Statistical Model Training Technique for Speech Synthesis Based on Speaker Class

Yusuke Ijima, Noboru Miyazaki, Hideyuki Mizuno

NTT Media Intelligence Laboratories, NTT Corporation, Japan

Abstract

To allow the average-voice-based speech synthesis technique to generate synthetic speech that is more similar to that of the target speaker, we propose a model training technique that introduces the label of *speaker class*. Speaker class represents the voice characteristics of speakers. In the proposed technique, first, all training data are clustered to determine classes of speaker type. The average voice model is trained using the labels of conventional context and speaker class. In the speaker adaptation process, the target speaker's class is estimated and is used to transform the average voice model into the target speaker's model. As a result, the speech of the target speaker is synthesized from the target speaker's model and the estimated target speaker's speaker class. The results of an objective experiment show that the proposed technique significantly reduces the RMS errors of log F0. Moreover, the results of a subjective experiment indicate that the proposal yields synthesized speech with better similarity than the conventional method.
Index Terms: HMM-based speech synthesis, average voice model, speaker adaptation, speaker clustering

1. Introduction

Recent research on text-to-speech synthesis has focused on supporting arbitrary speakers given only a small amount of the target speaker's speech data. In HMM-based speech synthesis systems [1], the average-voice-based speech synthesis technique with model adaptation [2] has been proposed. Given only a few minutes of the target speaker's speech data, this technique can synthesize arbitrary texts by transforming the average voice model to the target speaker's model.However, it has been reported that the similarity of the synthesized speech to the target speaker's speech is degraded by model conversion if the acoustic features of the average voice model are distant from those of the target speaker [3]. One useful solution is creating an average voice model whose characteristics are close to those of the target speaker.

To realize this approach, a similar speaker selection based model training technique has been proposed [4]. In this technique, synthetic speech is made closer to that of the target speaker by training an average voice model from perceptually similar speakers (manually selected); note that speaker selection decreases the amount of training data. However, in the case of automatic similar speaker selection using acoustic features, it was reported that the similarity of synthesized speech is degraded due to this reduction. Although, these results indicate that the selection must identify perceptually similar speakers to improve the similarity of the synthesized speech, it is well known that this selection is very difficult. To avoid these problems, i.e., selecting perceptually-similar speakers and offsetting the decrease in amount of training data, it is desirable to create one average voice model that can take into account of multiple speaker characteristics with no decrease in the amount of training data.

So that model training can take into account of the various characteristics of the training data, some studies have proposed a model training technique that adds characteristics of training data to the usual context set of phonetic, prosodic, and linguistic features. [5] proposed a style-mixed modeling technique that utilizes speaking styles and emotional expressions as context. In addition, the gender-mixed modeling technique, which uses speaker gender as an additional context, was proposed to enhance average-voice-based speech synthesis, and its effectiveness was shown [6]. In this study, we propose to add *speaker class*, which better represents detailed speaker characteristics than gender, to the average-voice-based speech synthesis technique.

In the proposed technique, a speaker clustering technique is applied to the training data so as to group the acoustic features of all speakers used for average voice model training. The average voice model is trained using the label of speaker class. In the speaker adaptation process, the target speaker's speaker class is estimated, and the average voice model is transformed to the target speaker's model using the labels of conventional context and the estimated speaker class. The key to realizing our proposal is the robust estimation of the target speaker's class. If complex features that have high correlation with perceptual similarity are used for speaker clustering, we would face the same problem of perceptually-similar speaker selection as in [4]. To avoid this problem, we use very simple acoustic features of spectrum, F0, and phoneme duration for speaker clustering and speaker class estimation. Objective and subjective evaluations show the effectiveness of the proposed technique.

2. Speech synthesis system with speaker class label

2.1. Overview of proposed speech synthesis system

A block diagram of the proposed speech synthesis method is shown in Fig. 1. The proposed technique first trains an average voice model using training data labeled with speaker class and other conventional contexts. The speaker class of the target speaker is estimated from the speaker's training data and input to the average voice model to transform it to better suit the target speaker. Given the estimated target speaker's class and other conventional contexts, the target speaker's model synthesizes the target speaker's speech. The overall process of training, adaptation, and speech synthesis is summarized below.

Training part:

Step 1 Apply the speaker clustering technique to all training data and define a finite number of speaker classes.

Step 2 Train an average voice model using the training data

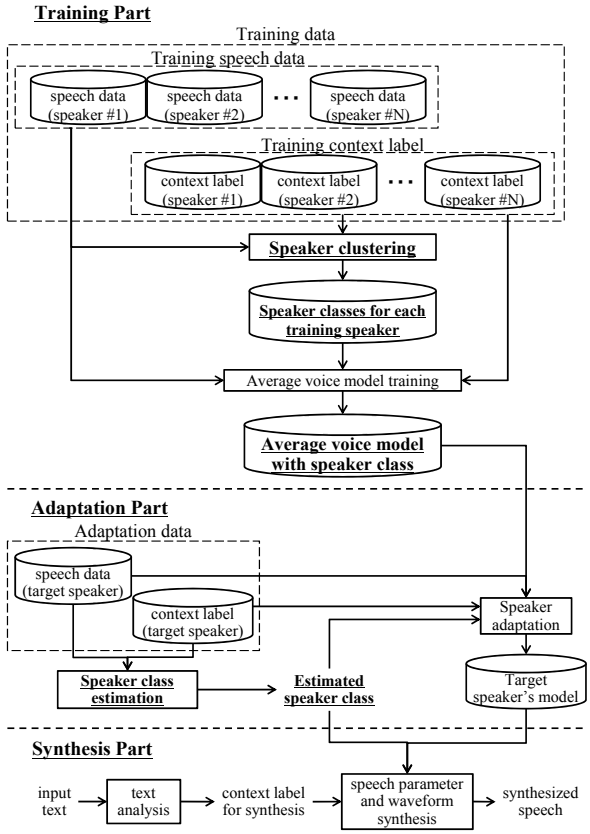

Figure 1: *Block diagram of the speech synthesis system.*

with labels of speaker class and conventional contexts.

Adaptation part:

Step 3 Estimate the speaker class of the target speaker using adaptation data.

Step 4 Transform the trained average voice model into the target speaker's model using the adaptation data and the estimated speaker class.

Synthesis part:

Step 5 Generate the context label from the result of text analysis.

Step 6 Generate the speech parameter sequence of the target speaker using the target speaker's model, the estimated speakers class and the generated context label.

Step 7 Synthesize the speech waveform of the target speaker.

In the proposed technique, if speaker class is estimated correctly, the leaf nodes that have similar speech characteristics to those of the target speaker are used for speaker adaptation and speech synthesis. Therefore, the output of the proposed technique is closer to the target speaker than is possible with the conventional technique. Details of this technique are described below.

2.2. Speaker class

We apply a speaker clustering technique to cluster the acoustic features of the speakers in the training data. For this, it might be thought necessary to use acoustic features that are highly correlated with perceptual similarity. Many previous studies showed that perceptual similarity is influenced by prosodic features, consisting of F0 and phoneme duration, and acoustic features, consisting of cepstral coefficients and the aperiodic component [4, 7–9]. However, since the key to realizing this approach is estimating the speaker class of the target speaker robustly, such complex features are not desirable.

In this study, in order to robustly estimate the speaker class of the target speaker, we utilize three simple features, average mel-cepstral coefficients, average logarithmic F0 (log F0), and speaking rate; they represent the features of spectrum, F0, and phoneme duration respectively. Speaker clustering, based on the k-means algorithm, is performed for each of the three features in isolation. The three acoustic features are described as follows.

2.2.1. Average mel-cepstral coefficients

Average mel-cepstral coefficients of all training data are used for spectrum-based speaker clustering. Because spectrum-oriented speaker characteristics are chiefly presented by voiced phonemes rather than unvoiced phonemes, the average mel-cepstral coefficients are obtained from only voiced frames as detected by TEMPO [10].

2.2.2. Average log F0

Average log F0 of all training data are used for F0-based speaker clustering. As per Sect. **2.2.1**, the average log F0 was obtained from only voiced frames as detected by TEMPO [10].

2.2.3. Speaking rate

Average speaking rates of all training data are used for phoneme-duration-based speaker clustering. The speaking rate is obtained from manually segmented phoneme boundaries of all training data. The speaking rate of speaker i (SR_i) is given by

$$SR_i = \frac{Mora_i}{UttLen_i} \quad (1)$$

where, $Mora_i$ and $UttLen_i$ represent, respectively, the number of mora of speaker i and the utterance length of speaker i.

2.3. Context clustering using speaker class label

Generally, in the average voice model training, decision tree-based context clustering for each model, i.e., mel-cepstrum, log F0, and phoneme duration, is performed using common questions. However, since our proposal adds speaker class to the other conventional contexts, using common questions may lead to negative effects on the tree structure. To avoid this problem, we also use model-specific questions. For instance, questions intended to identify the speaker class (speech rate) are used for context clustering for the phoneme duration model only. In this paper, the context clustering was performed before SAT.

2.4. Estimating speaker class of target speaker

To estimate the speaker class of target speaker, we use the very simple approach of Euclidean distance between the target speaker's features and the centroids of all clusters. Given the adaptation data of the target speaker, we first obtain the three features for the speaker, i.e., the average mel-cepstral coefficients, average log F0, and average speaking rate. Finally, the

Table 1: *The number of leaf nodes of decision trees for each feature.*

# of speaker class	model		
	mel-cepstrum	log F0	duration
1 (conventional)	4954	20941	2971
2	5260	29454	3054
4	5766	35120	2939
8	6607	37952	2952

Table 2: *Mel-cepstral distortion [dB] between original and synthetic speech for each target speaker (closed target speaker).*

# of speaker class	target speaker				
	#1	#2	#3	#4	#5
1	5.78	**5.90**	**5.85**	5.28	**5.83**
2	5.76	5.93	5.87	5.27	5.92
4	**5.73**	5.93	5.91	5.27	5.85
8	5.75	5.96	5.90	**5.25**	5.87

Table 3: *RMS errors of log F0 [cent] between original and synthetic speech for each target speaker (closed target speaker).*

# of speaker class	target speaker				
	#1	#2	#3	#4	#5
1	203.3	216.3	218.8	135.8	182.1
2	202.9	209.4	211.2	133.6	174.1
4	187.1	**199.5**	**204.3**	131.9	172.8
8	**183.1**	203.7	212.2	**127.2**	**169.1**

Table 4: *RMS errors of phoneme duration [ms] between original and synthetic speech for each target speaker (closed target speaker).*

# of speaker class	target speaker				
	#1	#2	#3	#4	#5
1	25.18	24.28	25.16	30.15	24.66
2	24.43	24.37	25.47	29.46	25.42
4	**23.55**	23.08	25.24	**28.70**	**24.19**
8	23.56	**22.73**	**24.66**	29.54	26.07

three subclasses (one per feature) of the target speaker are estimated to be those that have the smallest Euclidean distance between the input feature and cluster centroids.

3. Experiments

3.1. Experimental conditions

In the following experiments, we used the speech data gathered from 88 non-professional Japanese female speakers'. This database contains about 120 phonetically balanced sentences for each speaker. The speakers' ages ranged from 18 to 39.

The sampling frequency of the speech was 16 kHz and the quantization bit rate was 16 bits. We used STRAIGHT analysis [10] for speech feature extraction, and extracted spectral envelope, F0, and aperiodic components. The analysis frame shift was 5 ms. The spectral envelope was then converted to mel-cepstral coefficients using a recursion formula. The aperiodicity feature was also converted to average values for five frequency sub-bands, i.e., 0–1, 1–2, 2–4, 4–6, and 6–8 kHz. As a result, the feature vector consisted of 40 mel-cepstral coefficients including the zeroth coefficient, log F0, and five-band aperiodicity values with delta and delta-delta coefficients. The total dimensionality was 138. We used a five-state left-to-right hidden semi-Markov model with no skip topology. The output distribution in each state was modeled as a single Gaussian density function, and the covariance matrices were assumed to be diagonal.

For training the average voice model, one hundred sentences uttered by 85 of the 88 speakers were used. For its adaptation to the target speaker, twenty sentences uttered by the target speaker were used as adaptation data. We used the combined technique of CSMAPLR and MAP adaptation as the speaker adaptation algorithm [11].

In order to evaluate the effectiveness of the proposed speaker class approach, we created 4 trained average voice models with different numbers of speaker class, 1, 2, 4, and 8. The average voice model with 1 class represents the conventional average voice model.

3.2. The number of leaf nodes in the decision trees

In order to confirm the impact of the speaker class proposal has on the model structure, we investigated the number of leaf nodes in the decision trees for each of the four models. Table 1 lists the number of leaf nodes for each average voice model and each acoustic feature. The number of leaf nodes for the aperiodic feature is not shown because speaker class context determined from the aperiodic feature is not used in speaker clustering. We can see that the number of leaf nodes increases as the number of speaker class increases except for phoneme duration. This is because the amount of training data for phoneme duration is smaller than that for the other two features.

Furthermore, from the decision trees of each model, speaker classes associated with average log F0 tended to be split at the node close to the root node of the tree. On the other hand, speaker classes associated with the two other features tend to be split at nodes close to leaf nodes.

3.3. Objective evaluation

To objectively evaluate the proposed technique, we measured the mel-cepstral distortion, the RMS errors of log F0, and the RMS errors of phoneme duration between original and synthetic speech. To evaluate the influence of speaker class estimation, we used two types of target speakers (closed and open target speakers). The five closed target speakers were among those used for average voice model training, and their speaker classes for speaker adaptation were given correctly. The three open target speakers were not included in the average voice model training, and their speaker classes were estimated automatically. These eight speakers have different speaker classes about at least one feature. Twenty sentences uttered by each target speaker and used as the reference data were not included in the average voice model training data or speaker adaptation.

Table 2–4 show, respectively, the mel-cepstral distortion, the RMS errors of log F0, and the RMS errors of phoneme duration for each closed target speaker and each average voice model. From these results, we can see that the RMS errors of log F0 and phoneme duration are decreased by increasing

Table 5: *Mel-cepstral distortion [dB] between original and synthetic speech for each target speaker (open target speaker).*

# of speaker class	target speaker #1	#2	#3
1	**5.39**	5.87	6.03
2	5.41	**5.82**	6.02
4	5.44	5.86	**5.99**
8	5.45	5.87	6.03

Table 6: *RMS errors of log F0 [cent] between original and synthetic speech for each target speaker (open target speaker).*

# of speaker class	target speaker #1	#2	#3
1	184.9	217.3	205.6
2	174.9	211.4	**190.2**
4	**169.2**	**202.7**	193.3
8	173.8	208.9	193.2

the number of speaker classes. This indicates that the proposed technique enhances the effectiveness of the model's tree structure for log F0 and phoneme duration. On the other hand, the mel-cepstral distortions were not directly impacted by speaker class. This is because the speaker class yielded by average mel-cepstral coefficients does not adequately represent spectrum-oriented speaker characteristics. Therefore, to suppress mel-cepstral distortion, we have to determine which feature can best represent the spectrum-based characteristics of the speaker.

Table 5–7 also show, respectively, the mel-cepstral distortion, the RMS errors of log F0, and the RMS errors of phoneme duration for each open target speaker and each average voice model. These results demonstrate tendencies similar to those from the closed speakers test. However, when the number of speaker class is 8, RMS errors of most target speakers were higher to those with 4 classes. This is considered to be due to over-training. Therefore, we used the model with 4 classes in the following subjective experiment.

3.4. Subjective evaluation

We conducted a XAB test to evaluate voice characteristics and prosodic features of the synthesized speech using the model adapted from the conventional average voice model and the proposed average voice model. All permutations of synthetic sentence pairs matching each target speaker were created and presented in both orders (XAB and XBA), to eliminate bias in the order of stimuli. The subjects were ten persons, and each was presented synthesized speech samples and then asked which sample was similar to the reference speech. The reference speech was synthesized by a STRAIGHT vocoder. As in the objective evaluation of the open speakers, we used three open speakers as the target speaker, and twenty sentences as the evaluation sentences.

Figure 2 shows the preference scores for each target speaker. We can see that the proposed technique has better performance than the conventional technique. This indicates that the proposed technique based on speaker class can synthesize speech that is closer to the target speaker than the conventional alternative even though only three simple acoustic features are used for speaker clustering. However, since no performance comparison that changed the acoustic feature used for the speaker clustering was performed, it is necessary to evaluate the performance of speech synthesis by using other acoustic features that have high correlation with perceptual similarity as shown in [4].

Table 7: *RMS errors of phoneme duration [ms] between original and synthetic speech for each target speaker (open target speaker).*

# of speaker class	target speaker #1	#2	#3
1	22.11	19.78	21.60
2	23.78	19.71	22.95
4	**21.87**	**18.72**	**19.59**
8	21.92	19.06	20.14

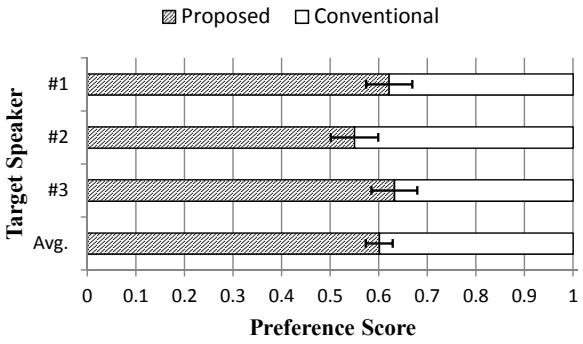

Figure 2: *Preference score from the subjective evaluation. (Error bars show the 95% confidence intervals.)*

4. Conclusion

In this paper, we proposed a model training technique that utilizes speaker class. This technique realizes robust speaker class estimation by using three simple features, the average mel-cepstral coefficients, average log F0, and speaking rate. Objective and subjective experiments showed that the proposed technique can synthesize speech that is closer to that of the target speaker than the conventional method. In particular, this technique can significantly reduce the RMS errors of log F0.

In future work, we will investigate other acoustic features and other speaker clustering techniques to improve the technique's speech synthesis performance. Applying the proposed technique to style adaptation [12] will also be investigated.

5. References

[1] H. Zen, K. Tokuda, T. Masuko, T. Kobayashi and T. Kitamura, "A Hidden semi-Markov model-based speech synthesis system," IEICE Trans. Inf. and Syst., vol.E90-D, no.5, pp.825–834, May. 2007.

[2] J. Yamagishi and T. Kobayashi, "Average-voice-based speech synthesis using HSMM-based speaker adaptation and adaptive training," IEICE Trans. Inf. and Syst. vol.E90-D, no.2, pp.533–543, Feb. 2007.

[3] J. Yamagishi, O. Watts, S. King and B. Usabaev, "Roles of the Average Voice in Speaker-adaptive HMM-based Speech Synthesis," in Proc. Interspeech 2010, pp.418–421, Sept. 2010.

[4] R. Dall, MC. Veaux, J. Yamagishi and S. King, "Analysis of speaker clustering strategies for HMM-based speech synthesis," in Proc. INTERSPEECH 2012, Sept. 2012.

[5] J. Yamagishi, K. Onishi, T. Masuko and T. Kobayashi, "Acoustic modeling of speaking styles and emotional expressions in HMM-based speech synthesis," IEICE Trans. Inf. & Syst., E88-D(3), pp.502–509, 2005.

[6] J. Yamagishi, T. Nose, H. Zen, Z. Ling, T. Toda, K. Tokuda, S. King, and S. Renals. "Robust speaker-adaptive HMM-based text-to-speech synthesis," IEEE Trans. Audio, Speech & Language Process., 17(6), pp.1208–1230, 2009.

[7] N. Higuchi and M. Hashimoto, "Analysis of acoustic features affecting speaker identification," in Proc. Eurospeech '95, pp.435–438, 1995.

[8] K. Amino, T. Sugawara and T. Arai, "Speaker Similarity in Human Perception and their Spectral Properties," in Proc. WESPAC IX, 2006.

[9] Y. Adachi, S. Kawamoto, S. Morishima and S. Nakamura, "Perceptual similarity measurement of speech by combination of acoustic features," in Proc. ICASSP 2008, pp.4861–4864, 2008.

[10] H. Kawahara, I. Masuda-Katsuse and A. Cheveigne, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," Speech Communication, 27, pp.187–207, 1999.

[11] J. Yamagishi, T. Kobayashi, Y. Nakano, K. Ogata and J. Isogai. "Analysis of Speaker Adaptation Algorithms for HMM-based Speech Synthesis and a Constrained SMAPLR Adaptation Algorithm," IEEE Trans. Audio, Speech & Language Process., 17(1), pp.66–83, 2009.

[12] M. Tachibana, J. Yamagishi, T. Masuko and T. Kobayashi, "A style adaptation technique for speech synthesis using HSMM and suprasegmental features," IEICE Trans. Inf. and Syst. vol. E89-D, no. 3, pp.1092–1099, Mar. 2006.