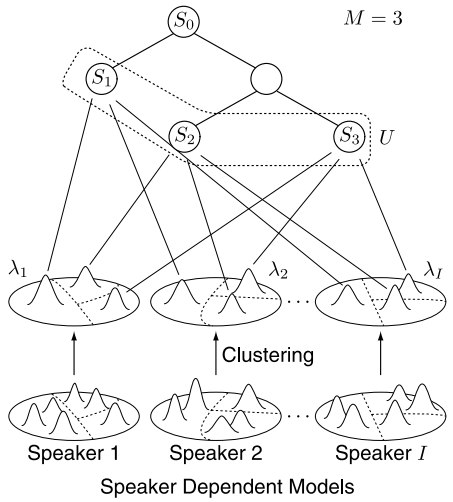


Fig. 3 Context clustering for average voice model using a decision tree common to the speaker dependent models.

3. Shared Decision Tree Context Clustering

3.1 Training of Average Voice Model

A block diagram of the training stage of average voice model using the proposing technique is shown in Fig. 2. First, context dependent models without context clustering are separately trained for respective speakers to derive a decision tree for context clustering common to these speaker dependent models. Then, the decision tree, which we refer to as a shared decision tree, is constructed using an algorithm described in Sect. 3.3 from the speaker dependent models. Finally, all speaker dependent models are clustered using the shared decision tree. A Gaussian pdf of average voice model is obtained by combining all speakers' Gaussian pdfs at every node of the tree. After the reestimation of parameters of the average voice model using training data of all speakers, state duration distributions is obtained for each speaker. Finally, state duration distributions of the average voice model is obtained by applying the same procedure.

3.2 Description Length of Average Voice Model

In the following, we will describe the case where the MDL (minimum description length) criterion [12] is used for selecting nodes to be split. However, it is also possible to use other criteria such as the ML (maximum likelihood) criterion [11].

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. 3). 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 $\{S_1, S_2, \cdots, S_M\}$ is defined as $\lambda_i(S_1, S_2, \cdots, S_M) = \{\mathcal{N}_{i1}, \mathcal{N}_{i2}, \ldots, \mathcal{N}_{iM}\}$.

The log-likelihood of λ_i for the training data is given by

$$L(\lambda_i) = -\frac{1}{2}\sum_{m=1}^{M} \Gamma_{im}\left(K + K\log(2\pi) + \log|\mathbf{\Sigma}_{im}|\right), \quad (1)$$

where K is the dimensionality of the data vector, Γ_{im} and $\mathbf{\Sigma}_{im}$ are the state occupancy count and the covariance matrix of Gaussian pdf of speaker i at node S_m, respectively. Using (1), the description length of λ_i is given by

$$\begin{aligned} D(\lambda_i) &= -L(\lambda_i) + cKM\log W_i + C \\ &= \frac{1}{2}\sum_{m=1}^{M} \Gamma_{im}\left(K + K\log(2\pi) + \log|\mathbf{\Sigma}_{im}|\right) \\ &\quad + cKM\log W_i + C, \end{aligned} \quad (2)$$

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

Using (2), we define the description length for model U as follows:

$$\begin{aligned} \widehat{D}(U) &= \sum_{i=1}^{I} D(\lambda_i) \\ &= \frac{1}{2}\sum_{i=1}^{I}\sum_{m=1}^{M} \Gamma_{im}\left(K + K\log(2\pi) + \log|\mathbf{\Sigma}_{im}|\right) \\ &\quad + c\sum_{i=1}^{I} KM\log W_i + IC, \end{aligned} \quad (3)$$

where I is the total number of speakers.

3.3 Construction of Shared Decision Tree

Suppose that node S_m of model U is split into two nodes S_{mqy} and S_{mqn} by applying a question q. Let 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:

$$\begin{aligned}\widehat{D}(U') = &\frac{1}{2}\sum_{i=1}^{I}\sum_{\substack{m'=1\\ m'\neq m}}^{M}\Gamma_{im'}\left(K + K\log(2\pi) + \log|\boldsymbol{\Sigma}_{im'}|\right)\\ &+ \frac{1}{2}\sum_{i=1}^{I}\Gamma_{imqy}\left(K + K\log(2\pi) + \log|\boldsymbol{\Sigma}_{imqy}|\right)\\ &+ \frac{1}{2}\sum_{i=1}^{I}\Gamma_{imqn}\left(K + K\log(2\pi) + \log|\boldsymbol{\Sigma}_{imqn}|\right)\\ &+ c\sum_{i=1}^{I}K(M+1)\log W_i + IC, \end{aligned} \tag{4}$$

where the number of nodes of U' is $M+1$, Γ_{imqy}, Γ_{imqn} and $\boldsymbol{\Sigma}_{imqy}$, $\boldsymbol{\Sigma}_{imqn}$ are the state occupancy counts and the covariance matrices of Gaussian pdfs of speaker i at nodes S_{mqy} and S_{mqn}, respectively.

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

$$\begin{aligned}\delta_m(q) &= \widehat{D}(U') - \widehat{D}(U)\\ &= \frac{1}{2}\sum_{i=1}^{I}(\Gamma_{imqy}\log|\boldsymbol{\Sigma}_{imqy}| + \Gamma_{imqn}\log|\boldsymbol{\Sigma}_{imqn}|\\ &\quad - \Gamma_{im}\log|\boldsymbol{\Sigma}_{im}|) + c\sum_{i=1}^{I}K\log W_i. \end{aligned} \tag{5}$$

The procedure of construction of the shared decision tree is summarized as follows:

1. Define a set composed of root node S_0 as model U, i.e., $U = \{S_0\}$.
2. Find the node $S_{m'}$ in model U and the question q' which minimizes $\delta_{m'}(q')$.
3. Terminate if $\delta_{m'}(q') > 0$.
4. Split the node $S_{m'}$ by the question q', and replace U by the resultant node set.
5. Go to step 2.

Note that only the questions which are applicable to all speaker dependent models are adopted in step 2. The last term on the right-hand side of (5) corresponds to the node splitting threshold for the increase in log-likelihood.

After the construction of the shared decision tree, we obtain Gaussian pdfs of the average voice model by combining Gaussian pdfs of speaker dependent models. The mean vector $\boldsymbol{\mu}_m$ and the covariance matrix $\boldsymbol{\Sigma}_m$ of the Gaussian pdf at node S_m are calculated as follows:

$$\boldsymbol{\mu}_m = \frac{\sum_{i=1}^{I}\Gamma_{im}\boldsymbol{\mu}_{im}}{\sum_{i=1}^{I}\Gamma_{im}} \tag{6}$$

$$\boldsymbol{\Sigma}_m = \frac{\sum_{i=1}^{I}\Gamma_{im}\left(\boldsymbol{\Sigma}_{im} + \boldsymbol{\mu}_{im}\boldsymbol{\mu}_{im}^{\top}\right)}{\sum_{i=1}^{I}\Gamma_{im}} - \boldsymbol{\mu}_m\boldsymbol{\mu}_m^{\top} \tag{7}$$

where $\top$ denotes matrix transpose, and $\boldsymbol{\mu}_{im}$ is the mean vector of the Gaussian pdf of speaker i at node S_m.

4. Experiments

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 16 kHz and windowed by a 25ms Blackman window with a 5ms shift. Then mel-cepstral coefficients were obtained by mel-cepstral analysis [14], [16]. F0 values were extracted using ESPS `get_F0` program [17]. 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 from 50 to 300 sentences of each speaker's speech data. Speakers were 3 females (FKN, FKS, FYM) and 3 males (MHO, MHT, MYI). Tables 1 and 2 show the number of sentences per speaker and corresponding sentence sets used for training. The sentence sets A–I consist of 50 sentences,

Table 1 Sentences per speaker and sentence sets used for training for the case of the same sentence sets.

Sentences per Speaker	Female			Male		
	FKN	FKS	FYM	MHO	MHT	MYI
50	A	A	A	A	A	A
100	A, B	A, B	A, B	A, B	A, B	A, B
150	A-C	A-C	A-C	A-C	A-C	A-C
200	A-D	A-D	A-D	A-D	A-D	A-D
250	A-E	A-E	A-E	A-E	A-E	A-E
300	A-F	A-F	A-F	A-F	A-F	A-F
450	A-I	A-I	A-I	A-I	A-I	A-I

Table 2 Sentences per speaker and sentence sets used for training for the case of the individual sentence sets.

Sentences per Speaker	Female			Male		
	FKN	FKS	FYM	MHO	MHT	MYI
50	A	B	C	D	E	F
100	A, B	B, C	C, D	D, E	E, F	F, G
150	A-C	B-D	C-E	D-F	E-G	F-H
200	A-D	B-E	C-F	D-G	E-H	F-I
250	A-E	B-F	C-G	D-H	E-I	A, F-I
300	A-F	B-G	C-H	D-I	A, E-I	A, B, F-I

respectively. Table 1 shows the case in which the same sentence sets were used for all speakers, and Table 2 shows the case in which the individual sentence sets were used for respective speakers. A model trained using 450 sentences per speaker is used as a reference model of the subjective evaluations in Sect. 4.4. The following contextual factors are taken into account:

- the number of morae in sentence
- position of breath group in sentence
- the number of morae in {preceding, current, succeeding} breath group
- position of current accentual phrase in current breath group
- the number of morae and accent type in {preceding, current, succeeding} accentual phrase
- {preceding, current, succeeding} part-of-speech
- position of current mora in current accentual phrase
- difference between position of current mora and accent type
- {preceding, current, succeeding} phoneme

It is noted that a unit of position is mora.

Average voice models are trained using the conventional technique (described in Sect. 2) [6], [7] and the proposed technique (described in Sect. 3). In the proposed technique, the total number of parameters of all speaker dependent models is considered, while the number of parameters of only one speaker independent model is considered in the conventional technique. This causes increase of the last term on the right-hand side of (5), and results in a higher node splitting threshold for the increase in log-likelihood than the conventional technique. Consequently, if the weighting factor c of the description length of the proposed technique is set to unity, as in the conventional technique, the number of the leaf nodes of the proposed decision tree becomes considerably small. The considerable decrease of the leaf nodes of the decision tree makes the synthetic speech unnatural. Therefore, we adjust the weighting factor c to increase the number of leaf nodes of the proposed decision tree. From the results of preliminary experiments, we set the weighting factor c of the description length to 1 for the conventional models and 0.4 for the proposed models, respectively.

4.2 Results of Context Clustering

Tables 3 and 4 show the number of leaf nodes of the decision trees constructed using the conventional and proposed techniques. Table 3 shows the result for the case of the same sentence sets, and Table 4 shows the results for the case of the individual sentence sets.

Table 5 shows the number of leaf nodes which did not have training data of all speakers and its percentage when the average voice models were trained using the individual 50-sentence sets of each speaker. In Table 5, (A) shows the number of leaf nodes lacking one or more speakers' data and its percentage, and (B) shows the number of leaf nodes which had only one speaker's data and its percentage. From Table 5, it can be seen that 50% of leaf nodes of the conventional decision tree for F0 lacked one or more speakers' data and 19% of leaf nodes had only one speaker's data. On the other hand, theoretically, every leaf node of the proposed decision tree has the training data of all speakers. Therefore, there is no node lacking one or more speakers' data.

Table 3 The number of leaf nodes of decision trees for the case of the same sentence sets.

Sentences per Speaker	Conventional			Proposed		
	Spec.	F0	Dur.	Spec.	F0	Dur.
50	405	690	584	608	907	811
100	622	1126	934	998	1597	1372
150	799	1418	1287	1339	1605	1825
200	1004	1794	1660	1599	2578	2230
250	1163	2060	1923	1895	2977	2754
300	1310	2270	2271	2138	3433	3098
450	1697	2887	2892			

Table 4 The number of leaf nodes of decision trees for the case of the individual sentence sets.

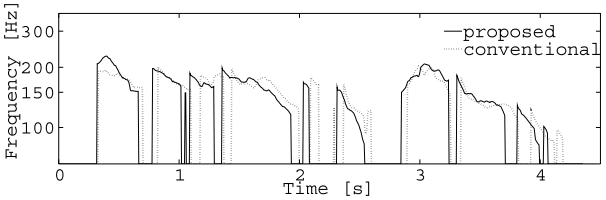
Sentences per Speaker	Conventional			Proposed		
	Spec.	F0	Dur.	Spec.	F0	Dur.
50	419	1011	911	548	818	814
100	670	1674	1416	913	1416	1497
150	834	2026	1820	1252	2009	2073
200	1015	2261	1974	1504	2438	2502
250	1158	2419	2293	1779	2908	2931
300	1284	2472	2369	2002	3257	3203

Table 5 The number of leaf nodes which did not have training data of all speakers. (A) shows the number of leaf nodes lacking one or more speakers' data and its percentage. (B) shows the number of leaf nodes which had only one speaker's data and its percentage.

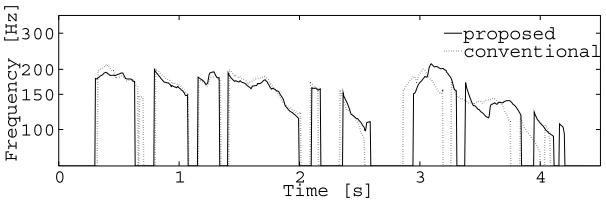
	Conventional		Proposed	
	(A)	(B)	(A)	(B)
Spectrum	37 (8%)	14 (3%)	0 (0%)	0 (0%)
F0	505 (50%)	197 (19%)	0 (0%)	0 (0%)

Figures 4 and 5 show examples of generated F0 contours for a Japanese sentence /he-ya-i-ppa-i-ni-ta-ba-ko-no-no-mu-ga-ta-chi-ko-me-pau-yu-ru-ya-ka-ni-u-go-i-te-i-ru/ (meaning "Cigarette smoke fills the whole room, and is moving gently," in English) which is not included in the training sentences. Figure 4 shows the result for the case of the same sentence sets, and Fig. 5 shows the results for the case of the individual sentence sets. In Figs. 4 and 5, (a) and (b) show the F0 contours generated from the average voice models trained using 50 sentences and 300 sentences per speaker. Dotted line and solid line show the F0 contours generated from the average voice models clustered using conventional and proposed techniques, respectively.

From Fig. 4, it can be seen that the conventional and proposed techniques provide similar results when

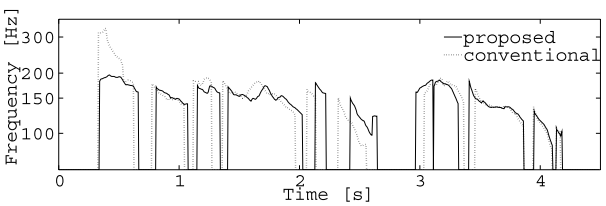


(a) 50 sentences per speaker.

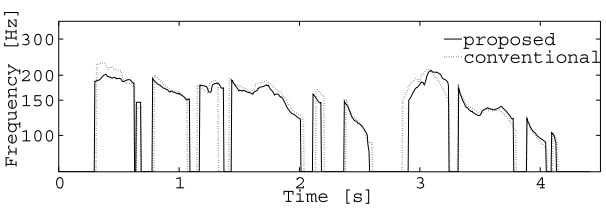


(b) 300 sentences per speaker.

Fig. 4 Comparison of F0 contours generated from average voice models constructed using conventional and proposed techniques for the case of the same sentence sets.



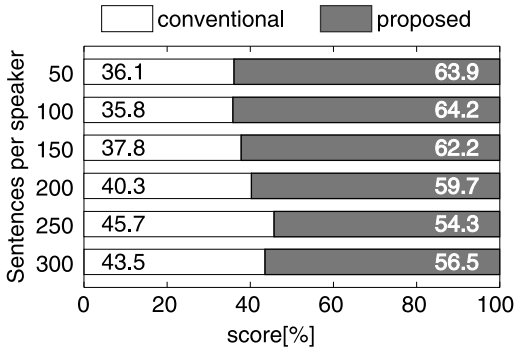
(a) 50 sentences per speaker.



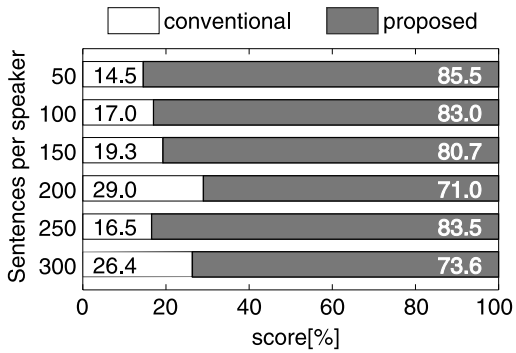
(b) 300 sentences per speaker.

Fig. 5 Comparison of F0 contours generated from average voice models constructed using conventional and proposed techniques for the case of the individual sentence sets.

the same sentence sets were used. This is because the intersection of context sets contained in the respective speakers' training data are large when the sentence sets were the same[†]. On the other hand, from Fig. 5 (a), we can see that the F0 contours generated from the conventional and proposed models are quite different at the beginning of the sentence; the values of F0 generated from the conventional model are unnaturally high, whereas there is no obviously unnatural part in the F0 contour generated from the proposed model. This is due to the fact that leaf nodes of the conventional model corresponding to the beginning of the sentence had only one female speaker's training data. However, from Fig. 5 (b), we can see that there is no significant difference between the F0 contours generated from the conventional and proposed models. This is due to the fact that the size of the intersection of sentence sets



(a) The same sentence sets.



(b) The individual sentence sets.

Fig. 6 Result of the paired comparison test.

increases as the number of sentences for each speaker increases. For example, when the number of sentences for each speaker is 300, the sentence set F is included by all speakers' sentence sets. As a result, the number of leaf nodes biased to a speaker or a gender decreases in the conventional model.

4.3 Subjective Evaluations

We conducted paired comparison tests for synthetic speech generated from the average voice models trained using the conventional and proposed techniques. Subjects were eleven males. For each subject, eight test sentences were chosen at random from the 53 test sentences which were not contained in the 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.

Figure 6 shows the results of the paired comparison test. In Fig. 6, (a) shows the results for the case of the same sentence sets, and (b) shows the results for the case of the individual sentence sets. The horizontal axes indicate the preference score, and the bars indicate the

[†]The context sets of respective speakers' data do not always the same even if the sentence sets are the same, since some contextual factors, such as position of pause and accentual type, are not determined by text and vary depending on speakers.

results for the models trained using 50, 100, 150, 200, 250 and 300 sentences per speaker, respectively.

From these figures, it can be seen that the average voice generated from the proposed models sound more natural than the average voice from conventional models regardless of the number of training sentences and sentence sets. It can also be seen that differences between the scores of proposed and conventional models are greater in the case of the individual sentence sets for respective speakers than the case of the same sentence sets. Moreover, the difference becomes greater as the number of training sentences decreases. Especially, when individual sentence set for respective training speakers were used and the number of sentences for each speaker is less than 150, the scores of the proposed technique attained more than 80%.

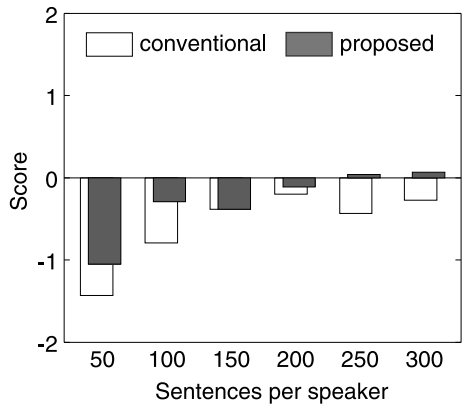
This can be due to the following reason. When the sentence sets for respective training speakers are different, context sets of respective speakers' data become quite different, and the intersection of context sets becomes smaller as the number of sentences decreases. Even if the same sentence set is used for all speakers, the context sets of respective speakers' data are not usually identical. Using the conventional technique, as the context sets of respective speakers' data becomes more different, the number of leaf nodes lacking one or more speakers' data increases, and quality of average voice generated from conventional models tends to degrade. On the other hand, since the proposed technique is robust to difference of context sets between training speakers' data, quality of average voice generated from proposed models does not degrade seriously.

4.4 Comparison to the Model Trained Using a Large Amount of Speech Data

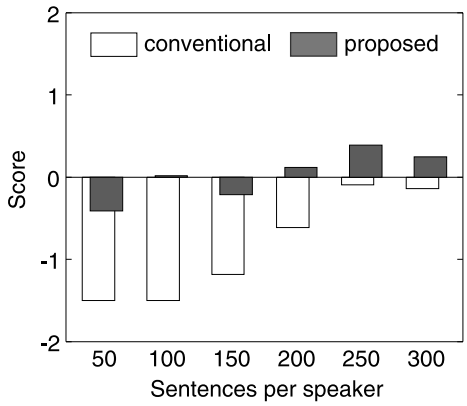
We conducted a comparison category rating test to evaluate the naturalness of the average voice generated from the model trained using the proposed technique. As a reference model, we used a conventional model trained using 450 sentences (sentence set A–I) per a speaker. A subject was required to judge quality of test speech on a seven point scale (3: much better, 2: better, 1: slightly better, 0: almost the same, −1: slightly worse, −2: worse, −3: much worse) compared to reference speech on naturalness and intelligibility. Subjects were seven males. For each subject, four test sentences were chosen at random from the 53 test sentences which were not contained in training data.

Figure 7 shows the result of the evaluation of the naturalness. In Fig. 7, (a) shows the results for the case of the same sentence sets, and (b) shows the results for the case of the individual sentence sets. The vertical axes indicate the average score and the horizontal axes indicate the number of sentences.

From this figure, it is seen that naturalness of the average voice of the proposed technique is higher than



(a) The same sentence sets.



(b) The individual sentence sets.

Fig. 7 Result of evaluation of naturalness.

the conventional technique. Comparing Figs. 7 (a) and (b), when the number of sentences is limited, scores for the conventional models using the individual sentence sets are lower than the conventional models using the same sentence sets, whereas scores for proposed models using the individual sentence sets are higher than proposed models using same sentence sets. Moreover, using proposed technique and the individual sentence sets, there is only a little degradation on naturalness of the average voice even when training data is limited. In fact, the average voice which was trained using only 50 sentences per speaker is almost equivalent on naturalness to the average voice trained using 450 sentences by the conventional technique. This is due to the facts that difference between context sets does not cause degradation of naturalness of average voice for proposed models, and that when the size of database is almost the same, the average voice model trained using "context-rich" database can generate more natural sounding speech than model trained using "context-poor" database.

4.5 Evaluations of the Model with F0 Normalization

To show effectiveness of the proposed technique, we

Table 6 Result of the evaluation of average voice model trained using speech data with F0 normalization. Score shows the average number of sentences which are judged to be clearly unnatural.

Clustering Method	Conventional		Proposed	
F0 Normalization	No	Yes	No	Yes
Score	21.0	14.2	7.0	1.3

compared it with an F0 normalization technique. F0 normalization was achieved by shifting F0 contours in logarithmic domain so that the mean value of F0 of each speaker is equal to mean value of F0 of all training speakers. Then average voice models were trained using individual 50-sentence sets. Subjects were five males and required to judge whether or not test speech was clearly unnatural. The test sentences were 53 sentences which were not contained in the training data.

Table 6 shows the result of the evaluation. In the table, each score shows the average number of sentences which are judged to be clearly unnatural. It can be seen that the average voices using the training data with F0 normalization sound more natural than those without F0 normalization. It is due to the fact that the influence of leaf nodes biased to a speaker or a gender is reduced in the decision tree of F0. It can also be seen that the average voices using the proposed technique sound more natural than the conventional technique with the F0 normalization. It has been observed from the informal listening tests that the proposed technique reduces the influence of leaf nodes biased to a speaker or a gender in the decision tree of spectrum and state duration, as well as F0.

5. Conclusion

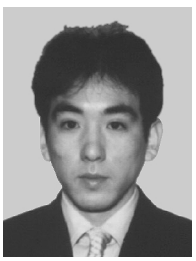
In this paper, we have proposed a new context clustering technique, named shared decision tree context clustering, for an HMM-based speech synthesis system. An advantage of the technique is that every node of the decision tree always has the data of all speakers. In other words, there is no node lacking one or more speakers' data. 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. Training using the proposed technique and SAT (Speaker Adaptive Training) [18] at the same time is also our future work.

References

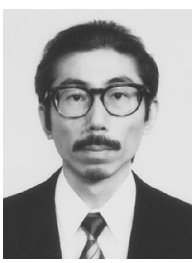
[1] T. Masuko, K. Tokuda, T. Kobayashi, and S. Imai, "Speech synthesis using HMMs with dynamic features," Proc. ICASSP-96, pp.389–392, May 1996.
[2] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis," Proc. EUROSPEECH-99, pp.2374–2350, Sept. 1999.
[3] K. Tokuda, T. Kobayashi, and S. Imai, "Speech parameter generation from HMM using dynamic features," Proc. ICASSP-95, pp.660–663, May 1995.
[4] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," Proc. ICASSP 2000, pp.1315–1318, June 2000.
[5] M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi, "Speaker adaptation for HMM-based speech synthesis system using MLLR," The Third ESCA/COCOSDA Workshop on Speech Synthesis, pp.273–276, Nov. 1998.
[6] M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi, "Adaptation of pitch and spectrum for HMM-based speech synthesis using MLLR," Proc. ICASSP 2001, pp.805–808, May 2001.
[7] M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi, "Text-to-speech synthesis with arbitrary speaker's voice from average voice," Proc. EUROSPEECH 2001, pp.345–348, Sept. 2001.
[8] C.J. Leggetter and P.C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," Computer Speech and Language, vol.9, no.2, pp.171–185, 1995.
[9] J. Takami, T. Kosaka, and S. Sagayama, "Automatic generation of speaker-common hidden Markov network by adding the speaker splitting domain to the Successive State Splitting algorithm," Proc. 1992 Autumn Meeting of the Acoustical Society of Japan, 3-1-8, pp.155–156, Oct. 1992.
[10] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi, "Hidden Markov models based on multi-space probability distribution for pitch pattern modeling," Proc. ICASSP-99, pp.229–232, March 1999.
[11] S.J. Young, J. Odell, and P. Woodland, "Tree-based state tying for high accuracy acoustic modeling," Proc. ARPA Human Language Technology Workshop, pp.307–312, March 1994.
[12] K. Shinoda and T. Watanabe, "MDL-based context-dependent subword modeling for speech recognition," J. Acoust. Soc. Jpn. (E), vol.21, pp.79–86, March 2000.
[13] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Duration modeling for HMM-based speech synthesis," Proc. ICSLP-98, pp.29–32, Dec. 1998.
[14] T. Fukada, K. Tokuda, T. Kobayashi, and S. Imai, "An adaptive algorithm for mel-cepstral analysis of speech," Proc. ICASSP-92, pp.137–140, March 1992.
[15] K. Tokuda, T. Masuko, T. Yamada, T. Kobayashi, and T. Kitamura, "An algorithm for speech parameter generation from continuous mixture HMMs with dynamic features," Proc. EUROSPEECH-95, pp.757–760, Sept. 1995.
[16] K. Tokuda, T. Kobayashi, T. Fukada, H. Saito, and S. Imai, "Spectral estimation of speech based on mel-cepstral representation," IEICE Trans. Fundamentals (Japanese Edition), vol.J74-A, no.8, pp.1240–1248, Aug. 1991.
[17] ESPS Programs Version 5.0, Entropic Research Laboratory, 1993.
[18] T. Anastasakos, J. McDonough, R. Schwartz, and J. Makhoul, "A compact model for speaker adaptive training," Proc. ICSLP-96, pp.1137–1140, Oct. 1996

Junichi Yamagishi received the B.E. degree in electrical and electronic engineering from Tokyo Institute of Technology, Tokyo, Japan, in 2002. He is currently a graduate student of the Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology, Yokohama, Japan. His research interests include speech synthesis and speech recognition. He is a member of ISCA and ASJ.



Masatsune Tamura received the B.E. degree in electrical and electronic engineering, and M.E., and Dr.Eng. degrees in intelligence science from Tokyo Institute of Technology, Tokyo, Japan, in 1997, 1999 and 2002, respectively. His research interests include speech synthesis, speech coding, and multimodal interface. He is currently a researcher with Toshiba Corporation, Tokyo, Japan. He is a member of ASJ.

Takashi Masuko received the B.E. degree in computer science, and M.E. degrees in intelligence science from Tokyo Institute of Technology, Tokyo, Japan, in 1993 and 1995, respectively. In 1995, he joined the Precision and Intelligence Laboratory, Tokyo Institute of Technology as a Research Associate. He is currently a Research Associate at the Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology, Yokohama, Japan. He is a co-recipient of both the Best Paper Award and the Inose Award from the IEICE in 2001. His research interests include speech synthesis, speech recognition, speech coding, and multimodal interface. He is a member of IEEE, ISCA and ASJ.



Keiichi Tokuda was born in Nagoya, Japan, in 1960. He received the B.E. degree in electrical and electronic engineering from the Nagoya Institute of Technology, Nagoya, Japan, the M.E. and Dr.Eng. degrees in information processing from the Tokyo Institute of Technology, Tokyo, Japan, in 1984, 1986, and 1989, respectively. From 1989 to 1996 he was a Research Associate at the Department of Electronic and Electric Engineering, Tokyo Institute of Technology. Since 1996 he has been with the Department of Computer Science, Nagoya Institute of Technology as Associate Professor. He is a co-recipient of both the Best Paper Award and the Inose Award from the IEICE in 2001, and the TELECOM System Technology Prize from the Telecommunications Advancement Foundation Award, Japan, in 2001. His research interests include speech coding, speech synthesis and recognition and multimodal signal processing. He is a member of IEEE, ISCA, IPSJ, ASJ and JSAI.

Takao Kobayashi received the B.E. degree in electrical engineering, the M.E. and Dr.Eng. degrees in information processing from Tokyo Institute of Technology, Tokyo, Japan, in 1977, 1979, and 1982, respectively. In 1982, he joined the Research Laboratory of Precision Machinery and Electronics, Tokyo Institute of Technology as a Research Associate. He became an Associate Professor at the same Laboratory in 1989. He is currently a Professor of the Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology, Yokohama, Japan. He is a co-recipient of both the Best Paper Award and the Inose Award from the IEICE in 2001, and the TELECOM System Technology Prize from the Telecommunications Advancement Foundation Award, Japan, in 2001. His research interests include speech analysis and synthesis, speech coding, speech recognition, and multimodal interface. He is a member of IEEE, ISCA, IPSJ and ASJ.