論文 / 著書情報
Article / Book Information

題目(和文)	
Title(English)	A Study on Acoustic Modeling of Speech for Personalized Speech Interface
著者(和文)	井島勇祐
Author(English)	Yusuke Ijima
出典(和文)	学位:博士（工学）, 学位授与機関:東京工業大学, 報告番号:甲第9881号, 授与年月日:2015年3月26日, 学位の種別:課程博士, 審査員:小林　隆夫,伊東　利哉,小池　康晴,杉野　暢彦,篠崎　隆宏
Citation(English)	Degree:, Conferring organization: Tokyo Institute of Technology, Report number:甲第9881号, Conferred date:2015/3/26, Degree Type:Course doctor, Examiner:,,,,
学位種別(和文)	博士論文
Category(English)	Doctoral Thesis
種別(和文)	要約
Type(English)	Outline

A Study on Acoustic Modeling of Speech for Personalized Speech Interface

Yusuke Ijima

March 2015

Summary

This thesis presents novel approaches to acoustic modeling of speech to achieve a personalized speech interface: automatic speech recognition (ASR) and text-to-speech synthesis (TTS).

Chapter 1 describes general background of the thesis.

Chapter 2 describes a rapid model adaptation technique for emotional speech recognition. This technique utilizes the MRHMM framework for the model adaptation and the style vector which corresponds to the degree or intensity of expressivity of styles. In the recognition stage, the HMM is adapted to the input style using the estimated style vector. It has been confirmed that the proposed technique reduces the error rate by 11% of the style-independent HMM. Furthermore, the technique can obtain not only linguistic information but also the degree of expressivity of emotional speech.

Chapter 3 presents an emotional speech recognition technique based on speaker adapted MRHMM. Using a speaker-independent neutral style model, the MRHMM is trained with a small amount of target speaker's data. The experimental results indicated that the performance of the proposed technique in both speech recognition and style estimation is promising for simulated emotional speech.

Chapter 4 describes an average voice model training technique that utilizes speaker classes representing the voice characteristics of speakers. In the speaker adaptation process, the speaker class of the target speaker is estimated and used for speaker adaptation and speech parameter generation. 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.

Chapter 5 presents a similar speaker selection technique as the first step

for further improving the similarity of the synthesized speech. The technique first trains a transform matrix based on distance metric learning using the perceptual voice quality similarity. Given an input speech, acoustic features of the input speech are transformed using the trained transform matrix. The results indicate that the proposed technique reduces the speaker selection error rate by about 53.9%.

Chapter 6 describes conclusions and future works of this thesis.

Acknowledgments

First, I would like to express my thanks to Professor Takao Kobayashi, Tokyo Institute of Technology, for all of his support, encouragement, and guidance. I would also like to thank Professors Toshiya Itoh, Yasuharu Koike, Nobuhiko Sugino, and Takahiro Shinozaki of Tokyo Institute of Technology for their kind suggestions.

Over the years, I have benefited greatly from interaction with members of the Kobayashi Laboratory. There are too many people to mention all of them individually, but I must thank Dr. Makoto Tachibana (currently with Yamaha Corporation), and Dr. Takashi Nose (currently with Tohoku University).

I have also benefited greatly from interaction with members of NTT Media Intelligence Laboratories, Nippon Telegraph and Telephone Corporation. There are too many people to mention all of them individually, but I must thank Dr. Satoshi Takahashi, Dr. Sumitaka Sakauchi, Dr. Hideyuki Mizuno, Noboru Miyazaki, and Mitsuaki Isogai (currently with NTT Communications Corporation). Without their help, I could not possibly have completed this work. I would also like to thank Professor Naomitsu Ikeda of Kumamoto National College of Technology. Without his help, I could not possibly have started working in the field of speech information processing.

Finally, I would like to give my special thanks to my family for all their support over the years.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 General background

Speech is one of the most important ways in which humans communicate. Over the years, many efforts have been made to incorporate speech into human-computer communication systems. In recent years, improvements in computer performance have produced greatly improved performance in speech interface systems, i.e., automatic speech recognition (ASR) and text-to-speech synthesis (TTS) systems. Consequently, a number of applications using speech interfaces have become indispensable for people's daily lives, such as car navigation systems, computer telephony integration systems (e.g., Interactive Voice Response; IVR), and transcription systems for use in parliament [1]. Furthermore, many new services using speech interfaces have been developed in the past several years with the rapid spread of sophisticated mobile devices such as smart phones and tablet computers. Examples of widely used services of this type are those provided by web search systems using speech (such as Google Voice Search [2]), spoken dialog systems (such as Apple Siri [3]), and automatic speech translation systems (such as NICT VoiceTra [4]).

In mobile device services, the speech interface plays a rather different role than in conventional applications such as IVR systems and transcription systems. The conventional applications require only the highly accurate transcriptions obtained from ASR systems. However, the new applications

also require paralinguistic and/or nonlinguistic information about speakers (emotional expressions, ages, illnesses, etc.) for comparatively new processes such as dialog control and machine translation. Furthermore, while the conventional applications do not need to handle various types of synthesized speech, the new applications need to handle types such as desirable speaker characteristics and emotional expressions. However, since these differ from user to user it would be difficult to handle them by using the same speech interface for all users. One realistic approach to overcoming this problem would be to enable mobile phone users to personalize their speech interface.

In the field of ASR, acoustic features of speech are affected by speaking styles and emotions as well as speaker characteristics. As a result, recognition performance is seriously degraded when emotional or spontaneous speech utterances are recognized. One useful approach to alleviating this problem is to adapt the acoustic models. Since the degree of expressivity of speaking styles and emotional expressions changes with every phrase, it is desirable to perform the model adaptation online. This implies that difficulties will be encountered when adapting an acoustic model while using only a limited amount of data; more specifically, one sentence or one phrase. For making such adaptations, rapid adaptation techniques based on a small number of control parameters would be more promising than the one based on maximum likelihood linear regression (MLLR) [5] because MLLR generally requires a certain amount of adaptation data to attain good performance. Such low dimensional parameter space-based adaptation techniques include vocal tract length normalization (VTLN) [6], Eigenvoice [7], and multiple-regression HMM (MRHMM) [8].

In the field of TTS, there are two main streams of corpus-based speech synthesis: the unit selection approach [9] and the HMM-based approach [10]. Unit selection speech synthesis enables speech to be synthesized with a high degree of naturalness since the synthesized speech is generated by directly using target speech waveforms of the target speaker. However, a considerable amount of speech data uttered by the target speaker is required to obtain sufficient performance. Therefore, the HMM-based approach is preferable since HMM-based speech synthesis enables an arbitrary speaker's speech to be synthesized from a smaller amount of the speaker's data than that required

in the unit selection approach. In the HMM-based approach, an average-voice-based speech synthesis technique with model adaptation [11] has been proposed. Given only a few minutes of the target speaker's speech data, this technique can synthesize arbitrary texts by adapting an average voice model to the target speaker's model. However, it has been reported that the model adaptation degrades the similarity of the synthesized speech to the target speaker's speech if the acoustic features of the average voice model are distant from those of the target speaker [12].

1.2 Scope of thesis

This thesis presents novel approaches to two methods (ASR and TTS) we proposed for acoustic modeling of speech to achieve a personalized speech interface. It will first describe a novel emotional speech recognition technique we proposed that can obtain not only linguistic information but also emotional expressions. It will also present a technique we proposed for generating synthesized speech of arbitrary speakers using a small amount of speech data.

Chapter 2 presents the novel adaptation technique we proposed, which is based on a quite small number of control parameters for enabling recognition of emotional and styled speech. The technique utilizes the MRHMM framework for model adaptation, but the approach to modeling speech is fundamentally different from that of [8]. In the original MRHMM, an additional acoustic feature, that is, fundamental frequency (F0), is used as the explanatory variable of the regression. In contrast, the proposed technique uses the degree or intensity of expressivity of emotional and styled speech, which is called the style vector, as the explanatory variable rather than specific acoustic features. The key idea of the technique is based on the style estimation [13] and style control [14] of speech. We first estimate the value of a style vector that represents the degree of expressivity of emotion or style of input speech. Then the model adaptation is conducted by setting the value of the explanatory variable to the estimated style vector and calculating the new mean vectors of output distribution functions of HMMs. As a result, we can obtain paralinguistic information, that is, the degree of expressivity of emotional and styled speech, as well as linguistic information after the recog-

nition process. The chapter will also show how the proposed technique ' s effectiveness was verified from results of phoneme recognition experiments for acted emotional speech.

Chapter 3 describes the MRHMM-based emotional speech recognition technique we also proposed, which uses only a small amount of training data of the target speaker because MRHMM training requires a considerable amount of training data for each style uttered by the target speaker. For the MRHMM training, we use a speaker-independent (SI) neutral style model that can be obtained much more easily than speaker-dependent style models. The SI model is adapted to target speakers' style-dependent models through simultaneous adaptation of speaker and style with a small amount of speech data uttered by the target speaker. Then, the MRHMM of the target speaker is trained from the obtained style-dependent models. We will show the assessment of the technique's performance in recognizing simulated emotional speech uttered by both professional narrators and non-professional speakers.

Chapter 4 presents the average voice model training technique we proposed, which is based on a speaker clustering approach, to generate synthetic speech with enhanced similarity to the target speakers' speech. A novel point of the technique is the use of the speaker characteristics (called "speaker class"), obtained from unsupervised clustering, as the additional contextual factor for average-voice-based speech synthesis. In the model training process, speaker clustering is first performed to all speakers used for model training to obtain the speaker class for each of them. The average voice model with multiple speaker characteristics is trained by using the obtained speaker classes. For the speaker adaptation and speech parameter generation, the speaker class of the target speaker is estimated using the Euclidean distance between the centroids of each cluster and the target speaker's feature. The use of the estimated speaker class makes it possible to utilize model parameters that have similar speaker characteristics to those of the target speaker for speaker adaptation and speech parameter generation. The results of objective and subjective experiments indicated the proposed technique can synthesize speech with improved similarity and naturalness.

Chapter 5 describes the speaker selection technique based on distance

metric learning (DML) [15] we proposed as the first step to achieving improved synthesized speech quality. It also describes experiments we conducted to ascertain its performance. We first conducted a large-scale subjective experiment using 62 female speakers to identify perceptual voice quality similarity. To exclude the influence of prosody, we used speech modified so as to exhibit exactly the same prosody (F0 and phoneme duration). Several acoustic features highly correlated to perceptual voice quality similarity were found by regression analysis of the subjective experiment results. With the proposed technique, the transform matrix is first trained on a data manipulation language (DML) basis to convert the acoustic feature space to perceptual voice quality similarity space. The acoustic features of a given speech sample are transformed using a trained transform matrix. Then, a similar speaker is chosen using Euclidean distances on the transformed acoustic feature space. An experiment was also performed to compare the technique ' s performance to that of speaker selection on an acoustic feature space without transformation.

Chapter 2

Emotional Speech Recognition Based on Speaker Dependent Multiple-Regression HMM

This chapter describes a model adaptation technique for emotional speech recognition based on multiple-regression HMM (MRHMM) to achieve the personalized ASR. We use a low-dimensional vector called style vector which corresponds the degree of expressivity of emotional speech as the explanatory variable of the regression. In the proposed technique, first, the value of the style vector for input speech is estimated. Then, using the estimated style vector, new mean vectors of the output distributions of HMM are adapted to the input style. The style vector is estimated every input utterance, and an on-line adaptation can be done in each utterance. We perform phoneme recognition experiments for professional narrators' acted speech and evaluate the performance by comparing with style-dependent and style-independent HMMs. Experimental results show the proposed technique reduced the error rates by 11% of the style-independent model.

A part of this chapter was presented at INTERSPEECH 2008 [16].

2.1 Introduction

To bring human-computer or -robot interaction more natural and realistic, we need an ASR system that can accept speech with various speaking styles and emotional expressions. In order to achieve such an ASR system, personalization of the ASR system would be required because emotional expressions and speaking styles are different with each speaker. In addition, acoustic features of speech are affected by speaking styles and emotions as well as speaker characteristics and linguistic factors. This fact causes serious deterioration of the performance on recognition of emotional or spontaneous speech. One of useful approaches to alleviating such a problem is to adapt the acoustic model. Since the degree of expressivity of speaking styles and emotional expressions would change in every utterance or even in a phrase, it is desirable to perform the model adaptation on line. This implies that we encounter a difficulty of adapting the acoustic model with using only a limited amount of data, more specifically, one sentence or one phrase.

For this purpose, rapid model adaptation techniques based on a small number of control parameters would be promising than that based on maximum likelihood linear regression (MLLR) [5], because MLLR generally requires a certain amount of adaptation data to attain considerable performance. Such low dimensional parameter space-based adaptation techniques include vocal tract length normalization (VTLN) [17], Eigenvoice [7], and multiple-regression HMM (MRHMM) [8]. In this chapter, we propose a novel adaptation technique based on a quite small number of control parameters for emotional speech recognition.

The proposed technique utilizes the MRHMM framework for the model adaptation. However, the approach to the modeling of speech is fundamentally different from that of [8]. In the original MRHMM, an additional acoustic feature, that is, fundamental frequency (F0), is used as the explanatory variable of the regression. In contrast, the proposed technique uses the degree or intensity of expressivity of emotions and styles appeared in acoustic features of speech, which is called the style vector, as the explanatory variable rather than specific acoustic features. The key idea of the technique is based on the style estimation [13] and style control [14] techniques of speech.

We first estimate the value of the style vector, then conduct the model adaptation by setting the value of the explanatory variable to the estimated style vector and calculating new mean vectors of the output distribution functions of HMM. As a result, we can obtain para-linguistic information, that is, the degree of expressivity of emotional speech as well as linguistic information after recognition process. In the eigenvoice technique [7], since each axis of the eigenspace does not represent the degree of expressivity of emotional speech, it would not be easy to obtain such para-linguistic information directly. We show the effectiveness of the proposed technique from results of phoneme recognition experiments for acted emotional speech.

2.2 Multiple-regression HMM

2.2.1 Style modeling based on MRHMM

Let $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$ be the mean vector and covariance matrix of output distribution of HMM at state i. In this chapter, we assume that the mean vector of MRHMM is modeled using multiple regression as

$$\boldsymbol{\mu}_i = \boldsymbol{h}_0^{(i)} + \boldsymbol{A}_i \boldsymbol{v} = \boldsymbol{H}_i \boldsymbol{\xi} \tag{2.1}$$

where $\boldsymbol{H}_i = [\boldsymbol{h}_0^{(i)}, \cdots, \boldsymbol{h}_L^{(i)}]$, $\boldsymbol{A}_i = [\boldsymbol{h}_1^{(i)}, \cdots, \boldsymbol{h}_L^{(i)}]$, $\boldsymbol{\xi} = [1, \boldsymbol{v}^\top]^\top$, and $\boldsymbol{v} = [v_1, \cdots, v_L]^\top$ is a style vector. The component v_k of the style vector represents the degree of expressivity of a certain emotional expressions or speaking style in speech. In addition, $\boldsymbol{H}_i$ is the regression matrix of dimension $M \times (L+1)$ and M is the dimensionality of $\boldsymbol{\mu}_i$.

When the training data and corresponding style vectors are given, the parameters of MRHMM, i.e., $\boldsymbol{H}_i$ and $\boldsymbol{\Sigma}_i$ can be estimated using EM algorithm. These estimation formulas can be found in [14].

2.2.2 Style estimation

We consider a problem of estimating the style vector $\boldsymbol{v}$ for an input observation sequence $\boldsymbol{O} = (\boldsymbol{o}_1, \cdots, \boldsymbol{o}_T)$ given the trained MRHMM λ whose parameters $\boldsymbol{H}_i$ and $\boldsymbol{\Sigma}_i$ are fixed. The optimal style vector $\overline{\boldsymbol{v}}$ for the input

observation $\boldsymbol{O}$ is determined in ML sense as

$$\overline{\boldsymbol{v}} = \arg \max_{\boldsymbol{v}} P(\boldsymbol{O}|\lambda, \boldsymbol{v}). \tag{2.2}$$

The EM algorithm-based re-estimation formula of the style vector for output distribution is given by

$$\overline{\boldsymbol{v}} = \left(\sum_{i=1}^{N} \sum_{t=1}^{T} \gamma_t(i) \boldsymbol{A}_i^\top \boldsymbol{\Sigma}_i^{-1} \boldsymbol{A}_i \right)^{-1} \left(\sum_{i=1}^{N} \sum_{t=1}^{T} \gamma_t(i) \boldsymbol{A}_i^\top \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{o}_t - \boldsymbol{h}_0^{(i)}) \right) \tag{2.3}$$

where N is the number of states, and $\gamma_t(i)$ is the probability of being in state i at time t. The derivation of the equation can be found in [13]. Note that the estimation formula in [13] is derived in hidden semi-Markov model (HSMM) framework which has explicit state duration pdfs.

In this study, we assume that the input observation sequence $\boldsymbol{O}$ is a whole sentence, and estimate the style vector in each sentence.

2.3 Speech recognition using MRHMM

2.3.1 MRHMM training

A block diagram of the training part for the MRHMM is shown in Fig. 2.1. We first train triphone HMMs for respective styles, such as neutral, sad, and joyful styles, independently. Then we apply a shared decision tree context clustering (STC) technique [18] to these models to construct a common tree structure for all styles. After that, we further apply re-estimation process based on the EM algorithm to the resultant triphone HMM of each style. Finally, we obtain a single model with the common tree structure for all styles by incorporating the style vector into the re-estimation procedure based on the EM algorithm for MRHMM.

2.3.2 Overview of the recognition system

When the trained MRHMM and a specific style vector are given, an HMM having the new mean vectors calculated by Eq.(2.1) can be obtained. By

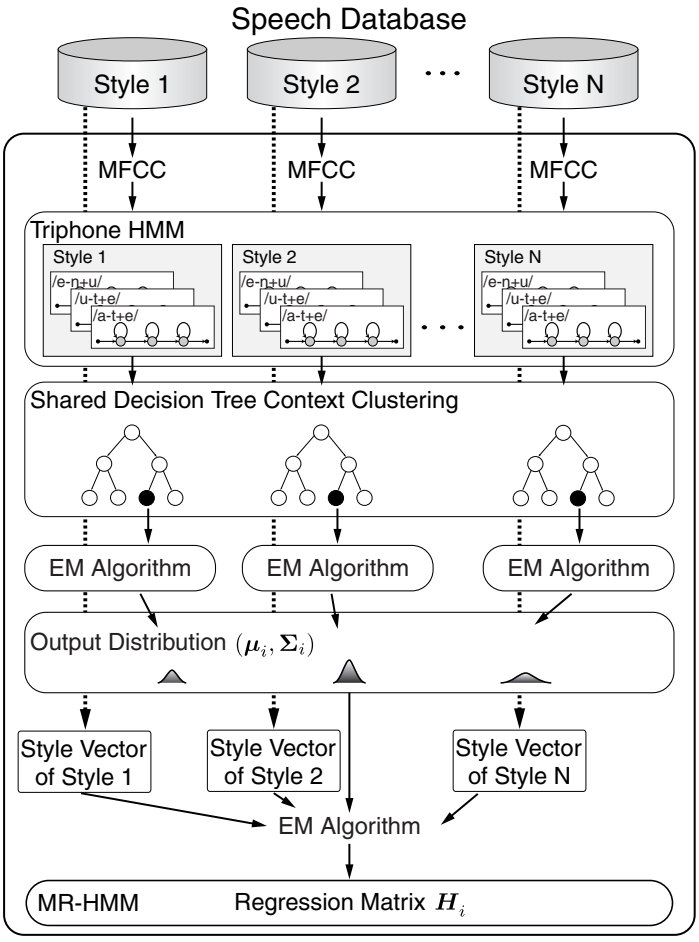


Figure 2.1: A block diagram of MRHMM training.

using the obtained HMM, we can straightforwardly perform ordinary speech recognition based on Viterbi algorithm. In the proposed technique, first, the style vector is estimated using style estimation technique mentioned in Section **2.2.2**, then, using the estimated style vector, the adapted HMM for recognition is obtained from the MRHMM. The style vector is estimated every input utterance, and the adapted HMM is modified in each utterance. This recognition process can be viewed as a kind of an on-line adaptation. When we perform the style estimation, we need a label sequence of the input speech. In [13], the triphone label is obtained by a pre-trained style-

independent HMM. In this study, for the given MRHMM, we use a two-pass recognition process summarized as follows:

Step 1 Obtain the *initial* HMM by setting the style vector equal to **0** in MRHMM.

Step 2 Perform phoneme recognition using the initial HMM.

Step 3 Estimate the style vector $\overline{\boldsymbol{v}}$ for input speech using phoneme label of the input speech obtained in **Step 2**.

Step 4 Obtain the *adapted* HMM from MRHMM by calculating the new mean vectors with the estimated style vector $\overline{\boldsymbol{v}}$.

Step 5 Perform phoneme recognition again using the adapted HMM and obtain the recognition result.

2.4 Experiments

2.4.1 Experimental conditions

We used three styles of professional narrators' acted speech – neutral, sad, and joyful styles. Speech database [19] of each style contains 503 phonetically balanced ATR Japanese sentences (about 50 minutes) uttered by two male and one female professional narrators, MMI, MJI, and FTY, respectively. The neutral, sad, and joyful style speech data were not real emotional speech data, but just read speech data with simulated styles. When we recorded the speech data, we directed the speakers not to vary so much the degree of expressivity in each style.

450 sentences were used for the training of MRHMM and fifty sentences not included in the training data were used as the evaluation data in each style. We performed a 10-fold cross-validation test. A one-dimensional style space was used and the style vectors of training data were set as (-1.0), (0.0), and (1.0) for the sad, neutral, and joyful styles, respectively. Although dimensional models such as the circumplex model [20] are widely used as the space representing emotional expressions, we did not use such style spaces to avoid the multicollinearity.

Table 2.1: Experimental conditions.

Sampling frequency	16 kHz
Frame length	25 ms
Frame shift	10 ms
Analysis window	Hamming window
Feature vector	12 MFCCs $+\Delta$
	Log of power $+\Delta$
Number of monophones	42
Model	left-to-right, 1mix., 3-state
	triphone HMM

To compare the recognition performance between MRHMM and HMMs, we also trained style-dependent HMMs (SD) and style-independent HMM (SI). The SD models were trained from 450 sentences of each style and the SI model was a single model trained by 1350 sentences combining 450 sentences of the three styles. In general, the tying topology of HMM affects the recognition performance. To compare the performance of the models under the same condition, we chose the topology of the HMMs and the MRHMM to be the same tree structure obtained by STC described in Section **2.3.1**. Other experimental conditions are shown in Table 2.1. Although the number of mixtures is generally set to 2 or more, we set it to 1 since the amount of the training data was not enough to increase the number of mixtures.

2.4.2 Recognition using neutral style HMM

To examine the influence of emotional speech on the recognition rate, we first performed phoneme recognition using the neutral style-dependent HMMs. Table 2.2 shows the recognition error rate of each speaker and style. In the table, the entry for "Average" means the average result of the three speakers. The error rate was calculated by

$$error(\%) = \left(1 - \frac{H}{H + D + S}\right) \times 100 \tag{2.4}$$

Table 2.2: Phoneme error rates (%) for neutral style-dependent HMM.

Input Style	Speaker			
	MMI	FTY	MJI	Average
Neutral	5.47	7.32	4.45	5.75
Sad	5.89	12.36	21.87	13.28
Joyful	14.90	12.32	18.83	15.34

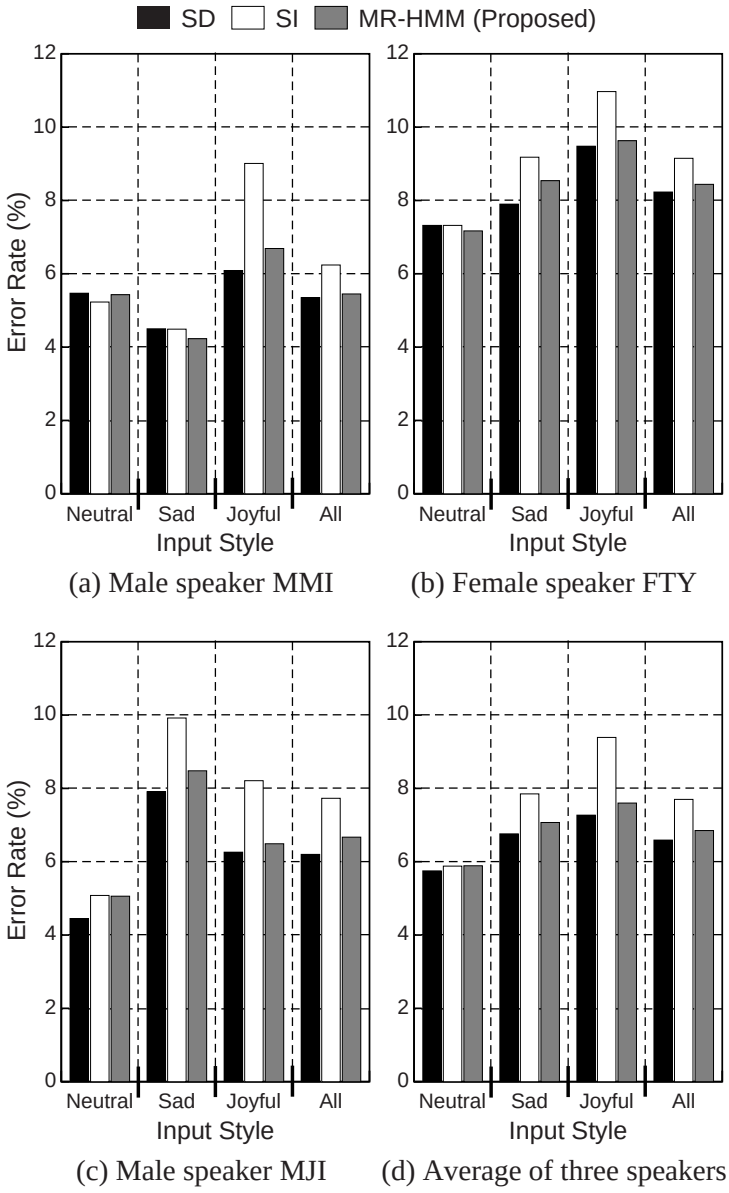
Table 2.3: Phoneme error rates (%) for initial and adapted HMMs.

Input Style	HMM	Speaker			
		MMI	FTY	MJI	Average
Sad	initial	4.79	9.61	10.86	8.42
	adapted	4.23	8.54	8.48	7.07
	correct label	4.20	8.51	8.38	7.03
Joyful	initial	9.44	11.19	8.74	9.79
	adapted	6.69	9.63	6.49	7.60
	correct label	6.56	9.49	6.38	7.48

where H, S, and D represent the number of correctly recognized phonemes, substitutions, and deletions, respectively. We can see that the error rates increased in sad and joyful styles compared to neutral style. From this result, we confirmed that the acoustic features were different in each style and that degrade the recognition performance.

2.4.3 Adaptation performance of MRHMM

Table 2.3 shows the recognition error rate of the initial HMM obtained in **Step 2**, and the adapted HMM obtained in **Step 4**. In the table, the entry for "correct label" represents the result when using the correct phoneme label of the input speech in the style estimation in **Step 3** and this corresponds to the upper limit of the performance of the proposed technique. It can be seen that the adapted HMM gives higher performance than the initial HMM. Moreover, the result of the adapted HMM is close to that of the "correct label" case.



(a) Male speaker MMI (b) Female speaker FTY

(c) Male speaker MJI (d) Average of three speakers

Figure 2.2: Phoneme recognition error rates (%).

2.4.4 Performance comparison between HMM and MRHMM

Figure 2.2 shows the recognition error rate of the style-dependent (SD) models, style-independent (SI) model, and MRHMM. In this figure, "All" represents the average result of the all three styles. As for the SD models, the

Table 2.4: Error reduction rates (%) from the SI model.

Input Style	Speaker			
	MMI	FTY	MJI	Average
Neutral	-3.82	2.05	0.39	-0.17
Sad	5.79	10.97	14.52	9.97
Joyful	25.75	12.22	20.95	19.06
All	12.66	7.76	13.71	11.04

style of the input speech was assumed to be known and the recognition result was obtained using the SD model of the input style. Thus, it is an ideal case and, in actual use, the style of input speech should be identified by some classification techniques. We can see that the error rates of the proposed technique decreased and were less than or comparable to that of the SD models. The error rates of SI model were reduced compared to that of neutral style-dependent model in Table 2.2, however, it is still higher than SD models and MRHMM.

Table 2.4 shows the error reduction rates of the proposed technique from the SI model. It can be seen that the MRHMM reduced the error rate by 11.04% on the average. Especially, the performance is improved in the joyful and sad styles.

2.4.5 Style estimation using MRHMM

Figures 2.3–2.5 show the distributions of the estimated value of the style vector of the test speech data. It can be seen that each style gives a different distribution and each distribution is near the value of the style vector that were set in the training. When we chose the classification threshold as from -1.5 to -0.5 for sad, from -0.5 to 0.5 for neutral, and from 0.5 to 1.5 for joyful, about 96% of speech data were classified as the correct style class of the input speech. However, there is a slight displacement between the mode of each distribution and the value of the style vector assumed in the training. This is because the acoustic features of the sad and joyful style included in

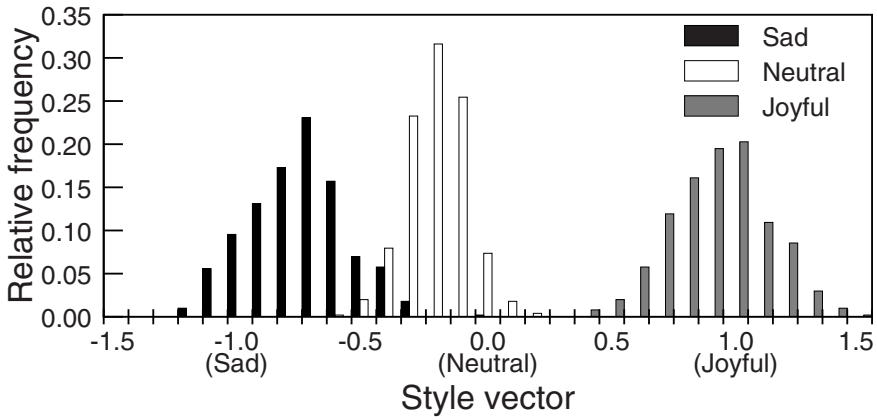


Figure 2.3: Histogram of the estimated value of the style vector (Male speaker MMI).

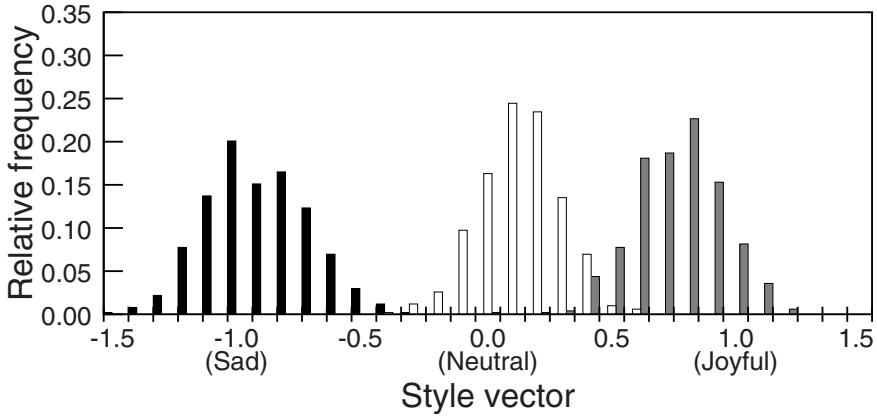


Figure 2.4: Histogram of the estimated value of the style vector (Female speaker FTY).

the database are not completely symmetric and that of neutral style are not absolutely located mid-point between the sad and joyful styles. As a result, the three styles were influenced by each other in the MRHMM training. Two-dimensional style space in which the sad style and joyful style are located in the independent axis might be one of the approaches to overcoming the problem.

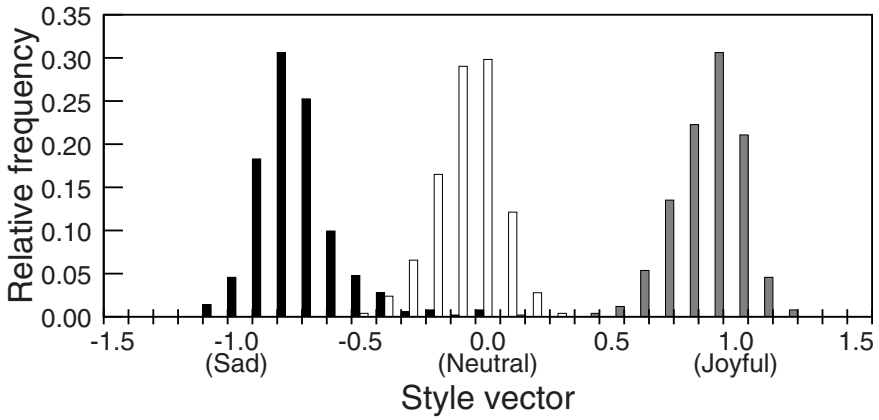


Figure 2.5: Histogram of the estimated value of the style vector (Male speaker MJI).

2.5 Conclusions

In this chapter, we have presented a speech recognition technique considering the degree of expressivity of speaking styles or emotional expressions. This technique utilizes the multiple-regression HMM (MRHMM) framework for the model adaptation and the style vector which corresponds to the degree or intensity of expressivity of styles as the explanatory variable of the regression. In the recognition stage, we adapt the HMM to the input style using the estimated style vector. We have shown that the proposed technique reduced the error rates by 11% of the style-independent HMM. Furthermore, we can obtain not only linguistic information but also the degree of expressivity of emotional and styled speech from the recognition process. This para-linguistic information would be also useful for human-computer or -robot interaction to detect user's emotional state and respond to the user more natural and appropriately.

Chapter 3

Emotional Speech Recognition Based on Speaker Adapted Multiple-Regression HMM

In this chapter, we propose an emotional speech recognition technique based on multiple-regression HMM (MRHMM) with a small amount of speech data uttered by the arbitrary speaker. In the previous chapter, MRHMM-based emotional speech recognition technique was proposed. However, this technique requires a large amount of speech data uttered by the target speaker to train MRHMM of the target speaker. In order to avoid this problem, speaker adaptation technique is applied to MRHMM-based emotional speech recognition in this chapter. For the MRHMM training, we use a speaker-independent (SI) neutral style model which can be obtained much easier than speaker-dependent style models. The SI model is adapted to target speaker's style-dependent models based on simultaneous adaptation of speaker and style with a small amount of speech data uttered by the target speaker. Then, the MRHMM of the target speaker is trained from the obtained style-dependent models. The recognition process consists of two stages. In the first stage, the style vector that represents the emotional expression category and the intensity of its expressiveness for the input speech is estimated on a sentence-by-sentence basis. Next, the acoustic models are adapted using the estimated

A part of this chapter was presented in IEICE Trans. Inf. and Syst. [21].

style vector, and then standard HMM-based speech recognition is performed in the second stage. We assess the performance of the proposed technique in the recognition of simulated emotional speech uttered by both professional narrators and non-professional speakers.

3.1 Introduction

In the previous chapter, we have presented a speech recognition technique considering the degree of expressivity of speaking styles or emotional expressions. This technique utilizes the multiple-regression HMM (MRHMM) framework for the model adaptation and the style vector which corresponds to the degree or intensity of expressivity of styles as the explanatory variable of the regression. As results of the evaluations, the proposed technique reduced the error rates by the style-independent HMM. Furthermore, we can obtain not only linguistic information but also the degree of expressivity of emotional and styled speech from the recognition process. However the technique has a problem that a considerable amount of speech data of the target speaker is required in advance to train the MRHMM. This leads to difficulty in recognition of arbitrary speaker's emotional speech. Although a possible approach to this problem is to use a speaker-independent MRHMM, the performance would be unsatisfactory because the emotional or style expressiveness varies sensitively on individual characteristics.

In this chapter, we propose a technique that enables us to easily obtain an arbitrary speaker's model and to adapt the model on-line. The on-line adaptation process of the proposed technique is the same as the MRHMM-based rapid model adaptation. However, for the MRHMM training, we use a speaker-independent (SI) neutral style model which can be obtained much easier than speaker-dependent style models. The SI model is adapted to target speaker's style-dependent models based on simultaneous adaptation of speaker and style with a small amount of speech data uttered by the target speaker. Then, the MRHMM of the target speaker is trained from the obtained style-dependent models. In the recognition stage, we first estimate the value of style vector for every sentence of the input speech based on a style estimation technique [13]. Then we adapt the model by calculating

new mean vectors of the probability density functions and perform standard HMM-based speech recognition.

In this chapter, we examine the effectiveness of the proposed technique under a condition where the types of emotion are limited, and also the amount of training data of the target speaker is very small.

3.2 Speech recognition based on multiple-regression HMM

3.2.1 Acoustic modeling of speech with multiple styles using MRHMM considering multiple mixture

In the MRHMM-based emotional speech recognition framework described in chapter 2, the acoustic model is represented by MRHMM, i.e., HMM with Gaussian probability density functions (pdfs) in which the mean vector of each pdf is expressed by a function of a low-dimensional vector called the style vector. Each component of the style vector corresponds to an intensity or quantity that represents how much the acoustic features are affected by a certain emotional expression or speaking style.

Here we consider a Gaussian mixture pdf as the output pdf. Let $\boldsymbol{\mu}_{im}$ be the mean vector of the m-th mixture component at state i. In the MRHMM, the mean vector is assumed to be represented by multiple regression of a style vector $\boldsymbol{v}$ as

$$\boldsymbol{\mu}_{im} = \boldsymbol{h}_0^{(im)} + \boldsymbol{A}_{im}\boldsymbol{v} = \boldsymbol{H}_{im}\boldsymbol{\xi} \tag{3.1}$$

where

$$\boldsymbol{A}_{im} = \left[\boldsymbol{h}_1^{(im)}, \cdots, \boldsymbol{h}_L^{(im)}\right] \tag{3.2}$$

$$\boldsymbol{H}_{im} = \left[\boldsymbol{h}_0^{(im)}, \cdots, \boldsymbol{h}_L^{(im)}\right] \tag{3.3}$$

$$\boldsymbol{v} = \left[v_1, v_2, \cdots, v_L\right]^\top \tag{3.4}$$

$$\boldsymbol{\xi} = \left[1, \boldsymbol{v}^\top\right]^\top. \tag{3.5}$$

$\boldsymbol{A}_{im}$ and $\boldsymbol{H}_{im}$ are $D\times L$- and $D\times(L+1)$-dimensional regression matrices, and D and L are the dimensionalities of $\boldsymbol{\mu}_{im}$ and $\boldsymbol{v}$, respectively. When training

data and corresponding style vectors are given, the regression matrix $\boldsymbol{H}_{im}$ of the MRHMM can be estimated using an EM algorithm. Let $\{\boldsymbol{O}^{(1)}, \cdots, \boldsymbol{O}^{(K)}\}$ and $\{\boldsymbol{v}^{(1)}, \cdots, \boldsymbol{v}^{(K)}\}$ be sets of observation sequences and style vectors for model training, where K is the total number of observation sequences, $\boldsymbol{O}^{(k)} = (\boldsymbol{o}_1^{(k)}, \cdots, \boldsymbol{o}_{T_k}^{(k)})$ is the k-th observation sequence, T_k is the number of frames of $\boldsymbol{O}^{(k)}$, and $\boldsymbol{v}^{(k)}$ is the style vector that corresponds to $\boldsymbol{O}^{(k)}$. The re-estimation formula of the regression matrix of the MRHMM can be derived in a similar way as that for the single mixture model case [14] based on a maximum likelihood (ML) criterion, and is given as follows.

$$\boldsymbol{H}_{im}^{ML} = \left(\sum_{k=1}^{K} \sum_{t=1}^{T_k} \sum_{m=1}^{M} \gamma_t(i,m) \boldsymbol{o}_t^{(k)} \boldsymbol{\xi}^{(k)\top} \right) \cdot \left(\sum_{k=1}^{K} \sum_{t=1}^{T_k} \sum_{m=1}^{M} \gamma_t(i,m) \boldsymbol{\xi}^{(k)} \boldsymbol{\xi}^{(k)\top} \right)^{-1} \tag{3.6}$$

where M is the number of mixtures of the MRHMM, $\boldsymbol{o}_t^{(k)}$ is an observation vector at time t in $\boldsymbol{O}^{(k)}$, and $\boldsymbol{\xi}^{(k)} = [1, \boldsymbol{v}^{(k)\top}]^\top$. In addition, $\gamma_t(i,k)$ is the probability of being in the m-th mixture component of state i at time t for given $\boldsymbol{O}^{(k)}$.

3.2.2 Style estimation for on-line model adaptation

We consider a problem of estimating the style vector $\boldsymbol{v}$ for an input observation sequence $\boldsymbol{O} = (\boldsymbol{o}_1, \cdots, \boldsymbol{o}_T)$ given the trained MRHMM λ whose parameters $\boldsymbol{H}_{im}$ and the covariance matrix $\boldsymbol{\Sigma}_{im}$ are fixed. The optimal style vector $\boldsymbol{v}^*$ for the input observation $\boldsymbol{O}$ is determined based on an ML criterion as

$$\boldsymbol{v}^* = \underset{\boldsymbol{v}}{\operatorname{argmax}} \, P\left(\boldsymbol{O}|\lambda, \boldsymbol{v}\right). \tag{3.7}$$

The EM algorithm-based re-estimation formula of the style vector for the output pdf is given by

$$\overline{\boldsymbol{v}} = \left(\sum_{i=1}^{N} \sum_{t=1}^{T} \sum_{m=1}^{M} \gamma_t(i,m) \boldsymbol{A}_{im}^{\top} \boldsymbol{\Sigma}_{im}^{-1} \boldsymbol{A}_{im} \right)^{-1} \left(\sum_{i=1}^{N} \sum_{t=1}^{T} \sum_{m=1}^{M} \gamma_t(i,m) \boldsymbol{A}_{im}^{\top} \boldsymbol{\Sigma}_{im}^{-1} (\boldsymbol{o}_t - \boldsymbol{h}_0^{(im)}) \right) \tag{3.8}$$

where N is the number of states of the MRHMM. The above formula is straightforwardly derived from the single mixture model case [13] where the estimation formula is derived within a hidden semi-Markov model (HSMM) framework, which is the model having explicit state-duration pdfs.

In this study, we assume that the input observation sequence $\boldsymbol{O}$ is a set of acoustic features for one sentence and we estimate the style vector in each sentence.

3.2.3 Training of MRHMM with a small amount of speech data using model adaptation

The MRHMM training generally requires a considerable amount of speech data to obtain reliable model parameters. However, it is unrealistic to prepare a sufficient amount of speech data of arbitrary speakers. In the style control and style estimation based on the multiple-regression HSMM (MRHSMM), we have shown that the use of average voice model [11] and simultaneous adaptation of speaker and style is promising for overcoming this problem [22], [23]. Thus we incorporate a similar approach into the MRHMM-based emotional speech recognition.

A block diagram of the model training is illustrated in Fig. 3.1. First, we train a speaker-independent (SI) neutral style model with a sufficient amount of neutral style speech of many speakers. Next, we adapt the SI neutral style model to a target speaker's respective styles using a model adaptation technique with a small amount of speech data uttered in advance by the target speaker. Then we obtain the target speaker's MRHMM based on a least squares estimation from the speaker- and style-adapted HMMs.

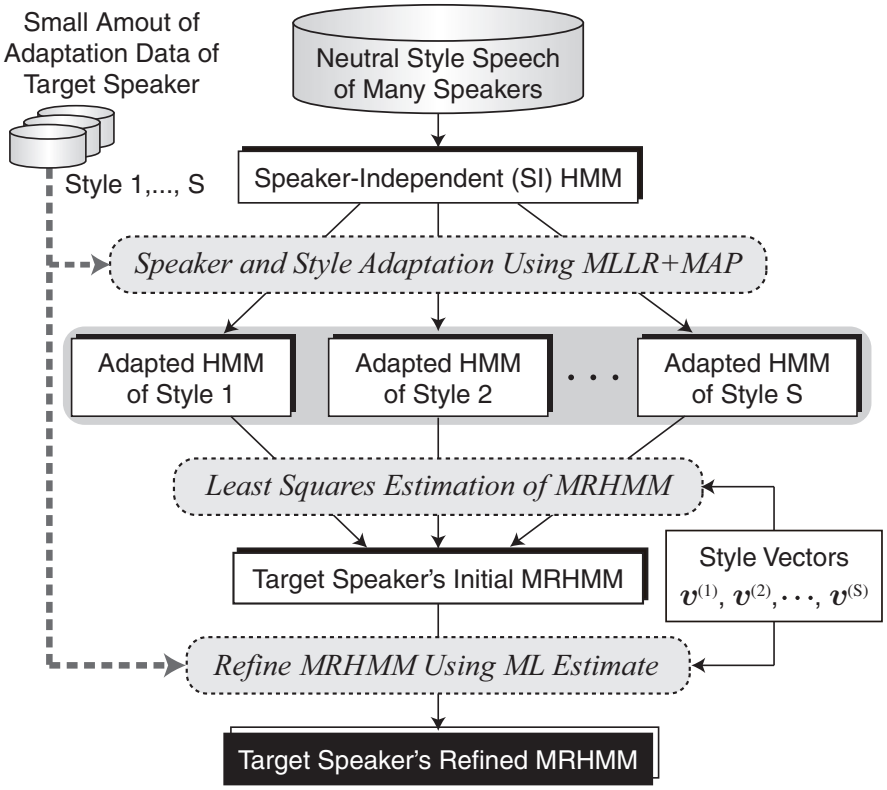


Figure 3.1: MRHMM training using SI neutral style model and model adaptation.

Suppose that the adaptation data contains speech uttered in S different styles. Let the mean vector of the m-th mixture pdf at state i of the style-adapted HMM of style s and the corresponding style vector be given by $\boldsymbol{\mu}_{im}^{(s)}$ and $\boldsymbol{v}^{(s)}$, respectively, for $1 \leq s \leq S$. We choose $\boldsymbol{H}_{im}$ that minimizes

$$E = \sum_{s=1}^{S} \left\| \boldsymbol{\mu}_{im}^{(s)} - \boldsymbol{H}_{im} \boldsymbol{\xi}^{(s)} \right\|^2 \tag{3.9}$$

as the regression matrix of the MRHMM [22], [23]. By differentiating E with respect to $\boldsymbol{H}_{im}$ and equating the result to zero, the optimal regression matrix $\boldsymbol{H}_{im}^{LS}$ is obtained as

$$\boldsymbol{H}_{im}^{LS} = \left(\sum_{s=1}^{S} \boldsymbol{\mu}_{im}^{(s)} \boldsymbol{\xi}^{(s)\top} \right) \left(\sum_{s=1}^{S} \boldsymbol{\xi}^{(s)} \boldsymbol{\xi}^{(s)\top} \right)^{-1}. \tag{3.10}$$

To improve the performance of the simultaneous adaptation of speaker and style using only a small amount of speech data, we refine the MRHMM parameter $\boldsymbol{H}_{im}$ as follows [22]:

$$\boldsymbol{H}_{im} = \frac{\tau\, \boldsymbol{H}_{im}^{LS} + \Gamma_{im}\, \boldsymbol{H}_{im}^{ML}}{\tau + \Gamma_{im}} \tag{3.11}$$

where $\boldsymbol{H}_{im}^{LS}$ is the regression matrix obtained by Eq. (3.10) and $\boldsymbol{H}_{im}^{ML}$ is the regression matrix estimated from the adaptation data in ML sense. In addition, τ is a positive parameter for controlling the modification weight and

$$\Gamma_{im} = \sum_{t=1}^{T} \gamma_t(i, m). \tag{3.12}$$

It is noted that the regression matrix $\boldsymbol{H}_{im}$ approaches to $\boldsymbol{H}_{im}^{ML}$ when enough adaptation data is available for the m-th mixture component of state i.

3.2.4 Emotional speech recognition using MRHMM-based on-line model adaptation

When the trained MRHMM and a specific style vector are given, an HMM having the new mean vectors calculated by Eq. (3.1) can be obtained. By using this HMM, we can straightforwardly perform ordinary speech recognition based on HMM.

In the proposed technique, first, the style vector is estimated using the style estimation technique mentioned in Sect. **3.2.2**, and then, using the estimated style vector, the adapted HMM for the recognition is obtained from the MRHMM. The style vector is estimated for every input utterance, and the adapted HMM is modified in each utterance. When we perform the style estimation, we need a phoneme label sequence of the input speech to calculate $\gamma_t(i, m)$ by the forward-backward algorithm. For this purpose, we use a two-pass recognition process. The overall recognition process is summarized as follows.

SI model training:

Step 0 Train SI neutral style HMM using neutral style speech data of many speakers.

MRHMM training:

Step 1 Convert the SI neutral style model into the target speaker's respective style models using a model adaptation technique.

Step 2 Construct the target speaker's MRHMM using Eq. (3.10).

Step 3 Refine the obtained MRHMM using Eq. (3.11).

MRHMM-based recognition:

Step 4 Obtain *neutral style* HMM by setting the style vector equal to $\mathbf{0}$, which is assumed to be the value of the style vector corresponding to the neutral style, in the training of the MRHMM.

Step 5 Perform phoneme recognition of input speech using the neutral style HMM.

Step 6 Estimate the style vector $\boldsymbol{v}^*$ for the input speech using the trained MRHMM and the phoneme sequence obtained in **Step 5**.

Step 7 Obtain *style-adapted* HMM from the trained MRHMM by calculating the new mean vectors with the estimated style vector $\boldsymbol{v}^*$ using Eq. (3.1).

Step 8 Perform speech recognition using the style-adapted HMM and obtain the final recognition result.

3.3 Experiments

3.3.1 Emotional speech database

In the following experiments, we used professional narrators' and non-professional speakers' speech. The professional narrators' speech data contains three styles of speech samples with simulated emotions — neutral, sad, and joyful styles, in which 503 phonetically balanced sentences taken from the ATR Japanese speech database were uttered by two males (MMI and MJI) and one female (FTY) narrators in the respective styles. The non-professional speakers' speech data consists of four styles of speech samples — neutral, sad, joyful, and angry styles, uttered with simulated emotions by

Table 3.1: Experimental conditions.

Sampling frequency	16 kHz
Frame length	25 ms
Frame shift	10 ms
Analysis window	Hamming window
Feature vector	12 MFCCs (with CMN) $+\Delta$
	Log of power $+\Delta$
Number of monophones	42
Model	left-to-right, 16-mixture, 3-state
	triphone HMM/MRHMM
	with diagonal covariance

nine graduate students (eight males and one female). Each style contains 100 sentences chosen from the above 503 sentences. The non-professional speakers had little experience in uttering the given sentence with such simulated styles. All the speech samples were recorded in a quiet room, and the speakers were directed to keep the degree of expressiveness of each style almost constant.

3.3.2 Experimental conditions

The SI neutral style model was trained from neutral style speech data of 209 speakers (106 males and 103 females) included in the Japanese Newspaper Article Sentences (JNAS) [24]. These speakers were different from the professional narrators and non-professional speakers mentioned above. The speech data used for the training of the SI neutral style model was about 50 sentences for each speaker, 10498 sentences in total. The parameters of the SI neutral style model were tied using a decision-tree-based context clustering with MDL criterion [25]. The total number of states in the SI neutral style model was 1875.

In the speaker and style adaptation, five sentences (around 20 seconds) of the respective styles were used for each target speaker. To alleviate the dependency of the choice of the adaptation data, the adaptation sentences

were randomly chosen and the experiments were conducted twice by changing the adaptation data. As for the model adaptation technique in **Step 1**, we applied a combined approach based on the MLLR and maximum a posteriori (MAP) adaptation (MLLR+MAP) [26]. Since the amount of adaptation data of each target style was small, we used a global transform in the MLLR. In this study, the covariance parameters were not adapted because the amount of adaptation data was very small. We set $\tau = 10$ in Eq. (3.11) on the basis of preliminary experimental results.

The speech recognition was performed based on the Viterbi algorithm using the decoder of the Hidden Markov Model Toolkit (HTK) [27]. We used phonetic networks based on Japanese phonetic concatenation rules in the recognition. The other experimental conditions are listed in Table 3.1. Although we have set the number of mixtures to 1 in chapter 2, we set it to 16 in the above experiments.

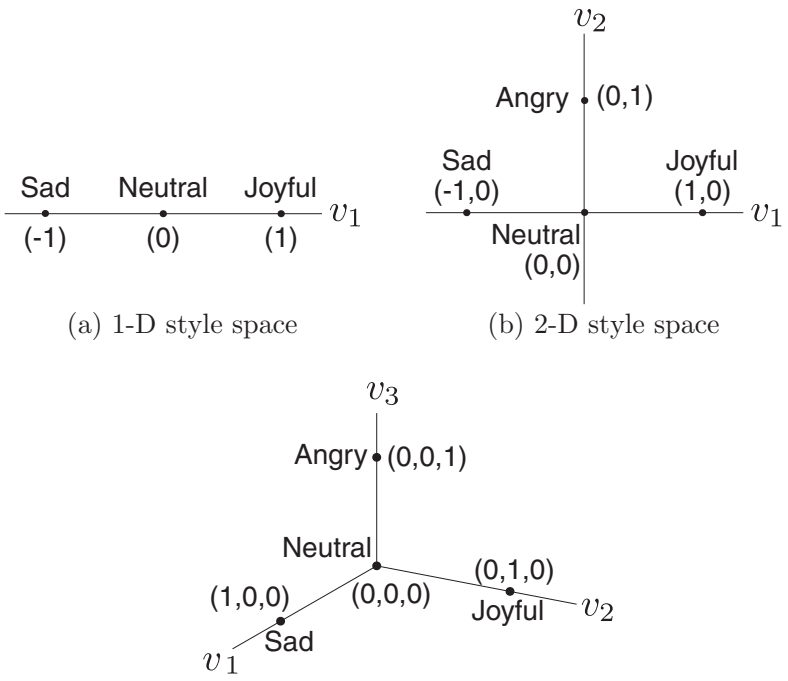
3.3.3 Performance evaluation with professional narrators

3.3.3.1 Performance of speaker and style adaptation of the MRHMM

We first evaluated the performance of the speaker and style adaptation by comparing the proposed MRHMM with four types of ordinary HMMs. In this experiment, we used three styles of the professional narrators' speech data. A one-dimensional style space (Fig. 3.2(a)) was used for the MRHMM. The style vectors for the adaptation data were set to fixed values, (-1), (0), and (1) for the sad, neutral, and joyful styles, respectively. We performed 10-fold cross-validation tests using 50 test sentences that were not included in the adaptation data.

Table 3.2 shows the average scores of the three speakers' phoneme recognition error rates, and the entry for "Overall" represents the average score of all the styles. The error rate was calculated by

$$error(\%) = \left(1 - \frac{H}{H + D + S}\right) \times 100 \tag{3.13}$$



(a) 1-D style space

(b) 2-D style space

(c) 3-D style space

Figure 3.2: Style spaces for MRHMM.

where H, S, and D represent the numbers of correctly recognized phonemes, substitutions, and deletions, respectively. In the table, the speaker-independent HMM is the SI neutral style model obtained in **Step 0**. The speaker-adapted neutral style HMM is the one adapted from the SI neutral style model using MLLR+MAP with the target speaker's five sentences of neutral style speech. Similarly, the speaker-adapted style-independent HMM is the one adapted from the SI neutral style HMM using MLLR+MAP with the target speaker's five sentences for each style, 15 sentences in total. The style-adapted HMMs are the ones obtained in **Step 1** using the target speaker's five sentences of the respective styles. It is noted that we assumed that the style of the input speech was known when using the style-adapted HMMs, and unknown for the other models. From the result, we can see that the error rates of the MRHMM significantly decreased compared with the speaker-independent HMM. Moreover, we confirmed the improvement of recognition performance to be statistically significant at the 1%

Table 3.2: Comparison of phoneme error rates (%) between ordinary HMMs and MRHMM.

Model	HMM				1-D MRHMM
Speaker	Independent	Adapted	Adapted	Adapted	Adapted
Style	Neutral	Neutral	Independent	Adapted	Adapted
Neutral	12.3	8.6	8.8	8.6	8.6
Sad	16.8	13.4	11.8	11.5	11.1
Joyful	19.4	16.7	14.9	13.7	13.5
Overall	16.2	12.9	11.8	11.2	11.0

level between the MRHMM and one of the ordinary HMMs except for the style-adapted HMMs. It should be again noted that the results for the style-adapted HMMs were obtained under the condition where the input speech's style was known. It has been found that the recognition performance of the style-adapted HMMs becomes worse when the style of input speech is unknown.

3.3.3.2 Results of style estimation and classification

We also evaluated the performance of the proposed technique in terms of the style estimation. The style classification test was conducted for the test speech samples using the following classification criterion: if the value of the style vector is less than -0.5, then the input speech is classified into sad style; if it is greater than 0.5, then joyful style; otherwise, neutral style. Table 3.3 shows the average classification rates of the respective styles for the test speech samples of three speakers. In total, about 89% of the speech data were classified as the correct style class of the input speech. This would be promising results in the sense that we estimated the degree of expressivity of the input speech without using prosodic features.

Figure 3.3 shows the histograms of the estimated values of the style vectors for the test speech samples. It can be seen that different styles give different distributions and the estimated values of the style vector are distributed around the values that were set in the training. However, there is

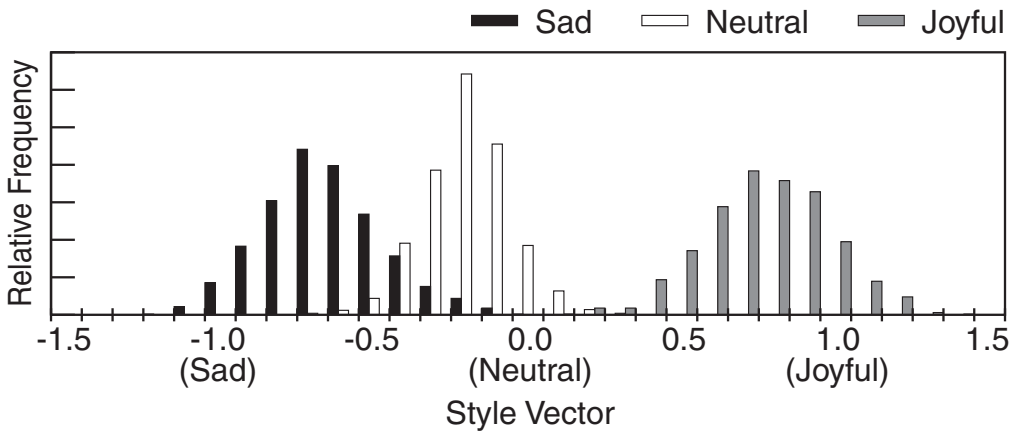
Table 3.3: Classification rates (%) for professional narrators' emotional speech.

Input Style	Classified Style		
	Neutral	Sad	Joyful
Neutral	98.2	0.4	1.4
Sad	14.8	85.2	0.0
Joyful	15.6	0.0	84.4

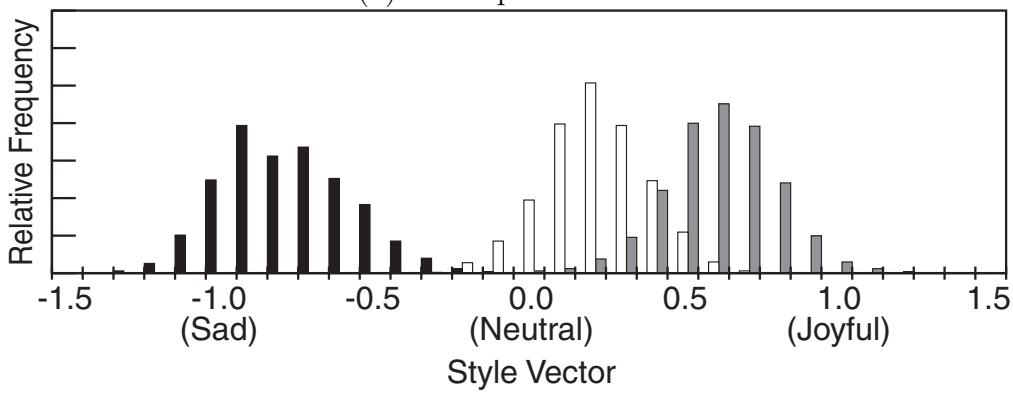
a slight displacement between the mode of each distribution and the value of the style vector assumed in the training. This is because the acoustic features of the sad and joyful styles included in the speech database are not completely symmetric and those of the neutral style are not absolutely located on the mid-point between the sad and joyful styles. As a result, the three styles were influenced by each other in the MRHMM training.

3.3.4 Performance evaluation with non-professional speakers

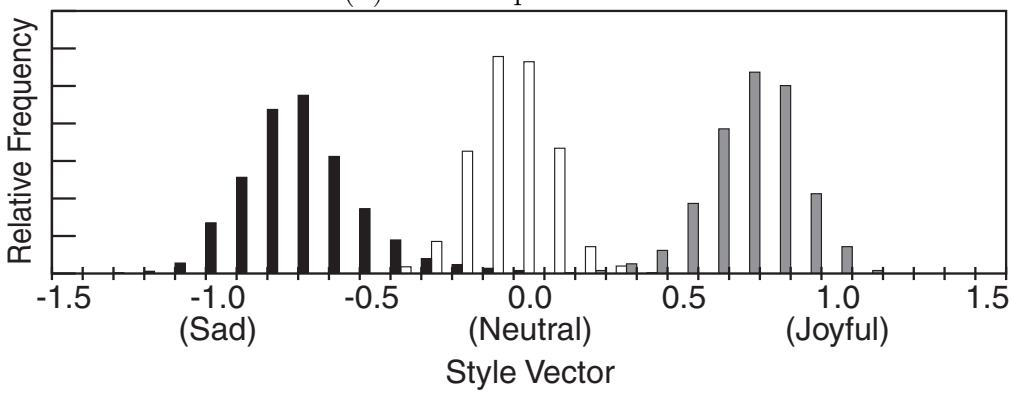
We next assessed the performance of the proposed technique using non-professional speakers' speech which is a little more realistic situation than focusing on the professional narrators' speech. We used four styles of the nine non-professional speakers' speech with simulated emotion. Two different style spaces, namely a two-dimensional space (Fig. 3.2(b)) and a three-dimensional one (Fig. 3.2(c)) were used for modeling MRHMMs. In the two-dimensional space, the style vectors for adaptation data were set to $(0, 0)$, $(1, 0)$, $(0, 1)$, and $(-1, 0)$ for the neutral, joyful, angry, and sad styles, respectively. In the three-dimensional space, the style vectors for adaptation data were set to $(0, 0, 0)$, $(1, 0, 0)$, $(0, 1, 0)$, and $(0, 0, 1)$ for the neutral, sad, joyful, and angry styles, respectively. We performed two-fold cross-validation tests using 50 test sentences that were not included in the adaptation data.



(a) Male speaker MMI



(b) Female speaker FTY



(c) Male speaker MJI

Figure 3.3: Histograms of the estimated values of the style vectors.

Table 3.4: Phoneme error rates (%) for non-professional speakers' emotional speech with different style spaces.

Model	HMM			
Speaker	Independent	Adapted	Adapted	Adapted
Style	Neutral	Neutral	Independent	Adapted
Neutral	15.1	11.2	11.4	11.2
Sad	18.6	15.7	14.8	14.0
Joyful	19.4	16.4	15.3	15.3
Angry	23.4	20.6	19.0	18.8
Overall	19.1	16.0	15.1	14.8

Model	2-D MRHMM	3-D MRHMM
Speaker	Adapted	
Style	Adapted	
Neutral	11.1	10.9
Sad	13.5	13.4
Joyful	14.7	14.7
Angry	17.7	17.7
Overall	14.3	14.2

3.3.4.1 Effect of the choice of style spaces for speaker and style adaptation performance

We first examined whether the choice of style spaces affect the recognition performance. Table 3.4 shows the average scores of the nine speakers' phoneme recognition error rates of respective styles. In the table, the entries for "2-D" and "3-D" represent the results for the MRHMM with the two-dimensional and three-dimensional style spaces, respectively. For comparison, we also evaluated the four types of ordinary HMMs described in Sect. **3.3.3.1**. The speaker-adapted style-independent HMM was obtained using adaptation data of the target speaker's five sentences for each style, 20 sentences in total. We again assumed that the style of the input speech was known when using the style-adapted HMMs, and unknown for the other

models. It can be seen that both of the 2-D and 3-D MRHMMs gave lower error rates than the ordinary HMMs. It was found that there are significant differences at the 1% level between the ordinary HMMs and the MRHMMs. As for the style spaces, the error rates are comparable in scores between the 2-D and 3-D style spaces, and it seems that the recognition performance is not sensitive to the choice of style spaces.

3.3.4.2 Results of Style Estimation and Classification Using Different Style Spaces

Next, we compared the estimation performance of the degree of emotional expressivity between the 2-D and 3-D style spaces. Figure 3.4 shows the distributions of the estimated values of the style vectors using the 2-D style space for all the test samples of one female and two male speakers who were arbitrarily chosen from the nine speakers, and Fig. 3.5 shows those using the 3-D style space. We can see that the distribution of the estimated style vectors belonging to the same style differs from those of other styles.

Table 3.5 shows the average classification rates of the styles for the test speech samples of nine speakers. The input speech samples were classified based on the Euclidean distance between the predetermined style vector used in the training (see Figs. 3.2(b) and (c)) and the estimated style vector. The overall correct classification rates of the MRHMM were 86.8% and 91.3% for the 2-D and the 3-D style spaces, respectively.

3.3.4.3 Performance evaluation in continuous speech recognition

We examined the performance of the proposed technique for non-professional speakers in terms of the word error rate in continuous speech recognition. Adaptation data and other experimental conditions are described in Sect. **3.3.1** and Sect. **3.3.2**. We performed 10-fold cross-validation tests using 50 test sentences that were not included in the adaptation data. The style space for the MRHMM was the 2-D one. For comparison, we again evaluated the speaker-independent HMM and the speaker-adapted style-independent HMM in Sect. **3.3.4.1**. We used Julius (ver. 4.1) [28] as a decoder. We used one of the sets of lexicons and language models contained in [29]. The

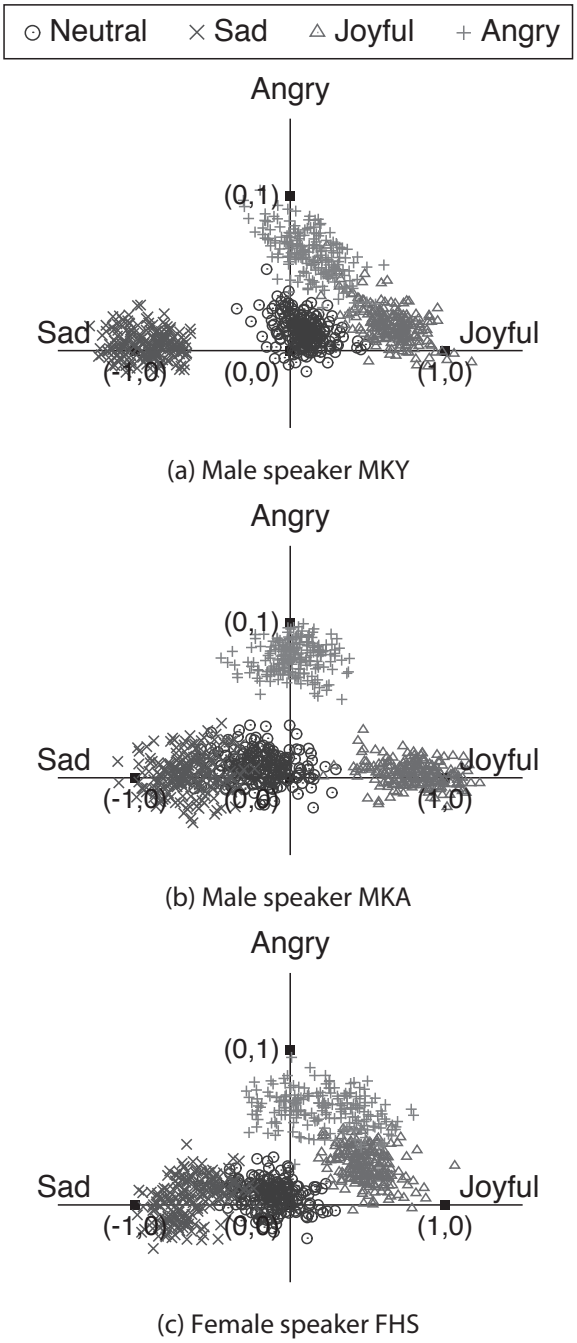
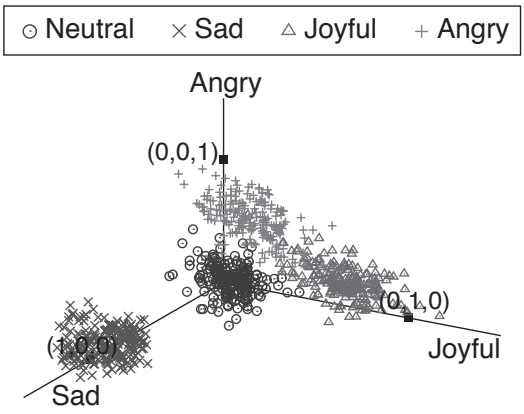
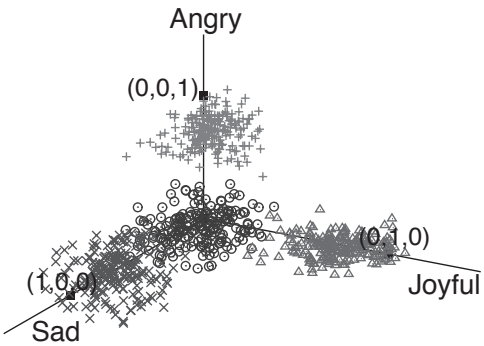


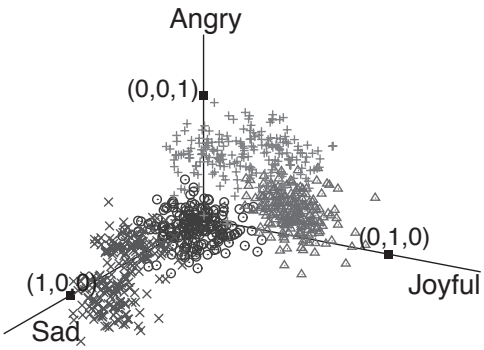
Figure 3.4: Examples of the distributions of the estimated style vector with 2-D style space.



(a) Male speaker MKY



(b) Male speaker MKA



(c) Female speaker FHS

Figure 3.5: Examples of the distributions of the estimated style vector with 3-D style space.

Table 3.5: Classification rates (%) for non-professional speakers' emotional speech with different style spaces.

(a) 2-D style space

Input Style	Classified Style			
	Neutral	Sad	Joyful	Angry
Neutral	99.6	0.3	0.1	0.0
Sad	6.2	93.8	0.0	0.0
Joyful	37.3	0.0	61.2	1.5
Angry	4.9	0.0	2.5	92.6

(b) 3-D style space

Input Style	Classified Style			
	Neutral	Sad	Joyful	Angry
Neutral	99.4	0.5	0.1	0.0
Sad	1.1	99.9	0.0	0.0
Joyful	22.7	0.0	76.3	1.1
Angry	7.6	0.0	2.0	90.4

vocabulary set of the lexicon contains 60k words, and consists of the most frequent words in Mainichi newspaper articles from the year 1991 to 1994 (45 months). The language models were 2-gram and backward 3-gram for the first and second pass, respectively, and obtained from above newspaper corpus. Although there are some out-of-vocabulary words in the lexicon for test speech sentences, we did not add any words to the lexicon.

Table 3.6 shows the word error rates for the respective models. We can see that the MRHMM gave the highest performance in all styles. The difference between the speaker-adapted style-independent HMM and the MRHMM is statistically significant at the 1% level. These results show that the proposed technique would also be effective in LVCSR.

Table 3.6: Comparison of word error rates (%) between ordinary HMMs and MRHMM.

Model	HMM		2-D MRHMM
Speaker	Independent	Adapted	Adapted
Style	Neutral	Independent	Adapted
Neutral	29.0	23.4	23.2
Sad	36.4	30.4	29.0
Joyful	37.2	31.3	30.1
Angry	44.2	38.2	35.9
Overall	36.7	30.8	29.5

3.4 Conclusion

This chapter proposed a technique for emotional speech recognition using rapid model adaptation, in which paralinguistic as well as linguistic information can be obtained. The technique utilizes a multiple-regression HMM (MRHMM) framework, and is based on style estimation and adaptation. Using a speaker-independent neutral style model, the MRHMM is trained with a small amount of target speaker's data. Furthermore, the acoustic models for speech recognition are adapted to the style of input speech from the trained MRHMM using the estimated style vector. From the experimental results of phoneme and continuous speech recognition, we found that the performance of the proposed technique in both speech recognition and style estimation is promising for simulated emotional speech.

Chapter 4

Average Voice Model Training Based on Speaker Class

This chapter describes an average voice model training technique using a speaker class label that represents the voice characteristics of speakers to generate synthetic speech with enhanced similarity to the target speakers' speech. In the proposed technique, first, all training speakers are clustered to determine the speaker classes of all training speakers. The average voice model is trained using the labels of conventional context and the obtained 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 consequence, the synthesized speech of the target speaker is generated from the target speaker's model and the estimated target speaker's speaker class. We conducted objective and subjective evaluations to compare the performance of the proposed technique with that of the conventional one. In the objective evaluations, we evaluated initial models and adapted models by changing the number of adaptation sentences. Subjective evaluation results showed the proposed technique can synthesize speech with improved similarity and naturalness.

A part of this chapter was presented at SSW8 [30].

4.1 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 [10], the average-voice-based speech synthesis technique with model adaptation [11] 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 [12]. 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 [31]. 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. [19] 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 [32]. 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 [31]. 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.

4.2 Speech synthesis system with speaker class label

4.2.1 Overview of proposed speech synthesis system

A block diagram of the proposed speech synthesis method is shown in Fig. 4.1. 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:

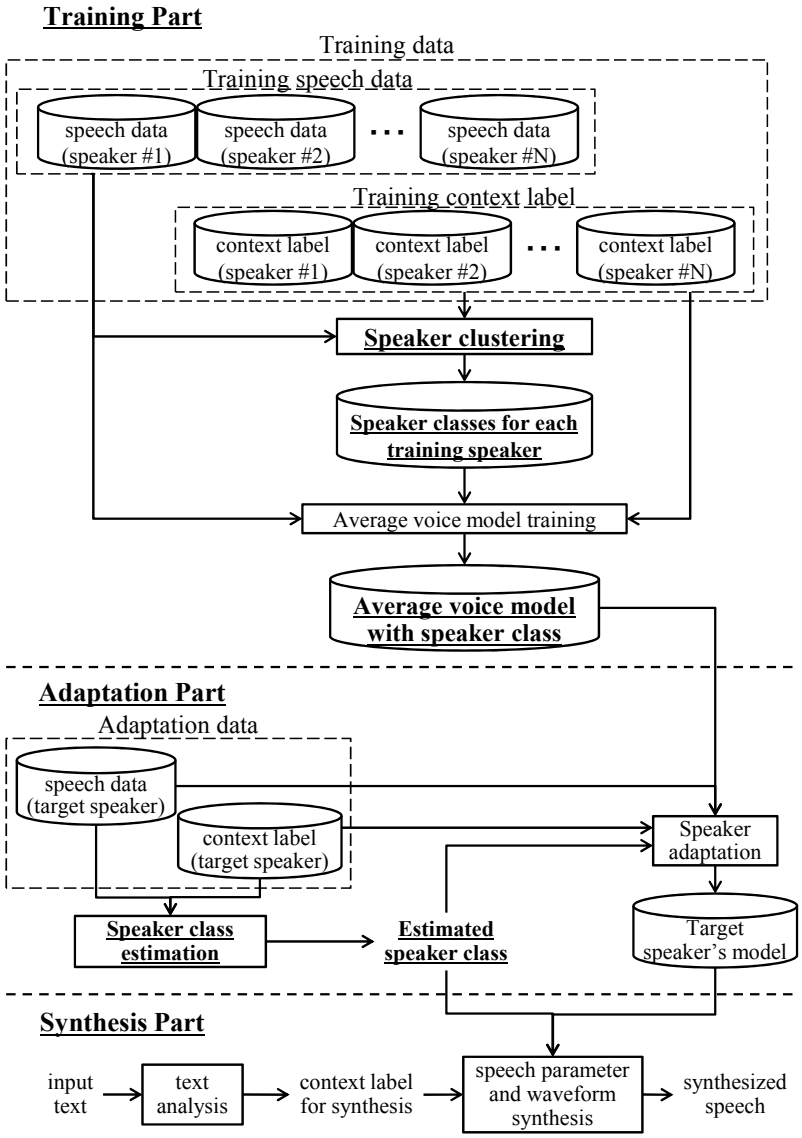


Figure 4.1: Block diagram of the speech synthesis system.

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 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.

4.2.2 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 [31], [33]–[35]. 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.

4.2.2.1 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 [36].

4.2.2.2 Average log F0

Average log F0 of all training data are used for F0-based speaker clustering. As per Sect. **4.2.2.1**, the average log F0 was obtained from only voiced frames as detected by TEMPO [36].

4.2.2.3 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

$$SR_i = \frac{Mora_i}{UttLen_i} \tag{4.1}$$

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

4.2.3 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.

4.2.4 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 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.

4.3 Experiments

4.3.1 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 [36] 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.

Table 4.1: 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

We used the combined technique of CSMAPLR and MAP adaptation as the speaker adaptation algorithm [37].

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.

4.3.2 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 4.1 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.

Table 4.2: 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 4.3: 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**

4.3.3 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 4.4: 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

Tables 4.2–4.4 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 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.

Tables 4.5–4.7 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.

Table 4.5: 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 4.6: 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

Table 4.7: 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

4.3.4 Subjective evaluation

We conducted a XAB test to evaluate voice characteristics and prosodic features of the synthesized speech using the model adapted from the con-

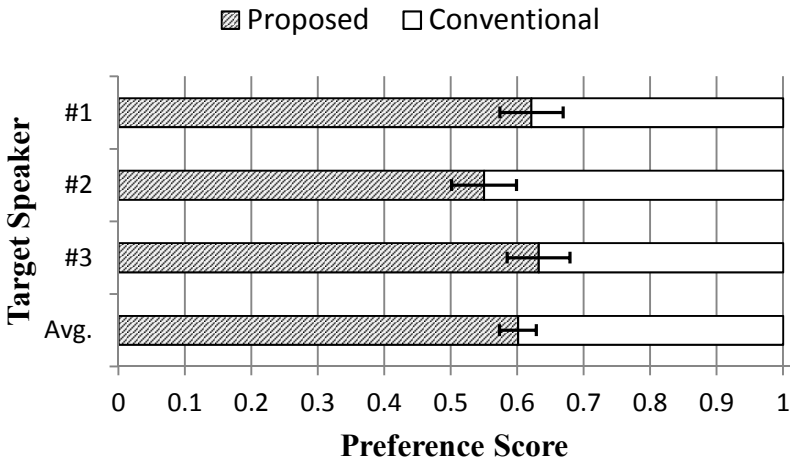


Figure 4.2: Preference score from the subjective evaluation. (Error bars show the 95% confidence intervals.)

ventional 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 4.2 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 [31].

4.4 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.

Chapter 5

Similar Speaker Selection Technique Based on Distance Metric Learning

This chapter analyzes the correlation between various acoustic features and perceptual voice quality similarity, and proposes a perceptually similar speaker selection technique based on distance metric learning. To analyze the relationship between acoustic features and voice quality similarity, we first conduct a large-scale subjective experiment using the voices of 62 female speakers and perceptual voice quality similarity scores between all pairs of speakers are acquired. Next, multiple linear regression analysis is carried out; it shows that four acoustic features are highly correlated to voice quality similarity. The proposed speaker selection technique first trains a transform matrix based on distance metric learning using the perceptual voice quality similarity acquired in the subjective experiment. Given an input speech, acoustic features of the input speech are transformed using the trained transform matrix, after which speaker selection is performed based on the Euclidean distance on the transformed acoustic feature space. We perform speaker selection experiments and evaluate the performance of the proposed technique by comparing it to speaker selection without feature space

A part of this chapter was presented in Proc. INTERSPEECH 2011 [38], Proc. INTERSPEECH 2012 [39] and IEICE Trans. Inf. and Syst. [40].

transformation. The results indicate that transformation based on distance metric learning reduces the error rate by 53.9%.

5.1 Introduction

In the previous chapter, we have presented the average voice model training technique based on speaker class. From results of evaluations, we have confirmed the quality of the synthesized speech is improved by the proposed technique, although the proposed technique utilized only simple acoustic features for speaker clustering. However, to achieve the further improvement of this technique, it would be required to select perceptual similar speakers to the target speaker. Although, a variety of approaches have been proposed to select similar speaker selection, these techniques employ acoustic feature similarities such as likelihood of Gaussian mixture models (GMMs) [41]. However, even if two speakers have similar acoustic features' distributions, their voice quality is not necessarily perceptually similar. In order to enhance the effectiveness of speaker selection, we need to identify perceptual similar speakers. To do this we rely on two key components: (1) identification of acoustic features that have high correlation with perceptual voice quality similarity, (2) a speaker selection technique that takes into account the similarity of the perceived voice quality, not merely acoustic similarity.

To achieve the first goal, a variety of approaches have been proposed to analyze the relationship between speaker characteristics and acoustic features [33]–[35]. Studies have shown that perceptual similarity is associated with prosodic features, consisting of fundamental frequency (F0) and phoneme duration, and acoustic features, consisting of cepstral coefficients and the aperiodic component. Because voice quality and prosody are evaluated simultaneously in subjective experiments, it was not clear which acoustic feature significantly influenced the human perception of voice quality [35]. Furthermore, because published similarity analyses considered only a dozen speakers at most, the range in voice qualities covered remains inadequate. For identifying the various relationships between acoustic features and perception in depth, it is essential to analyze the voices of many speakers.

Regarding the second goal, even if highly correlated acoustic features

with perceptual voice quality similarity are found, it is not appropriate to simply use the Euclidean distance of each acoustic feature. Multiple regression analysis is widely used since it can weight each feature, but "distance metric learning [15]" (DML) is more effective since it can take side information into account. Many studies on DML have demonstrated its usefulness in applications such as image retrieval [42], music retrieval [43], and sentence retrieval [44]. This technique can realize speaker selection if the side information is set properly. We used the perceptual voice quality similarity obtained from a subjective experiment as the side information. In addition, DML has also been used for feature space transformation in a number of studies. For instance, [42] used transformation of the original image space for image retrieval. In this chapter, since the perceptual voice quality similarity is used as the side information, DML can be considered to be transformation from acoustic feature space to perceptual voice quality similarity space.

In this study, our aims are to identify the acoustic features useful for the selection of perceptually similar speakers and to propose a speaker selection technique based on DML. We first conduct a large-scale subjective experiment using 62 female speakers to identify perceptual voice quality similarity. In the experiment, to exclude the influence of prosody, we use speech modified so as to exhibit exactly the same prosody (F0 and phoneme duration). Several acoustic features highly correlated to perceptual voice quality similarity are found by regression analysis of the results of the subjective experiment. In the proposed similar speaker selection technique, the transform matrix is first trained on the basis of DML to convert the acoustic feature space. Given a speech sample, the acoustic features of the sample are transformed using a trained transform matrix. Then, a similar speaker is chosen on the basis of Euclidean distances on the transformed acoustic feature space. To evaluate the proposed technique's performance, we compare it, in experiments, to speaker selection on an acoustic feature space without transformation. The results thus obtained demonstrated the technique's effectiveness.

5.2 Speech database and subjective experiment

We first conducted a subjective experiment to evaluate voice quality similarity between many speakers. Speech stimuli and details of the subjective evaluation are described below.

5.2.1 Speech database

We used the speech data of 62 female speakers included in the NTT-AT Japanese multi-speaker's speech database [45]. The sampling frequency of the speech was 16 kHz and the quantization bit rate was 16 bits. This database contains about 200 phonetically balanced sentences for each speaker. The speakers' ages ranged from 18 to 49.

5.2.2 Speech samples generated for the evaluation

For the subjective experiment, we used a single sentence, "Shoo enerugii ga sakebarete imasu", (in English "Energy savings are desired") spoken by 62 non-professional female speakers included in the NTT-AT database.

To analyze the relationship between perceptual voice quality similarity and acoustic features, this evaluation removed the parameter of the prosody of speech. In this experiment, prosody modified speech with the prosody (F0 and phoneme duration) extracted from a speech uttered by a speaker other than the chosen 62 speakers in the NTT-AT database, was employed as the speech stimuli. To generate the speech stimuli with target prosody, original acoustic features (spectrum and aperiodic component) of each speech were linearly interpolated according to target duration and the F0 was modified to match the target F0. The interpolation was executed within each manually segmented phoneme boundary. We used the STRAIGHT [36] vocoder for speech analysis and synthesis. The analysis frame shift was 1 ms.

Table 5.1: Evaluation criteria.

Score	Description
3	very similar
2	slightly similar
1	dissimilar

5.2.3 Subjective experiment for evaluation of perceptual voice quality similarity

A subjective experiment using the 62 speech stimuli was carried out. Subjects heard 3844 pairs (62×62) of speech stimuli, and rated the similarity of the presented speech pair. In order to counter the bias created by the order of stimuli presentation, the stimuli were also presented in inverse order. The rating scale is shown in Table 5.1. The subjects were 32 people (14 males and 18 females) who were listening to the speech stimuli for the first time. Each pair was evaluated by eight persons. Let $s(i, j)$ be the perceptual voice quality similarity between speaker i and j averaged over the evaluation scores of eight people. The voice quality similarity matrix component $Sim(i, j)$ is represented as follows.

$$Sim(i,j) = \begin{cases} \dfrac{s(i,j) + s(j,i)}{2} & (i \neq j) \\ s(i,j) & (i = j) \end{cases} \tag{5.1}$$

This yielded the voice quality similarity matrix, $Sim(i, j)$, for the 62 speakers.

Since each speech stimulus was evaluated by several different people in the subjective evaluation, the obtained similarity matrix might have been affected by differences in the listeners' evaluation criteria. However, it would be difficult to avoid this problem by performing a larger scale experiment because of its cost. Furthermore, the huge amount of evaluations by the same people that would be obtained in such an experiment might result in a lack of consistency in the evaluations. Therefore, each speech stimulus was evaluated by eight people we consider a minimum of subjects for subjective evaluation.

5.3 Regression analysis between voice quality similarity and acoustic features

5.3.1 Acoustic features

In analyzing the relationship between the perceptual voice quality similarity and acoustic features, we focused on ten acoustic features as described below.

- Low dimensional (1 to 12 dimensions) cepstral coefficients (CepL).
- High dimensional (13 to 24 dimensions) cepstral coefficients (CepH).
- Low dimensional (1 to 12 dimensions) cepstral coefficients using log spectrum from 0 kHz to 4 kHz (Cep4k).
- 1 to 12 dimensional coefficients of DCT value of aperiodic component (AP).
- Average value of aperiodic component in full band (APm).
- Ratio of the power in each sub-band to the power in full band (PR1–PR5).

PR of i-th sub-band PRi is represented as follows.

$$PRi = \frac{mean(spec_i)}{mean(spec_{full})}. \tag{5.2}$$

where, $spec_i$ and $spec_{full}$ represent respectively the spectrum in i-th band and the spectrum in full band ($0-8$ kHz). In this study, the spectrum was divided into 5 sub-bands ($0-1$, $1-2$, $2-4$, $4-6$, and $6-8$ kHz), using a spectral division method similar to that used for the aperiodic component in HMM-based speech synthesis.

Although many auditory features that take human perception into consideration, such as the perceptual linear predictive (PLP) feature [46], have been proposed, the goal of this study is not only achieving similar speaker selection but also applying it to speech synthesis. For this reason, we chose to use the acoustic features generally used in speech synthesis, i.e., cepstrum

Table 5.2: Correlation coefficients between all acoustic features.

	Cep4k	CepH	APm	AP	PR1	PR2	PR3	PR4	PR5
CepL	0.448	-0.145	0.143	0.232	0.221	0.177	0.235	0.205	0.169
Cep4k		-0.140	0.107	0.376	0.363	0.331	0.304	0.238	0.176
CepH			-0.241	-0.286	-0.293	-0.118	-0.338	-0.286	-0.254
APm				0.415	0.381	0.106	0.492	0.118	-0.003
AP					0.413	0.372	0.346	0.242	0.177
PR1						0.402	**0.799**	**0.638**	0.489
PR2							0.153	0.022	0.063
PR3								0.356	0.257
PR4									0.501

coefficients and an aperiodic component. We also used simple acoustic features such as power ratios (PR1–PR5) since simple features can be converted easily when synthesizing speech. In addition, previous studies, such as [47], showed the cepstrum features of speech, especially for high order cepstra, are affected by prosodic features. However, since the purpose of this chapter is to identify acoustic features that have high correlation with perceptual voice quality similarity, we did not consider pitch information so as to exclude the effect of prosodic features. Furthermore, it should be noted that we used the cepstrum obtained from a lower-band rather than a higher-band log spectrum. This is because we believe that voice quality similarity would be more affected by the rough characteristics of the higher-band spectrum than affected by the detailed spectral shape of it. Since we used spectrum power ratio (PR1–PR5), we were able to take into account the rough characteristics of the higher-band spectrum in the following analysis.

As the acoustic distance measure of each speaker, we used the Euclidean distance of these acoustic features for each speaker's speech. First, an acoustic feature of the prosody modified speech used in the subjective experiment was extracted by STRAIGHT in every frame. Second, the Euclidean distance between the acoustic feature of one speaker and that of another speaker's speech was calculated in the frame; the average Euclidean distance is de-

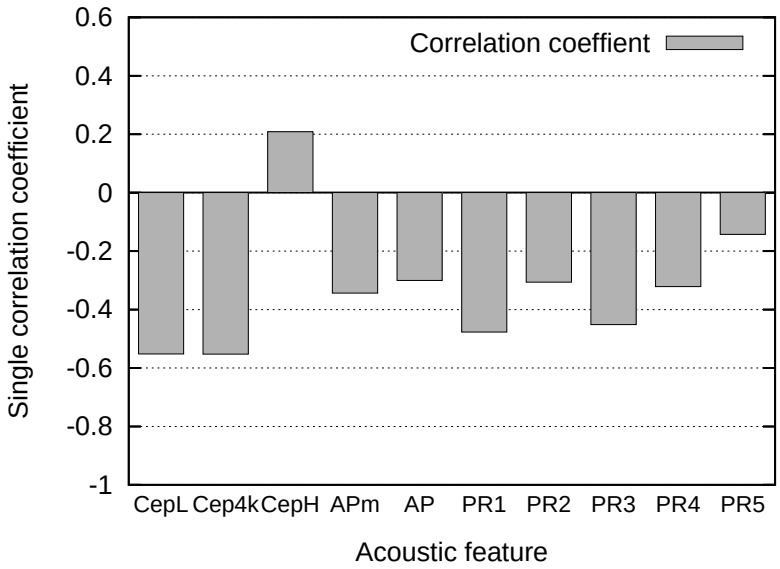


Figure 5.1: Single correlation coefficients between perceptual voice quality similarity and each acoustic feature.

fined as the distance between the two speakers. The analysis frame shift was 1 ms. Because voice quality characteristics are chiefly presented by voiced phonemes rather than unvoiced phonemes, the distance was calculated using only voiced frames as detected by TEMPO [36].

As a result, the distance matrix of each acoustic feature was obtained as well as the voice quality similarity matrix.

In order to analyze the relationship between the perceptual voice quality similarity and acoustic features, we performed single and multiple regression analysis. In all analyses, the voice quality similarity and the distance matrix were provided except for the combination of same speaker's speech.

5.3.2 Regression analysis

5.3.2.1 Single regression analysis

We first calculate single correlation coefficients between the perceptual voice quality similarity and each acoustic feature. Figure 5.1 shows single correlation coefficients for each acoustic feature. In this figure, because we calculated

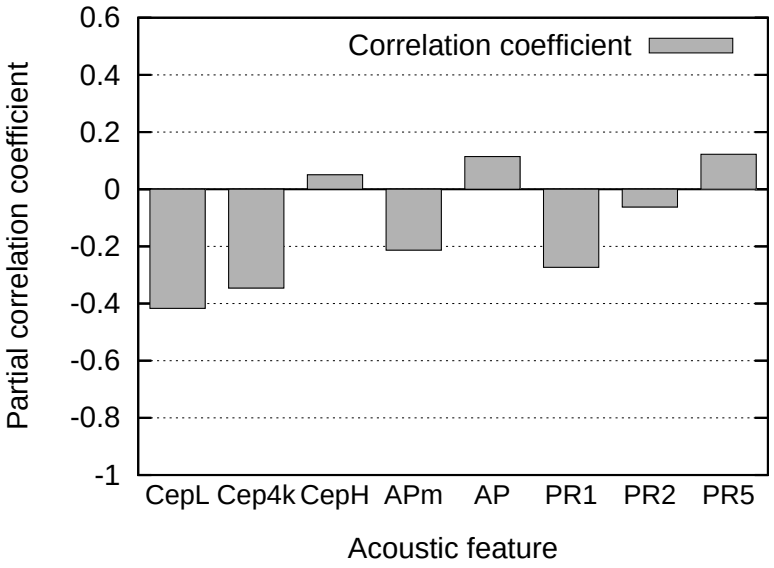


Figure 5.2: Partial correlation coefficients for each acoustic feature.

correlation coefficients between the distance of acoustic features and the perceptual voice quality similarity, acoustic features with negative correlation coefficient have high correlation with the perceptual voice quality similarity. The value of correlation coefficients shows that most acoustic features are correlated with perceptual voice quality similarity to some extent except for CepH and PR5. In particular, the four acoustic features CepL, Cep4k, PR1, and PR3 are correlated with perceptual voice quality similarity because the values of the correlation coefficients are around -0.5.

Table 5.2 lists the correlation coefficients between each acoustic feature. We can see that PR1 has high correlation with PR3(0.799) and PR4(0.638). It is not desirable to utilize these acoustic features simultaneously for multiple regression analysis because doing so may cause multicollinearity. Other combinations have lower correlation.

5.3.2.2 Multiple regression analysis

Next, we perform multiple regression analysis to investigate the effect of linearly combining multiple acoustic features. Eight acoustic features, i.e., CepL, Cep4k, CepH, APm, AP, PR1, PR2, and PR5, were utilized as the

Table 5.3: Bayesian information criterion values for each combination of acoustic features.

# of acoustic feature	combination of acoustic features	BIC value
1	(1) Cep4k	3418.2
2	(2) (1)+CepL	2776.1
3	(3) (2)+PR1	2323.4
4	(4) (3)+APm	2157.4
5	(5) (4)+CepH	2051.1
6	(6) (5)+PR5	1962.9
7	(7) (6)+AP	1900.6
8	(8) (7)+PR2	1895.4

explanatory variables of the regression. To avoid multicollinearity, PR3 and PR4 were not used since we confirmed from Table 5.2 they have high correlation with PR1. To obtain precise results from multiple regression analysis, it is necessary to avoid the use of these acoustic features simultaneously. We therefore used PR1, which has the highest single correlation coefficient. First, a multiple correlation coefficient was calculated using the above eight acoustic features. We confirmed that the perceptual voice quality similarity and the estimated one were highly correlated; the multiple correlation coefficient was "0.741". This result indicates that we can use these acoustic features to estimate voice quality similarity to some extent.

We also calculate the partial correlation coefficient for each acoustic feature. The results are shown in Fig. 5.2. The partial correlation coefficient values indicate that four acoustic features (CepL, Cep4k, APm, and PR1) have high correlation coefficients, which matches the results of Sect. **5.3.2.1**. On the other hand, the other four acoustic features, i.e., CepH, AP, PR2, and PR5, have low correlation coefficients.

Furthermore, in order to confirm the impact on similar speaker selection for each acoustic feature, we investigated the Bayesian information criterion (BIC) values for each combination of acoustic features. Table 5.3 shows the

BIC values. In this table, each column shows the combinations of acoustic features which have the minimum BIC value when changing the number of acoustic features. From this table, we can see that the BIC values decrease as the number of acoustic features increase. This implies that these eight acoustic features are effective for similar speaker selection to some extent. However, since the BIC value reductions are different according to each acoustic feature, the impact on similar speaker selection is considered to be large in the order corresponding to Cep4k, CepL, PR1, and APm. This result is consistent with the result of multiple regression analysis.

From these results, these four acoustic features, i.e., CepL, Cep4k, APm, and PR1, are considered to be acoustic features highly correlated with perceptual voice quality similarity. Thus, in the following speaker selection experiments, we used these four acoustic features. Although the other acoustic features, i.e., CepH, AP, PR2, and PR5, seem to be effective for speaker selection from the BIC values, we did not use these features since they correlated poorly with the perceptual voice quality similarity. We also calculated multiple correlation coefficients by using three (CepL+Cep4k+PR1) and four (CepL+Cep4k+PR1+APm) selected features. The multiple correlation coefficients obtained were 0.704 and 0.720, respectively.

5.4 Speaker selection technique based on distance metric learning

Next, we use distance metric learning (DML) to propose a similar speaker selection technique using the obtained acoustic features. An overview of our proposed selection system and its details are given below.

5.4.1 Overview of proposed similar speaker selection

A block diagram of the proposed method is shown in Fig. 5.3. In the proposed technique, we first employ distance metric learning to train a transform matrix using training data with speaker class. When an input utterance is given, the input utterance vector, described in Sect. **5.4.4**, extracted from the input utterance is transformed using the trained transform matrix. After

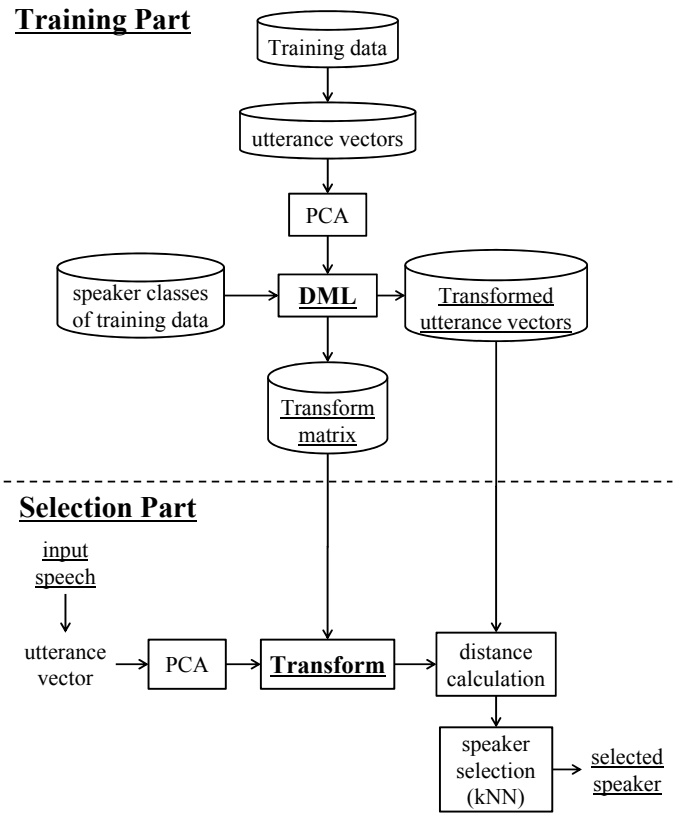


Figure 5.3: A block diagram of the speaker selection system based on distance metric learning.

that, k-nearest neighbor (kNN) [48] classifier-based speaker selection is performed by calculating the Euclidean distance between the transformed input utterance vector and the transformed utterance vectors extracted from all training data. The overall speaker selection process is summarized below.

Training part:

Step 1 Extract training utterance vectors for each utterance from all training data.

Step 2 Perform PCA using the extracted training utterance vectors for dimension reduction.

Step 3 Perform DML (RCA) to obtain the transform matrix $\boldsymbol{A}$ and trans-

formed utterance vectors (training vectors) using the speaker classes of training data and dimension-reduced training utterance vectors.

Selection part:

Step 4 Extract an input utterance vector from the input speech.

Step 5 Perform PCA using the extracted input utterance vector for dimension reduction.

Step 6 Transform the input utterance vector using the transform matrix **A** obtained from **Step 3**.

Step 7 Calculate the Euclidean distances between the transformed input utterance vector and the transformed training utterance vectors obtained from **Step 3**.

Step 8 Select one speaker as the most similar speaker, i.e., the speaker having the most frequent vectors among the k nearest neighbor vectors.

Because utterance vectors (described in Sect. **5.4.4**) are generally highly dimensional vectors, it is necessary to reduce the number of dimensions of the training vector to avoid the curse of dimensionality. For this reason, we perform PCA to achieve simple dimension reduction in **Step 2** and **Step 5**. After the dimension reduction, DML is performed using the dimension-reduced training vectors.

Details of each component, i.e., distance metric learning, the utterance vector, the speaker class, and kNN classifier-based speaker selection, are described as follows.

5.4.2 Distance metric learning

Let us denote a set of N vectors in d-dimensional space as $\boldsymbol{X} = \{x_i \in \mathbb{R}^d\}_{i=1}^N$, where the Mahalanobis distance between two vectors $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ is defined as

$$d(\boldsymbol{x}_i, \boldsymbol{x}_j) = (\boldsymbol{x}_i - \boldsymbol{x}_j)^T \boldsymbol{M} (\boldsymbol{x}_i - \boldsymbol{x}_j) \tag{5.3}$$

where $\boldsymbol{M}$ is a positive semi-definite matrix that satisfies valid metric properties. The goal of DML is to find an optimal Mahalanobis matrix $\boldsymbol{M}$ from

the side information. We can uniquely decompose any positive semi-definite matrix to $\boldsymbol{M} = \boldsymbol{A}^T\boldsymbol{A}$. This reduces Eq. (5.3) to

$$d(\boldsymbol{x}_i, \boldsymbol{x}_j) = \| \boldsymbol{A}(\boldsymbol{x}_i - \boldsymbol{x}_j) \|_2 \tag{5.4}$$

the Euclidean distance after transformation is $\boldsymbol{x}_i \rightarrow \boldsymbol{A}\boldsymbol{x}_i$. Thus DML is equivalent to transformation of the vector space using matrix $\boldsymbol{A}$.

In this study, in order to the avoid sparse data problem, we used Relevant Component Analysis (RCA) [49], a well known supervised distance metric learning method. Although a number of DML techniques, such as Neighborhood Component Analysis (NCA) [50] and Large Margin Nearest Neighbour (LMNN) [51], have been proposed to train a more precise transform matrix $\boldsymbol{A}$, these techniques generally require much training data. However, since our proposed technique requires the perceptual voice quality similarity obtained from the subjective evaluation, we cannot collect sufficient training data for such DML techniques. Therefore, we used RCA to train the transform matrix because this technique is simple and effective.

5.4.2.1 Relevant component analysis

Given a set of vectors, $\boldsymbol{X} = \{x_i\}_{i=1}^N$, and setting the K class for each vector, RCA trains global linear matrix $\boldsymbol{M}$ to minimize the distance between the vectors in each class. The optimal transformation by RCA is computed as $\boldsymbol{A} = \hat{\boldsymbol{C}}^{-1/2}$ and the Mahalanobis matrix is equal to the inverse of the average covariance matrix of classes, i.e., $\boldsymbol{M} = \hat{\boldsymbol{C}}^{-1}$, where $\hat{\boldsymbol{C}}$ is defined as follows:

$$\hat{\boldsymbol{C}} = \frac{1}{N}\sum_{j=1}^{K}\sum_{i=1}^{N_j}(\boldsymbol{x}_{ji} - \boldsymbol{\mu}_j)(\boldsymbol{x}_{ji} - \boldsymbol{\mu}_j)^T \tag{5.5}$$

here, $\boldsymbol{\mu}_j$ denotes the mean of the j-th class, and x_{ji} denotes the i-th vector in the j-th class; N and N_j are the total number of vectors and the total number of vectors in the j-th class, respectively.

To apply RCA to speaker selection, we need to define the class and the vector. In this chapter, the class and the vector are called the speaker class and the utterance vector, respectively.

5.4.3 Speaker class using perceptual voice quality similarity

To set the speaker class for each speaker, we adopt a speaker clustering technique based on perceptual voice quality similarity. We utilize the perceptual voice quality similarity matrix as the speaker vector obtained from Sect. **5.2.3**. Let $\boldsymbol{v}_i$ be the speaker vector of speaker i. It is represented as

$$\boldsymbol{v}_i = [Sim(i,1), \cdots, Sim(i,j), \cdots, Sim(i,N_s)] \tag{5.6}$$

where $Sim(i,j)$ represents the perceptual voice quality similarity between speakers i and j, and N_s represents the number of speakers participating in the subjective experiment. In this chapter, we set N_s to 62. Speaker clustering is done by applying the k-means algorithm to the speaker vectors.

5.4.4 Utterance vector

We utilized the GMM supervector [52] as the utterance vector to realize a text-independent similar speaker selection technique because its effectiveness in text-independent speaker recognition has been confirmed. The GMM supervector was created by concatenating the mean parameter of an individual GMM mixture. Given a speaker utterance, MAP adaptation is performed using a speaker-independent GMM that is trained in advance. Let μ_{ij} be the mean parameter of the adapted GMM's output distribution for mixture i and dimension j. The GMM supervector $\boldsymbol{m}$ is represented as

$$\boldsymbol{m} = [\mu_{11}, \cdots, \mu_{ij}, \cdots, \mu_{ML}] \tag{5.7}$$

where M and L represent, respectively, the number of GMM mixtures and the number of acoustic features' dimensions.

To advance the field of speaker recognition, i-vector [53] was proposed to improve the performance of text-independent speaker recognition. Because the purpose of this chapter is to confirm the effectiveness of applying distance metric learning to similar speaker selection, we used the GMM supervector only in the following experiments.

5.4.5 kNN classifier-based speaker selection

The k-nearest neighbor (kNN) classifier is the simplest classifier of all machine learning algorithms in pattern recognition. Because it is an effective and simple technique, it is used in various research fields. This chapter also uses this technique for speaker selection.

Given an input utterance vector and all training utterance vectors described in Sect. **5.4.4**, the Euclidean distances between an input vector and all training vectors are calculated. Next, the k training vectors that have the smallest distance from the input vector are chosen. Finally, the speaker that yields the greatest number of selected k training vectors is selected as the similar speaker.

5.5 Experiments

5.5.1 Experimental conditions

In the following experiments, we used the speech data of 62 female speakers as described in Sect. **5.2.1**. We used the perceptual voice quality similarity between all speaker pairs (62×62) as determined by the subjective experiment as described in Sect. **5.2.3**. Thirty sentences uttered by 61 of the 62 speakers were used for the training data and 30 sentences uttered by the other speaker not included in the training data were used as the evaluation data. In the selection experiment, we first select one speaker as the evaluated speaker, and one speaker was chosen from the remaining 61 training speakers. We performed a leave-one-out cross-validation test in order to ensure the validity of the results obtained.

We utilized the four acoustic features with the highest correlation with the perceptual voice quality similarity as identified in Sect. **5.3.2.2**, i.e., CepL, Cep4k, APm, and PR1. These acoustic features were extracted using STRAIGHT [36]. The analysis frame shift was 5 ms. Although the frame shift was 1 ms in Sect. **5.2** and **5.3**, we changed it to 5 ms because a 1 ms frame shift is generally too short for GMM supervectors. The following experiments were performed using only voiced frames as in Sect. **5.3.2**.

A speaker-independent GMM was trained from all speech data uttered

by the 62 female speakers (12400 utterances = 62 speakers × 200 utterances) to extract the GMM supervector. We set the number of nearest neighbors in the kNN classifier at 5.

To evaluate the speaker selection performance, we used "average similarity". The average similarity is calculated by the perceptual voice quality similarity between the input speaker and the selected speaker obtained from the above mentioned subjective experiment in Sect. **5.2.3**. Let $sel(utt_{ij})$ be the speaker identified by the speaker selection technique using utt_{ij}, which represents the j-th utterance uttered by speaker i. The average similarity is expressed as

$$\frac{1}{N_{eval}} \sum_{i=1}^{S} \sum_{j=1}^{U} Sim(i, sel(utt_{ij})) \tag{5.8}$$

where N_{eval}, S, and U represent, respectively, the number of evaluation utterances (S by U), the number of evaluated speakers, and the number of utterances per evaluated speaker; $Sim(i, sel(utt_{ij}))$ represents the perceptual voice quality similarity between input speaker i and the selected speaker from utt_{ij}.

5.5.2 Acoustic feature performance

To compare acoustic feature performances, we first performed speaker selection by changing the acoustic features. In this experiment, we did not use RCA to perform distance metric learning. In the proposed speaker selection methods, the optimal selection parameters (i.e., the number of GMM mixtures and the number of PCA dimensions) differ for each combination of acoustic features. Therefore, to set optimal parameters for each combination, we performed a preliminary experiment by changing these parameters. In the experiment, we set the number of GMM mixtures at 32, 64, and 128 and the number of dimensions of PCA from 10 to 40. From the obtained results, we respectively set the number of GMM mixtures for four combinations (i.e., CepL, CepL+Cep4k, CepL+Cep4k+PR1, and CepL+Cep4k+PR1+APm) at 32, 64, 64, and 128, and the number of PCA dimensions at 31, 30, 28 and 27.

Table 5.4: Average similarity for each acoustic feature.

Acoustic feature	Average similarity
CepL	2.35
CepL+Cep4k	2.43
CepL+Cep4k+PR1	2.44
CepL+Cep4k+PR1+APm	2.41

Table 5.4 shows the average similarity obtained for each acoustic feature. We can see that the average similarity increased by adding Cep4k and PR1. On the other hand, the selection performance hardly changed at all when APm was added. This is because the utterance vectors we used fail to make allowance for the temporal characteristics of acoustic features. In Sect. **5.3**, we used speech with exactly the same prosody (F0 and phoneme duration) to exclude the effect of the prosody. In this section, however, we used GMM supervector, which cannot represent temporal characteristics because it represents only the average characteristics of the whole utterance.

5.5.3 Performance comparison with distance metric learning

Next, we performed a speaker selection experiment by changing the number of speaker classes to investigate the effectiveness of distance metric learning in similar speaker selection. As suggested by the previous experiment, we used three acoustic features, i.e., CepL, Cep4k, and PR1. Figure 5.4 shows the average similarity for each speaker class. We can see that the average similarity for each acoustic feature was improved by distance metric learning using RCA.

However, it can be seen that for the case of two speaker classes, the average similarity decreased, but the change was slight, if at all, for four or more classes. This is because RCA fails to take into account the complexity according to the number of speaker classes. In general, when the number of speaker classes is increased, a transform matrix that can process the details of

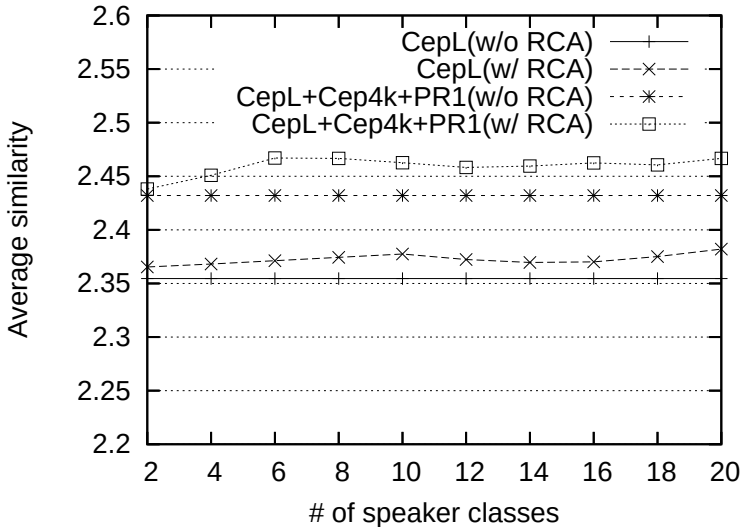


Figure 5.4: Average similarity versus the number of speaker classes.

the acoustic feature space is required. However, RCA can train only a global transform matrix, and so cannot take account of the complexity created by the increase in the number of speaker classes.

5.5.4 Overall performance

We investigated the overall speaker selection performance by combining acoustic features and distance metric learning. Figure 5.5 shows the histogram of the similarity between the selected speaker and the input speaker. The number of speaker classes was set to 8 from the results of Sect. **5.5.3**. It can be seen that the number of speakers having low similarity decreased with distance metric learning when acoustic features were added. To confirm the effectiveness of the proposed technique, we calculated the selection error rate by counting the number of speakers having low similarity. We set the threshold for calculating the error rate to 2.0 because this means that the number of listeners selecting "very similar" is larger than the number of listeners selecting "very dissimilar" in the subjective experiment. As a result, we found it was reduced by 53.9%, i.e., from 20.22% to 9.31%. A paired t-test we performed confirmed that the difference between the two methods

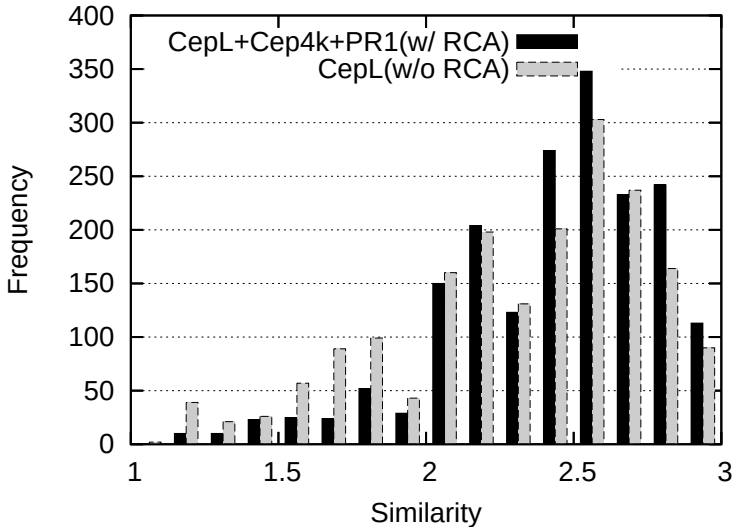


Figure 5.5: Histogram of the similarity between the selected speaker and the input speaker.

Table 5.5: Performance comparison with GMM-based speaker recognition.

Method	Average similarity
GMM	2.43
Proposed (w/o RCA)	2.44
Proposed (w/ RCA)	2.47

is statistically significant at the 1% level. This indicates that the proposed method can significantly reduce the speaker selection error rates.

5.5.5 Comparison with speaker recognition technique

Finally, we compared our proposed technique's speaker selection performance with that of a conventional speaker recognition technique based on GMM [41]. To obtain each speaker's GMM, we performed MAP adaptation from a speaker-independent GMM. As the speaker-independent GMM, we used the same model used in the proposed technique (described in Sect. **5.5.1**). We used CepL+Cep4k+PR1 as the acoustic feature, and 30 sentences uttered

by each speaker were used for MAP adaptation.

Table 5.5 shows the average similarity results obtained from the experiment. These results confirmed that the two methods (the proposed technique without RCA and the GMM-based one) have comparable performance. In addition, applying feature space transformation using RCA confirmed that the average similarity obtained in doing so is higher than that of the GMM-based method. From these results, we confirmed the effectiveness of the proposed technique compared with the conventional speaker recognition technique.

5.6 Conclusion

In this chapter, we analyzed the relationship between the perceptual voice quality similarity and various acoustic features for perceptually similar speaker selection. First, perceptual experiments using 62 female speakers' voices were designed and the perceptual voice quality similarity matrix between each speaker was determined. The results of multiple regression analysis showed that low dimensional cepstrum coefficient, low dimensional cepstrum coefficient under 4 kHz and the aperiodic component had high correlation to perceptual voice quality similarity; the multiple correlation coefficient was "0.741". Furthermore, we have presented a new speaker selection technique that takes perceptual voice quality similarity into account in the selection process. This technique utilizes distance metric learning to transform the acoustic feature space into the perceptual voice quality similarity space. Experiments showed that the proposed technique improves speaker selection performance. In particular, the proposed technique can significantly reduce the speaker selection error rates.

Chapter 6

Conclusions and Future Work

This thesis has presented new approaches to the acoustic modeling techniques automatic speech recognition (ASR) and text-to-speech synthesis (TTS) for achieving personalized speech interfaces. The proposed technique for TTS makes it possible to generate arbitrary speakers' voices. The proposed technique for ASR makes it possible to obtain not only linguistic information but paralinguistic information such as the speaker's emotional expressions.

To achieve personalized ASR, we have proposed a rapid model adaptation technique for emotional speech recognition that enables us to extract paralinguistic information as well as linguistic information contained in speech signals. This technique is based on style estimation and style adaptation using a multiple-regression hidden Markov model (MRHMM). The recognition process consists of two stages. In the first stage, the style vector that represents the emotional expression category and the intensity of its expressiveness for the input speech is estimated on a sentence-by-sentence basis. Next, the acoustic models are adapted using the estimated style vector, and then standard HMM-based speech recognition is performed in the second stage.

For achieving personalized TTS, we have also proposed an average voice model training technique using speaker class labels that represent the voice characteristics of speakers to generate synthetic speech with enhanced similarity to the target speakers' speech. With this technique, all training speakers are first clustered to determine the speaker classes for all of them. The

average voice model is trained using the labels of conventional context and the obtained speaker classes. In the speaker adaptation process, the target speaker's class is estimated and used to transform the average voice model into the target speaker's model. As a consequence, the synthesized speech of the target speaker is generated from the target speaker's model and the estimated target speaker's speaker class. We have also proposed a perceptual similar speaker selection technique based on distance metric learning as the first step for further improving the similarity of the synthesized speech. The technique first trains a transform matrix based on distance metric learning using the perceptual voice quality similarity acquired in a subjective evaluation. Given an input speech, acoustic features of the input speech are transformed using the trained transform matrix, after which speaker selection is performed using the Euclidean distance on the transformed acoustic feature space.

Through the experimental evaluations for the proposed ASR and TTS techniques, we have concluded that the techniques are flexible and statistically convincing approaches to achieving personalized speech interfaces.

6.1 Summary of thesis

Chapter 1 described the general background to the thesis. We overviewed applications using speech interfaces, i.e., ASR and TTS, and pointed out that in new personal device applications speech interfaces play a different role than in conventional applications. The conventional techniques based on model adaptation for achieving a personalized speech interface were introduced, as well as recent research on synthesizing the speech of arbitrary speakers. This was followed by an outline of the thesis scope. We described the basic idea of emotional speech recognition based on multiple-regression MRHMM, then introduced the concept of MRHMM-based emotional speech recognition based on model adaptation. We then described the idea of a model adaptation technique based on speaker class for achieving synthesized speech with enhanced similarity to the target speaker. We also described the idea of a perceptual similar speaker selection technique using distance metric learning.

Chapter 2 presented a technique we proposed for emotional speech recognition using rapid model adaptation, in which paralinguistic as well as linguistic information can be obtained. This technique utilizes the MRHMM framework for the model adaptation and a style vector that corresponds to the degree or intensity of expressivity of styles as the explanatory variable of the regression. In the recognition stage, we adapt an HMM to the input style using the estimated style vector. We showed that the technique reduced the error rates of a style-independent HMM by 11%. We also showed we can obtain not only linguistic information but also the degree of expressivity of emotional and styled speech from the recognition process.

Chapter 3 presented an MRHMM-based emotional speech recognition technique we proposed that uses only a small amount of speech data uttered by the target speaker. A speaker-independent neutral style model is used to train the MRHMM with a small amount of the target speaker's data. The acoustic models for speech recognition are adapted to the style of input speech from the trained MRHMM using the estimated style vector. From the experimental results of phoneme and continuous speech recognition, we found that the technique shows promising performance in both speech recognition and style estimation for simulated emotional speech.

Chapter 4 described an average voice model training technique we proposed that utilizes speaker classes representing the voice characteristics of speakers. In the speaker adaptation process, the speaker class of the target speaker is estimated and used for speaker adaptation and speech parameter generation. Objective and subjective experiments showed that the 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 chapter 5, we analyzed the relationship between perceptual voice quality similarity and various acoustic features for perceptually similar speaker selection. The results of multiple regression analysis showed that a low dimensional cepstrum coefficient (under 4 kHz) and the aperiodic component had high correlation to perceptual voice quality similarity; the multiple correlation coefficient was 0.741. We also presented a novel speaker selection technique that takes perceptual voice quality similarity into account in the

selection process. This technique utilizes distance metric learning to transform the acoustic feature space into a perceptual voice quality similarity space. Experiments showed that the technique improves speaker selection performance; in particular, it can significantly reduce the speaker selection error rates.

6.2 Future work

In the MRHMM-based ASR experiment we conducted, we used acted or simulated style speech due to the limitations of a speech corpus, but this is not a realistic situation. Therefore, we will explore the effectiveness of the technique we proposed by using more realistic speech data, such as spontaneous speech, and also develop a technique that would be effective for unknown emotions.

For the technique we proposed for model training using the speaker class context, we will investigate other acoustic features and other speaker clustering techniques to improve the technique's speech synthesis performance. Although we used only adult females for average voice model training and speaker adaptation, we will apply the technique to other types of speakers such as adult males, children, and elders. We will also investigate applying the technique to style adaptation [54].

For the similar speaker selection technique we proposed, we will investigate other distance metric learning techniques, other speaker classes, and other utterance vectors to improve the technique's speaker selection performance. Although we have selected acoustic features using speaker selection by regression analysis, a unified approach to feature selection (i.e., [55]) will also be performed to select acoustic features matching our selection method based on distance metric learning. In addition, since we took into account voice quality for selecting similar speakers, we will investigate a similar speaker selection technique that can take into account voice quality as well as prosodic features. Finally, we will investigate applying the technique to speech synthesis.

Bibliography

[1] T. Kawahara, "Transcription system using automatic speech recognition for the japanese parliament (diet)," IAAI, 2012.

[2] Y.Y. Wang, D. Yu, Y.C. Ju, and A. Acero, "An introduction to voice search," IEEE Signal Processing Magazine, vol.25, no.3, pp.28–38, 2008.

[3] "Apple - iOS 7 - Siri." `http://www.apple.com/ios/siri/`.

[4] S. Matsuda, X. Hu, Y. Shiga, H. Kashioka, C. Hori, K. Yasuda, H. Okuma, M. Uchiyama, E. Sumita, H. Kawai, *et al.*, "Multilingual speech-to-speech translation system: Voicetra," IEEE 14th International Conference on Mobile Data Management, pp.229–233, 2013.

[5] C.J. Leggetter and P. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models," Computer Speech & Language, vol.9, no.2, pp.171–185, 1995.

[6] E. Eide and H. Gish, "A parametric approach to vocal tract length normalization," ICASSP-96, pp.346–348, May 1996.

[7] R. Kuhn, J.C. Junqua, P. Nguyen, and N. Niedzielski, "Rapid speaker adaptation in eigenvoice space," IEEE Trans. Speech and Audio Process., vol.8, no.6, pp.695–707, 2000.

[8] K. Fujinaga, M. Nakai, H. Shimodaira, and S. Sagayama, "Multiple-regression hidden Markov model," ICASSP 2001, pp.513–516, 2001.

[9] A.J. Hunt and A.W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," ICASSP-96, pp.373–376, 1996.

[10] 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(5), pp.825–834, May 2007.

[11] 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(2), pp.533–543, Feb. 2007.

[12] J. Yamagishi, O. Watts, S. King, and B. Usabaev, "Roles of the average voice in speaker-adaptive HMM-based speech synthesis," INTERSPEECH 2010, pp.418–421, Sept. 2010.

[13] T. Nose, Y. Kato, and T. Kobayashi, "Style estimation of speech based on multiple regression hidden semi-Markov model," INTERSPEECH 2007, pp.2285–2288, Aug. 2007.

[14] K. Miyanaga, T. Masuko, and T. Kobayashi, "A style control technique for HMM-based speech synthesis," INTERSPEECH 2004-ICSLP, pp.1437–1440, Oct. 2004.

[15] L. Yang, "An overview of distance metric learning." http://www.cs.cmu.edu/ liuy/dist_overview.pdf, 2007.

[16] Y. Ijima, M. Tachibana, T. Nose, and T. Kobayashi, "An on-line adaptation technique for emotional speech recognition using style estimation with multiple-regression HMM," INTERSPEECH 2008, pp.1297–1300, Sept. 2008.

[17] E. Eide and H. Gish, "A parametric approach to vocal tract length normalization," ICASSP-96, pp.346–348, 1996.

[18] J. Yamagishi, M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi, "A context clustering technique for average voice models," IEICE Trans. Inf. and Syst., vol.E86-D(3), pp.534–542, 2003.

[19] 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. and Syst., vol.E88-D(3), pp.502–509, 2005.

[20] J.A. Russell, "A circumplex model of affect," Journal of personality and social psychology, vol.39, no.6, pp.1161–1178, 1980.

[21] Y. Ijima, M. Tachibana, T. Nose, and T. Kobayashi, "A rapid model adaptation technique for emotional speech recognition with style estimation based on multiple-regression HMM," IEICE Trans. Inf. and Syst., vol.E93-D(1), pp.107–115, Jan. 2010.

[22] M. Tachibana, S. Izawa, T. Nose, and T. Kobayashi, "Speaker and style adaptation using average voice model for style control in HMM-based speech synthesis," ICASSP 2008, pp.4633–4636, April 2008.

[23] T. Nose, Y. Kato, M. Tachibana, and T. Kobayashi, "An estimation technique of style expressiveness for emotional speech using model adaptation based on multiple-regression HMM," INTERSPEECH 2008, pp.2759–2762, Sept. 2008.

[24] "JNAS: Japanese newspaper article sentences." `http://www.milab.is.tsukuba.ac.jp/instruct.html`.

[25] K. Shinoda and T. Watanabe, "MDL-based context-dependent subword modeling for speech recognition," J. Acoust. Soc. Jpn. (E), vol.21, no.2, pp.294–300, March 2000.

[26] V. Digalakis and L. Neumeyer, "Speaker adaptation using combined transformation and Bayesian methods," IEEE Trans. Speech and Audio Process., vol.4, pp.294–300, July 1996.

[27] "The hidden markov model toolkit (HTK)." `http://htk.eng.cam.ac.uk/`.

[28] "Open-source large vocabulary CSR engine julius." `http://julius.sourceforge.jp/`.

[29] K. Shikano, K. Ito, T. Kawahara, K. Takeda, and M. Yamamoto, IT text: Speech recognition system (accompanying CD-ROM), Ohmsha, 2001.

[30] Y. Ijima, N. Miyazaki, and H. Mizuno, "Statistical model training technique for speech synthesis based on speaker class," 8th ISCA Workshop on Speech Synthesis, Barcelona, Spain, pp.161–165, August 2013.

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

[32] 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., vol.17(6), pp.1208–1230, 2009.

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

[34] K. Amino, T. Sugawara, and T. Arai, "Speaker similarity in human perception and their spectral properties," WESPAC IX, 2006.

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

[36] 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, vol.27, pp.187–207, 1999.

[37] 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., vol.17(1), pp.66–83, 2009.

[38] Y. Ijima, M. Isogai, and H. Mizuno, "Correlation analysis of acoustic features with perceptual voice quality similarity for similar speaker selection," INTERSPEECH 2011, pp.2237–2240, Aug. 2011.

[39] Y. Ijima, M. Isogai, and H. Mizuno, "Similar speaker selection technique based on distance metric learning with perceptual voice quality similarity," INTERSPEECH 2012, Sept. 2012.

[40] Y. Ijima and H. Mizuno, "Similar speaker selection technique based on distance metric learning using highly correlated acoustic features with perceptual voice quality similarity," IEICE Trans. Inf. and Syst., vol.E98-D(1), pp.157–165, Jan. 2015.

[41] D.A. Reynolds, "Speaker identification and verification using gaussian mixture speaker models," Speech Communication, vol.17(1-2), pp.91–108, Aug. 1995.

[42] H. Chang and D.Y. Yeung, "Kernel-based distance metric learning for content-based image retrieval," Image Vision Comput., vol.25, no.5, pp.695–703, May 2007.

[43] M. Slaney, K. Weinberger, and W. White, "Learning a metric for music similarity," ISMIR 2008, pp.313–316, Sept. 2008.

[44] D. Mochihashi, G. Kikui, and K. Kita, "Learning an optimal distance metric in a linguistic vector space," Systems and Computers in Japan, pp.12–21, 2006.

[45] NTT-AT, "Japanese speech database (in Japanese)." http://www.ntt-at.co.jp/product/denwa_j.

[46] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," The Journal of the Acoustic Society of America, vol.87, pp.1738–1752, 1990.

[47] N. Minematsu, K. Tsuda, and K. Hirose, "Quantitative analysis of F0-induced variations of cepstrum coefficients," ISCA Tutorial and Research Workshop on Prosody in Speech Recognition and Understanding, 2001.

[48] T. Cover and P. Hart, "Nearest neighbor pattern classification," IEEE Trans. Information Theory, vol.13, no.1, pp.21–27, 1967.

[49] N.S. A. Bar-Hillel, T. Hertz and D. Weinshall, "Learning distance functions using equivalence relations," ICML 2003, pp.11–18, 2003.

[50] J. Goldberger, S. Roweis, G. Hinton, and R. Salakhutdinov, "Neighbourhood components analysis," NIPS, 2005.

[51] K. Weinberger, J. Blitzer, and L. Saul, "Distance metric learning for large margin nearest neighbor classification," NIPS, 2006.

[52] W.M. Campbell, D.E. Sturim, and D.A. Reynolds, "Support vector machines using GMM supervectors for speaker verification," IEEE Signal Processing Letters, vol.13, no.5, pp.308–311, May 2006.

[53] N. Dehak, R. Dehak, P. Kenny, N. Brummer, P. Ouellet, and P. Dumouchel, "Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification," INTERSPEECH 2009, pp.1559–1562, Sept. 2009.

[54] 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(3), pp.1092–1099, 2006.

[55] H. Liu and H. Motoda, Feature selection for knowledge discovery and data mining, Springer, 1998.

List of Publications

Publications Related to This Thesis

Journal

1. Yusuke Ijima, Hideyuki Mizuno,
   "Similar speaker selection technique based on distance metric learning using highly correlated acoustic features with perceptual voice quality similarity,"
   IEICE Trans. on Information and Systems, vol.E98-D, 1, pp.157–165, Jan. 2015.

2. Yusuke Ijima, Takashi Nose, Makoto Tachibana, Takao Kobayashi,
   "A rapid model adaptation technique for emotional speech recognition with style estimation based on multiple-regression HMM,"
   IEICE Trans. on Information and Systems, vol.E93-D, 1, pp.107–115, Jan. 2010.

International Conference

1. Yusuke Ijima, Noboru Miyazaki, Hideyuki Mizuno,
   "Statistical model training technique for speech synthesis based on speaker class,"
   Proc. SSW8, pp.141–145, Barcelona, Spain, Aug. 2013.

2. Yusuke Ijima, Mitsuaki Isogai, Hideyuki Mizuno,
   "Similar speaker selection technique based on distance metric learning with perceptual voice quality similarity,"
   Proc. INTERSPEECH 2012, Portland, U.S.A., Sept. 2012.

3. Yusuke Ijima, Mitsuaki Isogai, Hideyuki Mizuno,
   "Correlation analysis of acoustic features with perceptual voice quality similarity for similar speaker selection,"
   Proc. INTERSPEECH 2011, pp.2237–2240, Florence, Italy, Aug. 2011.

4. Yusuke Ijima, Takeshi Matsubara, Takashi Nose, Takao Kobayashi,
   "Speaking style adaptation for spontaneous speech recognition using multiple-regression HMM,"
   Proc. INTERSPEECH 2009, pp.552–555, Brighton, U.K., Sept. 2009.

5. Yusuke Ijima, Makoto Tachibana, Takashi Nose, Takao Kobayashi,
   "Emotional speech recognition based on style estimation and adaptation with multiple-regression HMM,"
   Proc. ICASSP 2009, pp.4157–4160, Taipei, Taiwan, April 2009.

6. Yusuke Ijima, Makoto Tachibana, Takashi Nose, Takao Kobayashi,
   "An on-line adaptation technique for emotional speech recognition using style estimation with multiple-regression HMM,"
   Proc. INTERSPEECH 2008, pp.1297–1300, Brisbane, Australia, Sept. 2008.

IEICE Technical Report

1. Yusuke Ijima, Makoto Tachibana, Takashi Nose, Takao Kobayashi,
   "Acoustic Model Training Technique for Speech Recognition Using Style Estimation with Multiple-Regression HMM,"
   IEICE Technical Report, vol.108, no.338, SP2008-85, pp.37–42, Dec. 2008 (in Japanese).

2. Yusuke Ijima, Makoto Tachibana, Takashi Nose, Takao Kobayashi,
   "An On-line Acoustic Model Adaptation Technique Based on Style Estimation,"
   IEICE Technical Report, vol.108, no.142, SP2008-48, pp.31–36, July 2008 (in Japanese).

ASJ Meeting

1. Yusuke Ijima, Noboru Miyazaki, Hideyuki Mizuno,
   "Performance evaluation of statistical model training technique using speaker class context,"
   ASJ Spring meeting, 1-R5-12, pp.403–404, Mar. 2014 (in Japanese).

2. Yusuke Ijima, Noboru Miyazaki, Hideyuki Mizuno,
   "Statistical model training technique using speaker class context,"
   ASJ Autumn meeting, 1-P-17a, pp.343–344, Sept. 2013 (in Japanese).

3. Yusuke Ijima, Mitsuaki Isogai, Hideyuki Mizuno,
   "Similar speaker selection method based on distance metric learning,"
   ASJ Spring meeting, 2-11-2, pp.337–338, Mar. 2012 (in Japanese).

4. Yusuke Ijima, Mitsuaki Isogai, Hideyuki Mizuno,
   "Correlation analysis of acoustic features with perceptual voice quality similarity,"
   ASJ Autumn meeting, 3-Q-13, pp.383–384, Sept. 2011 (in Japanese).

5. Yusuke Ijima, Mitsuaki Isogai, Hideyuki Mizuno,
   "Analysis of acoustic features affecting perception of voice quality similarity,"
   ASJ Spring meeting, 1-7-8, pp.273–274, Mar. 2011 (in Japanese).

6. Yusuke Ijima, Makoto Tachibana, Takashi Nose, Takao Kobayashi,
   "Perfomance evaluation of acoustic model training technique for speech recognition using style estimation,"
   ASJ Spring meeting, 1-P-27, pp.187–188, Mar. 2009 (in Japanese).

7. Yusuke Ijima, Makoto Tachibana, Takashi Nose, Takao Kobayashi,
   "Performance evaluation of on-line acoustic model adaptation technique based on style estimation,"
   ASJ Autumn meeting, 2-P-10, pp.131–132, Sept. 2008 (in Japanese).

Other Publications

Journal

1. Hiroko Muto, Yusuke Ijima, Noboru Miyazaki, Hideyuki Mizuno, Sumitaka Sakauchi,
   "Pause insertion prediction using evaluation model of perceptual pause insertion naturalness,"
   Journal of Information Processing, Vol. 56, No. 3, Mar. 2015 (in Japanese).

2. Yu Maeno, Takashi Nose, Takao Kobayashi, Tomoki Koriyama, Yusuke Ijima, Hideharu Nakajima, Hideyuki Mizuno, Osamu Yoshioka.
   "Prosodic Variation Enhancement Using Unsupervised Context Labeling for HMM-based Expressive Speech Synthesis,"
   Speech Communication, Elsevier, Vol. 57, No. 3, pp. 144–154, Feb. 2014.

International Conference

1. Hiroko Muto, Yusuke Ijima, Noboru Miyazaki, Hideyuki Mizuno,
   "Pause insertion prediction using evaluation model of perceptual pause insertion naturalness,"
   Proc. Speech Prosody 2014, pp.558–562, Dublin, Ireland, May 2014.

2. Yu Maeno, Takashi Nose, Takao Kobayashi, Tomoki Koriyama, Yusuke Ijima, Hideharu Nakajima, Hideyuki Mizuno, Osamu Yoshioka,
   "HMM-based expressive speech synthesis based on phrase-level F0 context labeling,"
   Proc. ICASSP 2013, pp.7859–7863, Vancouver, Canada, May 2013.

3. Yu Maeno, Takashi Nose, Takao Kobayashi, Yusuke Ijima, Hideharu Nakajima, Hideyuki Mizuno, Osamu Yoshioka,
   "HMM-based emphatic speech synthesis using unsupervised context labeling,"
   Proc. INTERSPEECH 2011, pp.1849–1852, Florence, Italy, Aug. 2011.

IEICE Technical Report

1. Yu Maeno, Takashi Nose, Takao Kobayashi, Tomoki Koriyama, Yusuke Ijima, Hideharu Nakajima, Hideyuki Mizuno, Osamu Yoshioka,
   "A Study on Multi-Class Local Prosodic Context for Expressive Prosody Generation,"
   IEICE Technical Report,vol.112, no.422, SP2012-112, pp.85–90, Jan. 2013 (in Japanese).

2. Hosana Kamiyama, Yusuke Ijima, Mitsuaki Isogai, Hideyuki Mizuno,
   "Analysis of the correlation between various acoustic features and the audibility of speech with noise,"
   IEICE Technical Report, vol.112, no.81, SP2012-46, pp.69–74, June 2012 (in Japanese).

3. Yu Maeno, Takashi Nose, Takao Kobayashi, Yusuke Ijima, Hideharu Nakajima, Hideyuki Mizuno, Osamu Yoshioka,
   "A Study on Automatic Prosodic Context Labeling for Emphatic Speech Synthesis,"
   IEICE Technical Report, vol.112, no.81, SP2012-33, pp.1–6, June 2012 (in Japanese).

4. Takashi Nose, Takeshi Matsubara, Yusuke Ijima, Takao Kobayashi,
   "Speaking Style Classification of Spontaneous Speech Using Multiple-Regression HMM,"
   IEICE Technical Report, vol.109, no.139, SP2009-46, pp.31–36, July 2009 (in Japanese).

ASJ Meeting

1. Tadashi Inai, Sunao Hara, Masanobu Abe, Yusuke Ijima, Noboru Miyazaki, Hideyuki Mizuno,
   "Quality improvement of HMM-based speech synthesis utilizing phone-sized spectrum in high frequency band,"
   ASJ Spring meeting, 2-Q-36, Mar. 2015 (in Japanese).

2. Takuma Inoue, Sunao Hara, Masanobu Abe, Yusuke Ijima, Hideyuki Mizuno,
   "Evaluation of HMM-based speech synthesis using high-frequency component of speech waveform,"
   ASJ Spring meeting, 3-6-13, pp.349–352, Mar. 2014 (in Japanese).

3. Takuma Inoue, Sunao Hara, Masanobu Abe, Yusuke Ijima, Hideyuki Mizuno,
   "Quality improvement of HMM-based speech synthesis using high-frequency component of speech waveform,"
   ASJ Autumn meeting, 1-P-21a, pp.347–350, Sept. 2013..

4. Hiroko Muto, Yusuke Ijima, Noboru Miyazaki, Hideyuki Mizuno,
   "Analysis of ease of comprehension with pause insertion and pause duration,"
   ASJ Autumn meeting, 1-7-5, pp.263–264, Sept. 2013 (in Japanese).

5. Yusuke Ijima, Hideyuki Mizuno,
   "Statistical speech synthesis method generating high-quality speech with the database size,"
   ASJ Spring meeting, 3-P-10b, pp.477–478, Mar. 2013 (in Japanese).

6. Yusuke Ijima, Hideyuki Mizuno,
   "Speech synthesis technique based on similar speaker selection,"
   ASJ Spring meeting, 3-P-8d, pp.475–476, Mar. 2013 (in Japanese).

7. Takuma Inoue, Sunao Hara, Masanobu Abe, Yusuke Ijima, Hideyuki Mizuno,
   "Evaluation of mixture method by HMM-based speech synthesis and segment speech synthesis,"
   ASJ Spring meeting, 1-7-15, pp.297–298, Mar. 2013 (in Japanese).

8. Hosana Kamiyama, Yusuke Ijima, Mitsuaki Isogai, Hideyuki Mizuno,
   "Correlation analysis of acoustic features with speech clarity in noise,"
   ASJ Spring meeting, 1-10-5, pp.513–514, Mar. 2012 (in Japanese).

9. Yu Maeno, Takashi Nose, Takao Kobayashi, Yusuke Ijima, Hideharu Nakajima, Hideyuki Mizuno, Osamu Yoshioka,
   "A study on automatic context labeling of emphatic expression for expressive speech synthesis,"
   ASJ Autumn meeting, 3-8-4, pp.335–336, Sept. 2011 (in Japanese).

10. Yu Maeno, Takashi Nose, Takao Kobayashi, Yusuke Ijima, Hideharu Nakajima, Hideyuki Mizuno, Osamu Yoshioka,
    "A study on prosodic contextual factors for HMM-based speech synthesis with diverse speaking styles,"
    ASJ Spring meeting, 1-Q-28, pp.385–386, Mar. 2011 (in Japanese).

11. Takeshi Matsubara, Yusuke Ijima, Makoto Tachibana, Takashi Nose, Takao Kobayashi,
    "Speaking style classification of spontaneous speech based on style estimation,"
    ASJ Spring meeting, 1-P-28, pp.189–190, Mar. 2009 (in Japanese).

12. Yusuke Ijima, Naomitsu Ikeda, Tadashi Sakata, Yuichi Ueda, Akira Watanabe,
    "Correction for the center of gravity of formant trajectory on word recognition,"
    ASJ Spring meeting, 1-P-3, pp.129–130, Mar. 2007 (in Japanese).

Other Conference

1. Masato Sekine, Katsuhiko Ogawa, Narichika Nomoto, Yusuke Ijima, Osamu Yoshioka,
   "A design of naturally spoken dialogue by Speech Synthesis System,"
   IPSJ Interaction 2011, pp.355–358, Mar. 2011 (in Japanese).

2. Yusuke Ijima, Naomitsu Ikeda, Tadashi Sakata, Yuichi Ueda, Akira Watanabe,
   "A development of speech recognition system with speaker normalization,"
   JCEEE-Kyushu, 10-2P-07, p.417, Sept. 2005 (in Japanese).