論文 / 著書情報
Article / Book Information

題目(和文)	DNN音声合成の多様化と音声対話システムへの応用に関する研究
Title(English)	A Study on DNN-Based Speech Synthesis with Diverse Voices and Its Application to Spoken Dialogue Systems
著者(和文)	北条伸克
Author(English)	Nobukatsu Hojo
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11435号, 授与年月日:2020年3月26日, 学位の種別:課程博士, 審査員:小林 隆夫,山口 雅浩,奥村 学,杉野 暢彦,篠﨑 隆宏
Citation(English)	Degree:Doctor (Engineering), Conferring organization: Tokyo Institute of Technology, Report number:甲第11435号, Conferred date:2020/3/26, 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 DNN-Based Speech Synthesis with Diverse Voices and Its Application to Spoken Dialogue Systems

Nobukatsu Hojo

March 2020

Summary

This thesis provides a novel approach to text-to-speech (TTS) synthesis systems used for a spoken dialogue system. The research objective was to improve the naturalness of synthesized speech with respect to "how natural the system's intention is perceived via the synthesized speech". We call this measure for TTS "illocutionary act naturalness (IAN)" in this thesis.

In order to achieve this aim, first, a novel TTS method that enables DNN to produce diverse voices was investigated. In order to produce diverse voices with TTS, we investigated a technique that can improve quality of synthesized speech with a limited amount of training data for each speaker. This technique is necessary because it can reduce the total data cost. To achieve this aim, we propose DNN-based speech synthesis using speaker codes as a method to improve the performance of the conventional speaker dependent DNN-based method. In order to model speaker variation in the DNN, the augmented feature (speaker codes) is fed to the hidden layer(s) of the conventional DNN. Objective and subjective evaluation was conducted to compare the proposed method with conventional ones. The results of objective evaluation show that the proposed model outperforms the speaker-dependent DNNs and multi-speaker DNNs with shared hidden layers for both multi-speaker modeling and speaker adaptation tasks. The subjective evaluation results show that the proposed model can produce more natural-sounding speech than the conventional speaker dependent DNNs and HMM-based speaker adaptation method, particularly when using a small number of target speaker utterances.

Second, the proposed model is applied to TTS considering intentions and evaluated with respect to the IAN. In order to consider a system's intention, we proposed to utilize dialogue-act tags which are designed for non-task oriented

dialogue systems. In this work, we first created speech database with DA tags. The proposed method is based on that in Chapter 2, which uses DA codes instead of speaker codes. The proposed method was compared with the two HMM-based conventional methods (the style-mixed model and the model adaptation method). The results of the subjective evaluation reveal that the proposed method improves the IAN compared with the HMM-based conventional methods. In addition, a comparison of the subjective and objective evaluation results reveals that the improvement of the proposed method over the HMM-based style-mixed model method is due to the improved reproducibility of global prosodic characteristics. We also showed that the improvement of the proposed method over the HMM-based model adaptation method is due to the improved reproducibility of sentence-final tones. Further, the proposed method also shows a higher IAN compared with the conventional DNN-based speech synthesis that considers only linguistic information. Thus, this results show that IAN can be improved by considering DAs information. In the subjective evaluation experiments, the size of the test set was limited due to the cost for evaluation. Therefore, it was difficult to analyze efficacy and difficulty of the proposed method using utterances in various contexts.

Finally, the sentence-final tone label estimation is focused to precisely investigate the efficacy and difficulty of the approach to utilize DA for speech synthesis. Since objective evaluation is cost effective, it is possible to use all of the utterances in the speech corpus as the test set. Therefore, we can use utterances in various contexts for evaluation. The speech database constructed in Chapter 3 of this thesis is annotated with sentence-final tone labels. The proposed method utilizes DAs as well as conventional morpheme and parts-of-speech information derived from an utterance to estimate a sentence final tone label. The objective evaluation reveals that the proposed method improves Cohen's κ by 0.12. This result shows the effectiveness of considering DAs in the proposed method. Analysis results show that the proposed method is effective for expressing three types of intentions of the system; "self-question and consent", "interrogative by rising intonation" and "self-question/fillers". We also analyze incorrect estimation results by the proposed method. The results show that more detailed DA tags such as "self-disclosure with/without expressing a speaker's emotion" or "repeat with/without expressing surprise" are effective to improve estimation accuracy. It is considered effective to include these tags in the DA tag set and construct a spoken dialogue

system based on the tag set.

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. Also, I would like to thank Professor Masahiro Yamaguchi, Professor Manabu Okumura, Professor Nobuhiko Sugino and Professor Takahiro Shinozaki of Tokyo Institute of Technology for their kind suggestions.

In addition, I have benefited greatly from interaction with members of NTT Media Intelligence Laboratories and Communication Science Laboratories, Nippon Telegraph and Telephone Corporation. I must thank Dr. Yusuke Ijima, for his continuous advice and assistance on every research of this thesis. I would also like to thank Noboru Miyazaki, for his advice on the research subject and design of this thesis. I would also like to thank Dr. Hiroaki Sugiyama, for his great advice as an expert in the field of dialogue systems. I wish to acknowledge the help provided by Dr. Hideyuki Mizuno, particularly on the work in Chapter 2. I also wish to acknowledge the help provided by Dr. Takahito Kawanishi and Dr. Kunio Kashino, particularly on the work in Chapter 3. I would like to offer my special thanks to Dr. Hirokazu Kameoka, for providing me the great opportunity to work in this field. I would like to thank other lab members including Dr. Kou Tanaka and Takuhiro Kaneko.

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

The text-to-speech (TTS) synthesis technique is currently widely used in numerous applications such as car navigation systems and computer telephony integration systems (e.g., Interactive Voice Response; IVR). Among these applications, spoken dialogue systems (such as Siri and Alexa) have become rapidly popular in recent years. The subject of this thesis is to realize natural communication between a system and a user by tailoring TTS for spoken dialogue systems.

In human-to-human communication, a hearer infers the speaker's intention from every utterance [3], [4]. If the expression of the speaker's intention is unnatural, it increases the cognitive load for inference or causes inconsistency with context and misunderstanding. This results in the degradation of the quality of communication. For spoken dialogue systems as well, it is considered necessary to express the system's intention naturally in order to achieve high-quality communication. Since perceived intention differs depending on the prosodic characteristics (fundamental frequencies (F0) and phoneme durations (durations)) [5], [6], it is necessary for TTS to express intentions naturally by controlling prosodic characteristics.

In the speech synthesis field, the naturalness of the synthesized speech is often evaluated as the quality of speech samples separated from context [7], [8]. In other words, "how natural the system's intention is perceived via the synthesized speech" has not been investigated thus far. Based on the classification by the speech act

theory [9], [10], the former is rephrased as naturalness with respect to the act of to saying something (locutionary act), while the latter is rephrased as naturalness with respect to conveying intentions (illocutionary act). This thesis defines the former as "locutionary act naturalness (LAN)" and the latter as "illocutionary act naturalness (IAN)" and aims at improving the IAN of synthesized speech.

In order to improve IAN, it is considered necessary to reproduce the prosodic characteristics of the intentions of synthesized speech. Since the number of intentions can be large (e.g., 30 in this thesis), it is important to train a speech synthesis model for an intention from a small amount of data, as this reduces the cost for training. In the hidden Markov model (HMM)-based speech synthesis, numerous techniques have succeeded in generating speech from a small amount of data from the target speaker. A powerful method is an average-voice-based speech synthesis technique with model adaptation [11]. In this technique, average voice models are created from several speakers' speech data and are adapted to a small amount of speech data from a target speaker using model adaptation algorithms like constrained structural maximum a posteriori linear regression (CSMAPLR) [12].Another successful method is based on cluster adaptive training (CAT) [13]. This model has multiple compact decision trees that are interpolated to produce a huge variety of possible contexts; the model is trained using a multi-speaker speech corpus to improve speech quality.

Recent studies have also showed that deep neural network (DNN)-based speech synthesis [14]–[16] can produce more natural synthesized speech than conventional HMM-based speech synthesis. However, DNN-based speech synthesis requires a considerable amount of speech data spoken by the target speaker to achieve sufficient performance validity.The problem then becomes the high cost required to generate speech from various speakers from a DNN, as we need a considerable amount of speech data from all the speakers the system uses as well as annotations of phonetic and prosodic contextual information on them. If we could train a DNN using a small amount of speech data, this technique can be applied to intention expression.

1.2 Scope of thesis

This thesis describes DNN-based speech synthesis, considering intentions in order to improve the IAN of the synthesized speech generated by a spoken dialogue system. We first propose a novel multi-speaker modeling method based on DNNs. Then, we apply the technique for modeling intentions in order to improve the IAN. Thereafter, we present the sentence-final tone estimation method, whose main focus is to analyze the difficulty of the proposed scheme to use intention tags for speech synthesis.

In Chapter 2, we propose to use augmented features based on speaker codes as a relatively simple method for multi-speaker modeling and speaker adaptation for DNN-based speech synthesis. In order to model speaker variation in the DNN, the augmented feature (speaker codes) is fed to the hidden layer(s) of the conventional DNN. In Chapter 2, we investigate the effectiveness of introducing speaker codes to DNN acoustic models for speech synthesis to fulfill two tasks: multi-speaker modeling and speaker adaptation. For the multi-speaker modeling task, we propose a method that trains connection weights of the entire DNN using a multi-speaker speech corpus. When performing multi-speaker synthesis, the speaker code corresponding to the selected target speaker is fed to the DNN to generate the speaker's voice. When performing speaker adaptation, a set of connection weights of the multi-speaker model is re-estimated to generate a new target speaker's voice. We investigate the relationship between the prediction performance and architecture of the DNNs through objective measurements. The effectiveness of the proposed model architecture is investigated through objective and subjective evaluation.

In Chapter 3, the model architecture proposed in Chapter 2 is applied for modeling intentions and evaluated with respect to the IAN. In order to investigate this aspect, first, we construct a speech database with DA tags. Second, we build the proposed DNN-based speech synthesis system based on the database. Then, we evaluate the proposed method by comparing its performance with two conventional HMM-based speech synthesis systems—namely, the style-mixed modeling method and the style adaptation method. In this chapter, we also propose a method to evaluate the IAN of synthesized speech. Both objective and subjective evaluation reveal that the proposed method improves the IAN.

In Chapter 4, we focus on estimating appropriate sentence-final tone labels, as estimating them is considered essential for communicating the system's precise intentions to users through an utterance. In this chapter, we propose to utilize intention-related features as well as conventional features, including morphological information of the utterance text, to estimate sentence-final tone labels. We added sentence-final tone labels to the database constructed in Chapter 3, thereby ensuring that each utterance has all the information related to utterance text, DA, and sentence-final tone label. Based on this database, we build the proposed sentence-final tone estimation model. The objective evaluation results indicate that considering intentions are effective in estimating sentence-final tone labels.

Chapter 2

DNN-based speech synthesis using speaker codes

Deep neural network (DNN)-based speech synthesis can produce more natural-sounding synthesized speech than the conventional HMM-based speech synthesis. However, it is not revealed whether the synthesized speech quality can be improved by utilizing a multi-speaker speech corpus. To address this problem, this chapter proposes DNN-based speech synthesis using speaker codes as a method to improve the performance of the conventional speaker dependent DNN-based method. In order to model speaker variation in the DNN, the augmented feature (speaker codes) is fed to the hidden layer(s) of the conventional DNN. In this chapter we investigate the effectiveness of introducing speaker codes to DNN acoustic models for speech synthesis for two tasks: multi-speaker modeling and speaker adaptation. For the multi-speaker modeling task, the method we propose trains connection weights of the whole DNN using a multi-speaker speech corpus. When performing multi-speaker synthesis, the speaker code corresponding to the selected target speaker is fed to the DNN to generate the speaker's voice. When performing speaker adaptation, a set of connection weights of the multi-speaker model is re-estimated to generate a new target speaker's voice. We investigated the relationship between the prediction performance and architecture of the DNNs through objective measurements. Objective evaluation experiments revealed that the proposed model outperformed conventional methods (HMMs, speaker dependent DNNs and multi-speaker DNNs based on a shared hidden layer

structure). Subjective evaluation experimental results showed that the proposed model again outperformed the conventional methods (HMMs, speaker dependent DNNs), especially when a using small number of target speaker utterances.

2.1 Introduction

Recent studies have shown that deep neural network (DNN)-based speech synthesis [14]–[16] can produce more natural synthesized speech than the conventional hidden Markov model (HMM)-based speech synthesis. However, DNN-based speech synthesis requires a considerable amount of speech data uttered by the target speaker to obtain sufficient performance.

The problem then becomes the high cost required to generate speech from various speakers from a DNN because we need a considerable amount of speech data from all the speakers the system uses, and annotations of phonetic and prosodic contextual information on them.

In the field of HMM-based speech synthesis, many techniques have succeeded in generating speech from a smaller amount of target speaker data. A powerful method is an average-voice-based speech synthesis technique with model adaptation [11]. In this technique, average voice models are created from several speakers' speech data and are adapted to a small amount of speech data from a target speaker using model adaptation algorithms such as CSMAPLR [12]. Another successful method is based on cluster adaptive training (CAT) [13]. This model has multiple compact decision trees that are interpolated to produce a huge variety of possible contexts, and is trained using a multi-speaker speech corpus to improve the speech quality.

Motivated by these previous studies in HMM-based speech synthesis, we aimed in our work to improve the synthetic speech quality from a DNN by using a multi-speaker speech corpus. One possible approach to train a DNN using a multi-speaker speech corpus is adopting the shared hidden layer structure [1]. The model structure is composed of the same hidden layers shared among different speakers and the output layers composed of speaker-dependent nodes. It has been shown that this model structure can improve synthesized speech quality for both multi-speaker modeling and speaker adaptation tasks. Another approach to train

a DNN using a multi-speaker speech corpus is to feed augmented speaker specific features to the network in order to incorporate speaker-level information to the DNNs. In the speech recognition, for example, this approach has been shown to be an effective way to feed i-vectors [17] or speaker codes [18], [19] as augmented features. As for DNN-based speech synthesis, a recent study [20] conducted an experimental analysis using i-vector based feature augmentation. This work is based on the assumption that speaker code-based features can also be effectively introduced to DNNs for speech synthesis.

In this chapter we propose to use augmented features based on speaker codes as a relatively simple method for multi-speaker modeling and speaker adaptation for DNN-based speech synthesis. Initially we proposed using speaker codes for DNN-based speech synthesis [21], and it has been shown that this approach can achieve better synthetic speech quality than can be obtained with conventional speaker dependent DNNs and speaker adaptation using HMMs. Luong et al. [22] also proposed using the features of speaker codes for DNN-based speech synthesis. Their main focus is to control the speech characteristics by modifying the augmented features and it has been shown that such modification can be performed effectively by manipulating the input code vectors. However, it still has not been revealed which approach, the augmented feature-based approach or the shared hidden layer approach, is better able to model multi-speaker characteristics for DNN-based speech synthesis. Adding to the contribution of our previous study [21], we describe in this chapter how we conducted an objective evaluation experiment to compare the performance of the proposed speaker code-based approach with that of the conventional shared hidden layer approach.

2.2 Conventional DNN-based speech synthesis

2.2.1 Conventional Speaker dependent DNN

The baseline model is a DNN acoustic model similar to the one described by Zen et al. [14]. The model is illustrated in Fig. 2.1. The DNN is used as a function to map the linguistic feature vectors to acoustic feature vectors. First, the input text is converted to the linguistic feature vector. The input features include binary answers to questions about linguistic contexts and numeric values. Then the

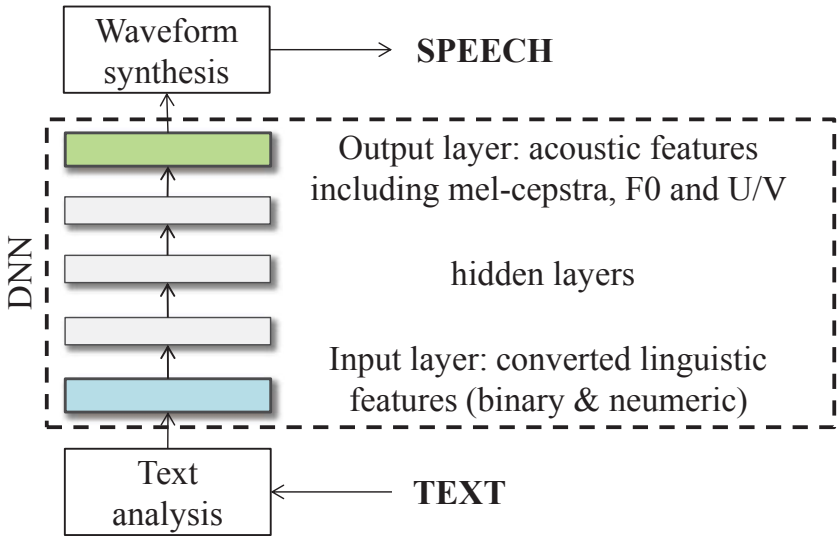


Figure 2.1: The model architecture of the conventional speaker dependent DNN.

linguistic feature vector is mapped to the output feature by forward propagation of DNN. The output features include spectral and excitation parameters and their time derivatives. The baseline model is trained using only the target speaker utterances in the training corpus.

2.3 Multi-speaker modeling using speaker codes

There are two possible model architectures to introduce speaker codes into the DNN-based speech synthesis. One is similar to the architecture proposed in the speech recognition field [18], [19], which is illustrated in Fig. 2.2. We call this architecture "embedded speaker code" in this chapter. The other is a simplified version of the embedded speaker code model illustrated in Fig. 2.3 . This model architecture does not embed the speaker ID vector and directly feeds it to hidden layers. We call this architecture "one-hot speaker code" in this chapter.

2.3.1 Embedded speaker code-based speech synthesis

As shown in Fig. 2.2, in this model architecture, a speaker ID is embedded to a speaker code $\mathcal{S}$ through a set of connection weights. Here the speaker code $\mathcal{S}$ represents the speaker information. The speaker codes are fed to all hidden layers. The speaker ID vector $\boldsymbol{v} = [v_1, \cdots, v_K]^{\mathrm{T}}$ for speaker m is set to the following fixed

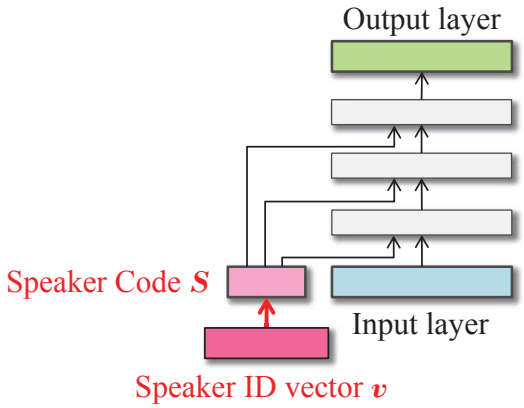


Figure 2.2: The model architecture of the embedded speaker code-based speech synthesis. The red bold lines indicate the re-estimated parameters for speaker adaptation.

1-of-K form:

$$v_k = \begin{cases} 1 & (k = m) \\ 0 & (k \neq m) \end{cases} \tag{2.1}$$

where K is the number of speakers in the training data.

For the formulation below, the connection weight from the speaker code $\mathcal{S}$ to the k-th hidden layer is represented by $W_{\mathcal{S}}^k$. The vector fed by speaker code $\mathcal{S}$ to the k-th hidden layer is represented by $\boldsymbol{y}^k$.

For embedded speaker codes, we can get

$$\begin{aligned} \boldsymbol{y}^k &= W_{\mathcal{S}}^k \mathcal{S} \\ &= W_{\mathcal{S}}^k W_E \boldsymbol{v}, \end{aligned} \tag{2.2}$$

where $\boldsymbol{v}$ is a one-hot vector that represents the speaker ID in (2.1) and W_E is the embedding connection weight from the speaker ID vector $\boldsymbol{v}$ to speaker codes $\mathcal{S}$. The size of each vector or matrix is:

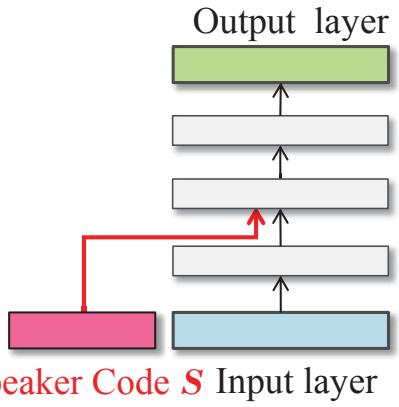
$$\boldsymbol{y}^k : D_h^k \tag{2.3}$$

$$W_{\mathcal{S}} : D_h^k \times D_{\mathcal{S}} \tag{2.4}$$

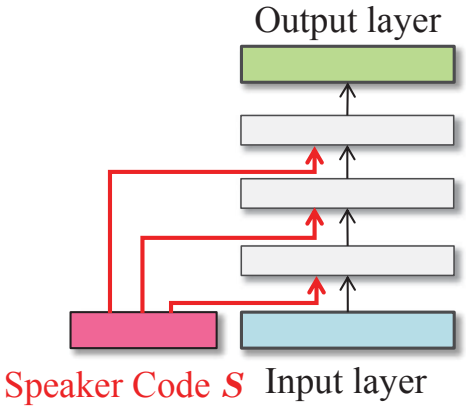
$$W_E : D_{\mathcal{S}} \times K \tag{2.5}$$

$$\boldsymbol{v} : K, \tag{2.6}$$

where D_h^k is the unit size of k-th hidden layer and $D_{\mathcal{S}}$ is a dimension of the speaker code $\mathcal{S}$.



(a) One-hot speaker code (connected to a single hidden layer)



(b) One-hot speaker code (connected to all hidden layers)

Figure 2.3: The model architecture of the one-hot speaker code-based speech synthesis. The red bold lines indicate the re-estimated parameters for speaker adaptation.

By setting $D_{\mathcal{S}} < D_h^k$ and $D_{\mathcal{S}} < K$, we can constrain the matrix $W_{\mathcal{S}}^k W_E$ to be low rank (=$D_{\mathcal{S}}$) and $\boldsymbol{y}^k$ to be in $D_{\mathcal{S}}$-dimensional linear subspace. This constraint is expected to make the model robust even when the amount of training data for a speaker is limited.

2.3.2 One-hot vector speaker code-based speech synthesis

In this model architecture, the speaker code $\mathcal{S} = [\mathcal{S}_1, \mathcal{S}_2, \cdots, \mathcal{S}_K]^{\mathrm{T}}$ for speaker m is set to the same formulation as the one-hot speaker ID vector $\boldsymbol{v}$ in eq. (2.1).

$$\mathcal{S}_k = \begin{cases} 1 & (k = m) \\ 0 & (k \neq m) \end{cases} \tag{2.7}$$

As shown in Fig.2.3, a speaker code $\mathcal{S}$ is fed through an additional set of connection weights to a certain hidden layer (Fig. 2.3a) or all hidden layers (Fig. 2.3b).

For one-hot vector speaker codes, we can get

$$\boldsymbol{y}^k = W_{\mathcal{S}}^k \mathcal{S}, \tag{2.8}$$

where $\mathcal{S}$ is a one-hot vector that represents the speaker ID in (2.7). The size of each vector or matrix is:

$$\boldsymbol{y}^k : D_h^k \tag{2.9}$$

$$W_{\mathcal{S}}^k : D_h^k \times K \tag{2.10}$$

$$\mathcal{S} : K. \tag{2.11}$$

It can be seen that $W_{\mathcal{S}}^k$ is full rank and $\boldsymbol{y}^k$ has no constraint, unlike the embedded speaker code model in (2.2). This means that one-hot vector speaker code models may model speaker characteristics precisely using larger number of parameters while they can be less robust for limited training data when compared with embedded speaker code models.

2.3.3 Training procedure for multi-speaker modeling using speaker codes

For multi-speaker modeling based on embedded speaker codes and one-hot speaker codes, the proposed methods train connection weights of the whole DNN using a multi-speaker speech corpus. When synthesizing a speech parameter sequence, a target speaker is chosen from the corpus and the speaker code corresponding to the selected target speaker is fed to the DNN to generate the speaker's voice. The proposed model is expected to generate more stable and natural speech compared with conventional speaker dependent DNNs because the networks from the linguistic feature to the acoustic feature are trained with a greater variety of contextual information by using a multi-speaker speech corpus.

2.4 Speaker adaptation using speaker codes

The multi-speaker modeling approach will need high computational cost every time we want to generate speech from a new target speaker. This is because the whole

model needs to be retrained using a corpus including the new speaker's utterances. It is considered that model adaptation by re-estimating a subset of model parameters using the target speaker utterances can address this problem. Following up on the previous research work that has been done in the field of speaker adaptation for speech synthesis [20], [23]–[25], we consider that the speaker code based DNN can also be used as a speaker adaptation method, since this approach has been shown to be effective in speech recognition [18], [19]. The set of parameters re-estimated for adaptation is set differently than that for the model architectures.

2.4.1 Speaker adaptation using embedded speaker codes

The adaptation procedure described in this section is similar to those in [18], [19]. First, the connection weights of the whole DNN are trained using a multi-speaker speech corpus. The speaker ID vector $\boldsymbol{v}$ is set in a form similar to that given in Sect. 2.3.1, but this time $\boldsymbol{v}$ is appended by an additional dimension to have $K + 1$ dimensions in total. This additional dimension is used to represent unseen target speaker information, and always set to 0 in the training procedure. Second, the model is adapted to a new target speaker using only the target speaker utterances as adaptation data. This time the additional dimension of $\boldsymbol{v}$ is set to 1 and other dimensions to 0, and only the connection weights for embedding are re-estimated to minimize the distance between the output features of the adaptation data and predicted values. When synthesizing, a one-hot vector $\boldsymbol{v}$ whose additional dimension is set to 1 is used. This adaptation procedure corresponds to updating only the $K + 1$-th column of W_E in (2.2).

The adaptation algorithm amounts searching a new point of $\boldsymbol{y}^k$ to model a new target speaker. Here, by setting $D_{\mathcal{S}} < D_h^k$ we can constrain the search space of $\boldsymbol{y}^k$ to be in the $D_{\mathcal{S}}$-th linear subspace. This constraint decreases the number of free parameters for adaptation. Therefore, it is expected to make the model robust even when the amount of adaptation data is limited. The adaptation procedure for embedded speaker codes described above was originally proposed in the speech recognition field [18], [19] in order to make an adapted model robust to a limited amount of adaptation data by reducing the number of adaptation parameters. Specifically, the number of adaptation parameters for embedded speaker codes is reduced to $D_{\mathcal{S}}$, while that in Sect. 2.4.2 is D_h. However, it is not clear whether

this reduction is effective for speech synthesis. In order to confirm this point, we conducted experiments to compare these two methods.

2.4.2 Speaker adaptation using one-hot speaker codes

While a DNN using one-hot speaker codes is adapted in the similar procedures similar to those of embedded speaker code models, this time the set of re-estimated parameters is different. First, the connection weights of the whole DNN are trained using a multi-speaker speech corpus. The speaker code $\mathcal{S}$ is set in a form similar to that of speaker ID vector $\boldsymbol{v}$ in the embedded speaker code. The vector $\mathcal{S}$ has $K+1$ dimensions in total. The additional dimension is always set to 0 in the training procedure and 1 in the adaptation and synthesis procedure. For the adaptation procedure, only the connection weights from the speaker codes $\mathcal{S}$ to the hidden layers are re-estimated to minimize the distance between the output features of the adaptation data and predicted values. This adaptation procedure corresponds to updating only the $K+1$-th column of $W_{\mathcal{S}}^{k}$ in (2.8). This time the number of parameters for adaptation is D_h^k.

This adaptation procedure is different from those reported in previous speech recognition studies [18], [19] and the procedure described in Sect. 2.4.1. These references for speaker adaptation for speech recognition mainly focus on fast and robust adaptation with a limited amount of adaptation data. On the other hand, for speech synthesis, it is necessary to model the speaker characteristics precisely to generate the target speaker's speech. The procedure for this model is expected to be flexible and more appropriate for speech synthesis, because it re-estimates a set of parameters whose dimensions are generally larger than those of $\mathcal{S}$ for adaptation.

Another possible adaptation procedure for one-hot speaker codes was proposed by Luong et al. [22]. They proposed to update only $\mathcal{S}$ in (2.8) for a new target speaker. This time the number of parameter for adaptation is K, which is usually smaller than the number for D_h^k. This constraint is expected to have effect similar to embedded speaker code models. We adopted the proposed adaptation procedure because it enabled us to see whether or not a large number of parameter for adaptation is needed by comparing the performance of one-hot and embedded speaker code models.

2.5 Experiments

As described in Sect. 2.3, there are various possible architectures for DNN-based speech synthesis using speaker codes. In this section, we will first explain how we investigated the relationship between the model architecture and the proposed method's performance through objective measurements. We determined the best architecture for the proposed method on the basis of the measurement results. We then conducted a subjective evaluation to confirm whether our method could improve the speech quality compared with the conventional speaker dependent DNN. We also compared the results with those obtained by HMMs as well as those by speaker dependent DNNs in these experiments. This is because prediction performance obtained by HMM is known to be superior to that obtained by speaker dependent DNN [14]. In Sect. 2.5.1, we present the experimental conditions we used. We describe the objective evaluation experiments in Sect. 2.5.2 and the subjective experiments in Sect. 2.5.3.

2.5.1 Experimental setup

In the experiments, we used speech data in Japanese obtained from 35 speakers (17 males and 18 females). Two speakers, one male and one female, were used as target speakers. The training corpus included 7,260 utterances (about 1,340 minutes) from 33 speakers apart from the two target speakers. We conducted objective evaluation for two tasks: multi-speaker modeling and speaker adaptation. For each task, we considered two training conditions: five utterances (about 1.2 minutes) and 300 utterances (about 63 minutes) for the training corpus for each target speaker. Twenty utterances were used as a testing set for each target speaker. The sampling rate of the corpus was 22.05 kHz. The STRAIGHT vocoder [26] was employed to extract 40 dimensional mel-cepstral coefficients, five band aperiodicities, and F0 in log-scale at 5 msec steps.

We compared the performance obtained with the following five acoustic models.

- HMM: The conventional HMM-based average voice model with adaptation.
- DNN_SD: The conventional speaker dependent model [14].

- SPKCODE_EMBED: The proposed method using embedded speaker codes described in Sect. 2.3.1.

- SPKCODE_ONE-HOT: The proposed method using one-hot speaker codes described in Sect. 2.3.2.

The input linguistic feature vector of a DNN contained 506 dimensional linguistic features. Each observation vector consisted of 40 mel-cepstral coefficients, log F0, five band aperiodicities, their delta and delta-delta features, and a voiced/unvoiced binary value. The input numeric features were normalized to the 0.01-0.99 range, and the output features were normalized by speaker-dependent mean and variance. The DNN systems had five hidden layers and each hidden layer had 1024 units. A sigmoid activation function was used in the hidden layers followed by a linear activation at the output layer. For the training procedure, the weights of the DNN were initialized randomly with Gaussian distribution with zero mean and variance $\frac{1}{\sqrt{D}}$, where D denotes the dimension of an input vectors. The initial values for biases of the DNN were set to zero. The weights and biases were then optimized to minimize the mean squared error between the output features of the training data and predicted values, using the Adam-based back-propagation algorithm [27]. The parameters for the Adam algorithm were set as $\alpha = 0.0001, \beta_1 = 0.9, \beta_2 = 0.999$ and $\epsilon = 1\mathrm{e} - 8$. Five percent of the utterances of the whole training data were used as a development set. The DNN_SD models were trained using only the target speaker utterances in the training corpus. We did not train DNN_SD models using five utterances because the amount of data in the training corpus was too small to train a model properly.

For the HMM, we used a five-state left-to-right hidden semi-Markov model with no skip topology. Each observation vector consisted of 138 features (40 mel-cepstral coefficients, log F0, five band aperiodicities, and their delta and delta-delta features). The output distribution in each state was modeled as a single Gaussian density function and the covariance matrices were assumed to be diagonal. For the multi-speaker training task, the average voice model was trained using a multi-speaker speech corpus including the target speaker utterances. Then the average voice model was adapted using the target speaker utterances again. For the speaker adaptation task, on the other hand, the target speaker utterances were excluded from the corpus for training the average voice model. The average

voice model was then adapted using the target speaker utterances. We used the combined technique of CSMAPLR and MAP adaptation as the speaker adaptation algorithm [12]. The model size was determined automatically by the minimum description length (MDL) criterion [28], where the control parameter of the model size was set to $\alpha = 1.0$.

For all of the four methods evaluated in these experiments, segmentations (phoneme durations) from natural speech were used instead of predicting duration. We applied MLPG [29] to the output features for all of the four methods. For DNN-based methods, we used pre-computed variances from the training data for MLPG. We did not apply spectral enhancement techniques such as global variance [30] to reduce factors considered in the experiments.

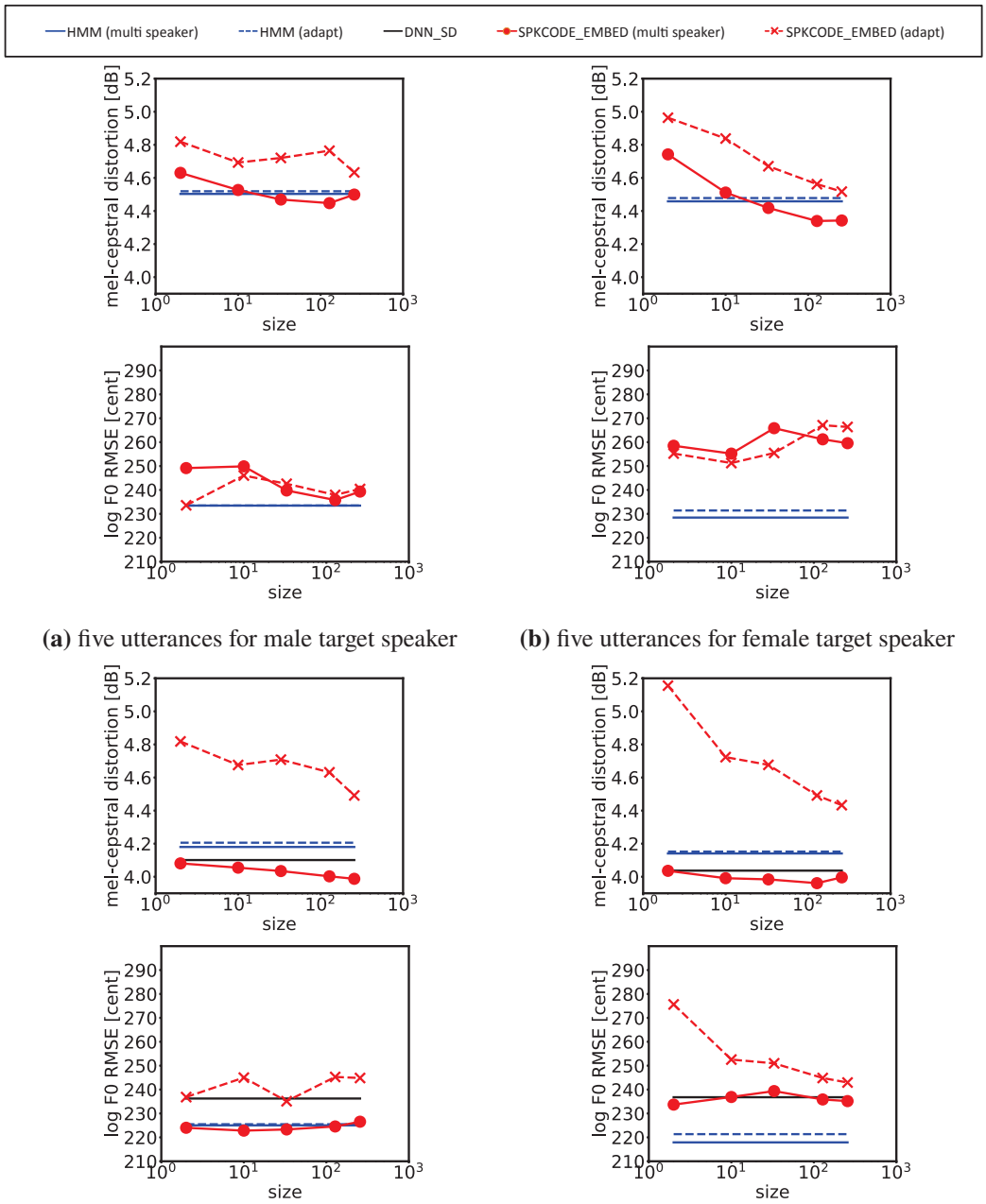
2.5.2 Objective evaluation

We first investigated the relationship between the prediction performance and the architectures of the proposed method. We changed the embedded speaker code dimension for SPKCODE_EMBED models. The embedding dimensions were set to 2, 10, 33, 128 and 256. For SPKCODE_ONE-HOT models, the input hidden layer(s) for speaker codes were changed. The input hidden layer(s) for SPKCODE_ONE-HOT models were set to the 1st, 2nd, 3rd, 4th, 5th hidden layer and all hidden layers. We used objective measurements to compare the performances we obtained in these cases with those obtained with conventional HMM and DNN_SD.

2.5.2.1 Objective evaluation for SPKCODE_EMBED

Figure 2.4 presents the mel-cepstral distortions (MCDs) and RMSEs of log F0 for SPKCODE_EMBED models. The horizontal axis indicates the size of the embedded speaker code vector. The results are presented with HMM and DNN_SD.

First, we compared the performance for the model structure of the SPKCODE_EMBED models. For both the conditions using five and 300 target speaker utterances, the experimental results showed the tendency that a larger embedding vector size achieves smaller MCDs. This tendency was common for both the multi-speaker modeling and speaker adaptation tasks and similar to that



(a) five utterances for male target speaker

(b) five utterances for female target speaker

(c) 300 utterances for male target speaker

(d) 300 utterances for female target speaker

Figure 2.4: The objective evaluation results for embedded speaker code model. The horizontal axis indicates the size (dimension) of the speaker codes.

reported by Luong et al. [22]. As for RMSEs of log F0, on the other hand, there was no consistent relationship between the size of the embedding vector and F0 RMSEs for both the multi-speaker modeling and speaker adaptation tasks. This

tendency contrasted with that reported in a previous study [22], in which it was stated that a larger embedding vector size effectively reduces RMSEs of log F0. We consider that this is because of the difference in the number of speakers in the training data. There were 33 speakers in our experiments, while there were 100 speakers in those conducted by Luong et al. [22]). Since the number of speakers was smaller in our experiments, the embedding speaker code vector space became sparse when the vector size was large, which led to over-fitting and diminished performance.

We then compared the performance of the conventional methods, i.e., HMM and DNN_SD. We found the relationship of

- DNN_SD < HMM
  (300 utterances, MCDs)
- HMM < DNN_SD
  (300 utterances, F0 RMSEs)

for both multi-speaker modeling and speaker adaptation tasks. The MCDs of DNN_SD were better than those of HMM because in modeling complex context dependencies DNN-based methods have an advantage over the tree-clustered HMM-based methods as Zen et al. described [14]. The F0 RMSEs of DNN_SD were higher than HMM, which is the same tendency as Zen et al. reported [14].

We then compared the performance of SPKCODE_EMBED with that obtained with these conventional methods. From the results of multi-speaker modeling tasks, we can see the relationship of

- SPKCODE_EMBED < HMM
  (5 utterances, MCDs)
- HMM < SPKCODE_EMBED
  (5 utterances, F0 RMSEs)
- SPKCODE_EMBED < DNN_SD < HMM (300 utterances, MCDs)
- HMM < SPKCODE_EMBED < DNN_SD (300 utterances, F0 RMSEs).

From these results for multi-speaker modeling, we confirm that the SPKCODE_EMBED models can slightly improve the parameter estimation accuracy compared with DNN_SD models by utilizing a multi-speaker speech corpus.

From the results obtained from the speaker adaptation task in Fig. 2.4, we found the relationship of

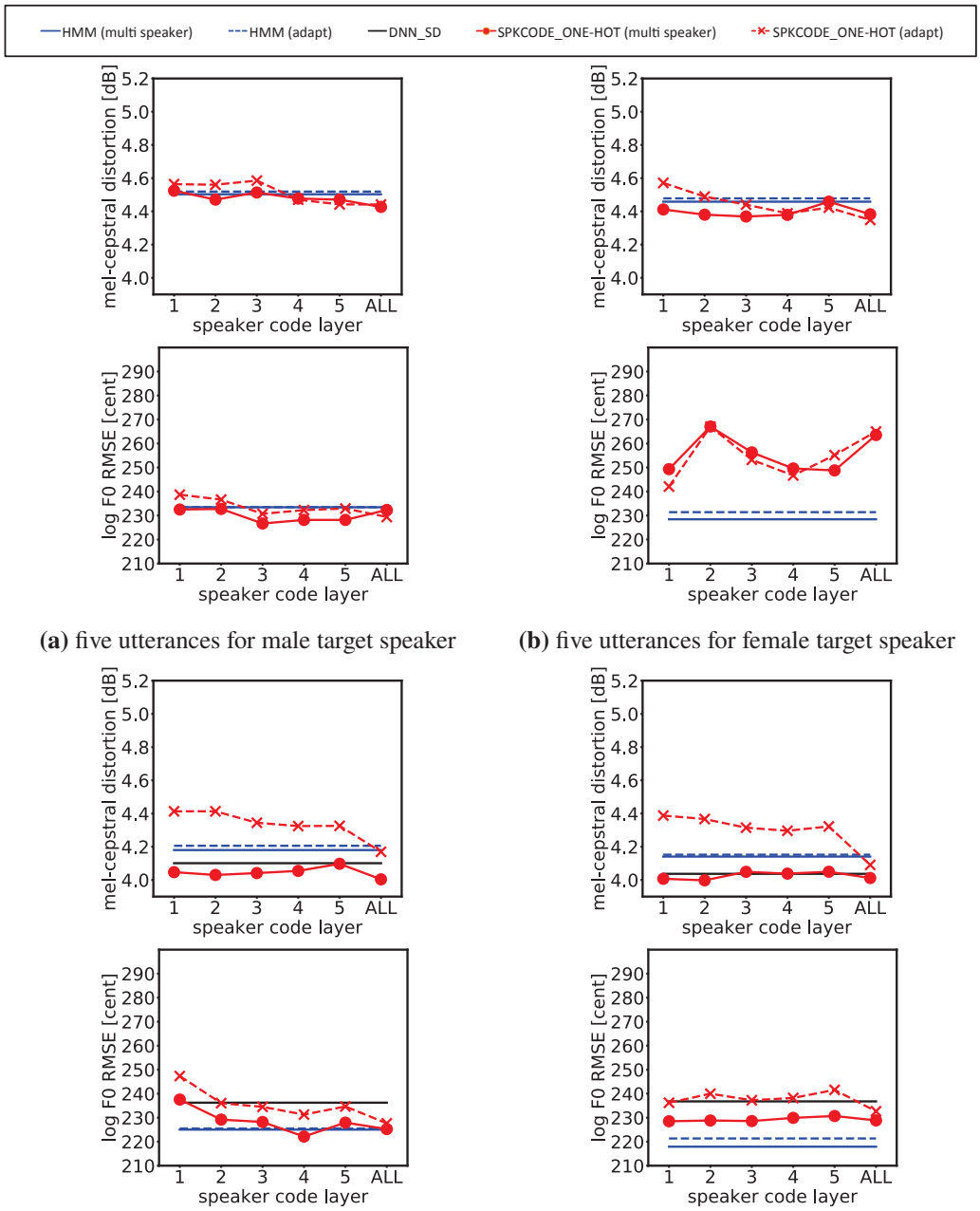
- HMM < SPKCODE_EMBED
  (five utterances, MCDs)
- HMM < SPKCODE_EMBED
  (five utterances, F0 RMSEs)
- DNN_SD < HMM < SPKCODE_EMBED (300 utterances, MCDs)
- HMM < DNN_SD < SPKCODE_EMBED (300 utterances, F0 RMSEs).

The performance of SPKCODE_EMBED was comparable to or worse than that obtained with the conventional HMM and DNN_SD. This is because the embedded speaker code model architecture has too small a number of free parameters for adaptation. The number of free parameters for adaptation is equal to the size of the embedded speaker codes in SPKCODE_EMBED (e.g., 2, 10, 33, 128 and 256 in our experiments). This too small a number of free parameters led to poor reproducibility of speaker characteristics. From these results, we confirmed the embedded speaker code-based approach for speaker adaptation task was not effective because it was worse than DNN_SD when using 300 target speaker utterances and also worse than HMM when using five target speaker utterances.

2.5.2.2 Objective evaluation for SPKCODE_ONE-HOT

Figure 2.5 presents the MCDs and F0 RMSEs for SPKCODE_ONE-HOT models. The horizontal axis indicates the speaker code input layer. The results for conventional HMM and DNN_SD are the same as those in Fig. 2.4.

First, we compared the performance for the model structure of the SPKCODE_ONE-HOT models. For both the multi-speaker modeling and speaker adaptation tasks, we found that the models using the 4th hidden layer gave consistently low MCDs and F0 RMSEs when using five target speaker utterances while the models using all hidden layers gave lower MCDs and F0 RMSEs when using 300 target speakers' utterances. From these results, we determined that the models using the 4th hidden layer were optimal when using five target speaker utterances



(a) five utterances for male target speaker

(b) five utterances for female target speaker

(c) 300 utterances for male target speaker

(d) 300 utterances for female target speaker

Figure 2.5: The objective evaluation results for one-hot speaker code models. The horizontal axis indicates the speaker code input layer.

while the models using all the hidden layers were optimal when using 300 target speaker utterances, for both multi-speaker training and speaker adaptation tasks.

We then compared the performance for the SPKCODE_ ONE-HOT models with that of the conventional methods. From the results for the multi-speaker modeling task, we found the relationship of

- SPKCODE_ONE-HOT < HMM
  (five utterances, MCDs)
- HMM < SPKCODE_ONE-HOT
  (five utterances, F0 RMSEs)
- SPKCODE_ONE-HOT < DNN_SD < HMM (300 utterances, MCDs)
- HMM < SPKCODE_ONE-HOT < DNN_SD (300 utterances, F0 RMSEs)

when the model structure for SPKCODE_ONE-HOT was set at optimal. For each condition, we found that the performance of DNN_ONE-HOT was consistently better than that of DNN_SD models. These results confirmed the prediction performance improvement of SPKCODE_ONE-HOT using a multi-speaker speech corpus. When SPKCODE_ONE-HOT is compared with HMM, it is found that MCDs of SPKCODE_ONE-HOT are improved. The advantage of DNN over HMM in modeling mel-cepstra was also confirmed for SPKCODE_ONE-HOT models. We can also see that F0 RMSEs of SPKCODE_ONE-HOT were comparable to or higher than those of HMM. These results revealed that, relative to HMM, there is still room to improve F0 prediction in SPKCODE_ONE-HOT models, although their performance is considerably better than that of conventional DNN_SD models.

For the speaker adaptation results, we can see found the relationship of

- SPKCODE_ONE-HOT < HMM
  (five utterances, MCDs)
- HMM < SPKCODE_ONE-HOT
  (five utterances, F0 RMSEs)
- DNN_SD < SPKCODE_ONE-HOT < HMM (300 utterances, MCDs)
- HMM < SPKCODE_ONE-HOT < DNN_SD (300 utterances, F0 RMSEs)

when the model structure for SPKCODE_ONE-HOT was set at optimal. When the performance of SPKCODE_ONE-HOT was compared with that of DNN_SD, it was found that MCDs were degraded while F0 RMSEs were improved. Since the multi-speaker speech corpus for SPKCODE_ONE-HOT model training has a larger variety of context than those for DNN_SD, SPKCODE_ONE-HOT models are especially advantageous in F0 prediction. From the results of MCDs' degradation results we obtained, we concerned that the adaptation by SPKCODE_ONE-HOT could not represent the precise speaker characteristics by effectively using the large amount of adaptation data. The results revealed that more flexible adaptation models were needed when given a large amount of adaptation data. To solve this problem, one promising approach would be a more elaborate speaker code-based adaptation method for which the number of parameters for adaptation can change in accordance with the amount of adaptation data, in the same way that for the HMM-based adaptation method, CSMAPLR reported by yamagishi et al. [11]. When we compared SPKCODE_ONE-HOT with HMM, we found the relationship was the same as that for the multi-speaker modeling task; MCDs were improved while F0 RMSEs were degraded.

Finally, we compared the performance of the two proposed models. For multi-speaker modeling, the performance of SPKCODE_ONE-HOT and SPKCODE_EMBED was comparable regardless of the amount of training data when the model structure was set optimal. For speaker adaptation, SPKCODE_ONE-HOT models showed superior performance to that of the SPKCODE_EMBED models regardless of the amount of adaptation data or the objective measure. The difference between the two models was especially large when the amount of the adaptation data was large. This is because the performance of SPKCODE_EMBED models saturated with the amount of adaptation data while SPKCODE_ONE-HOT models showed greater performance for a larger amount of adaptation data. SPKCODE_ONE-HOT models can represent the speaker characteristics in more detail because they have a larger number of parameters for each speaker. The number of parameters for adaptation for these models, whose speaker code vector is connected to all hidden layers, is equal to unit size $\times$ the number of hidden layers. In our experimental condition, for example, it is equal to $1024 \times 5 = 5120$. Since this number is greater than that for SPKCODE_EMBED models ($2, 10, 33, 128$ or 256 as described in Sect. 2.5.2.1), SPKCODE_ONE-HOT models show more

precise reproducibility for speaker characteristics. These results confirm that reducing adaptation parameters by adopting the SPKCODE_EMBED model was not effective for adapting speech synthesis models to speakers when five utterances were given as adaptation data. From the results of the objective evaluation experiment, we concluded that SPKCODE_ONE-HOT models generally perform better than the SPKCODE_EMBED and the best speaker code input layer for SPKCODE_ONE-HOT models for all the hidden layers.

2.5.3 Subjective evaluation

We conducted subjective evaluations with respect to the naturalness and similarity of synthesized speech to confirm the effectiveness of the proposed method. We used four models for these evaluations, DNN_SD, HMM (multi-speaker modeling), SPKCODE_ONE-HOT (multi-speaker modeling) and SPKCODE_ONE-HOT (speaker adaptation). The proposed SPKCODE_EMBED models were eliminated from these experiments because the objective evaluation experiments revealed that the proposed SPKCODE_ONE-HOT models perform better and they are not necessary to reveal the effectiveness of the proposed method. We used the optimal architectures for SPKCODE_ONE-HOT models as discussed in the last section, the model using the 4th hidden layer when using five target speaker utterances and the model using all hidden layers when using 300 target speaker utterances. The number of listeners was 24 for a naturalness test and 22 for a similarity test. We conducted five-point MOS and DMOS tests. The scale for the MOS test was five points for "very natural" and 1 points for "very unnatural". The scale for the DMOS test was five points for "very similar" and 1 point for "very dissimilar".

Figures 2.6 and 2.7 show the naturalness and similarity scores obtained in the subjective evaluations with confidence intervals of 95%. We found the relation of HMM < SPKCODE_ONE-HOT for both naturalness and similarity scores. Furthermore, the scores of SPKCODE_ONE-HOT using five target speaker utterances were equivalent to those of DNN_SD using 300 target speaker utterances. We can also see the relation of DNN_SD < SPKCODE_ONE-HOT for both naturalness and similarity when using 300 target speaker utterances. These results confirmed that the proposed method can improve the synthetic speech quality by using a

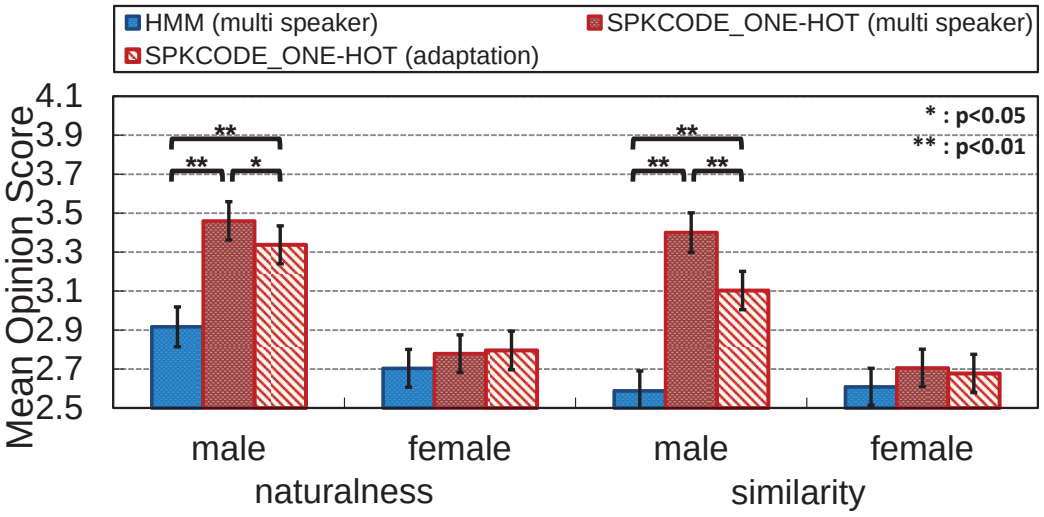


Figure 2.6: Naturalness and similarity test results with their 95% confidence interval. (The number of target speaker utterances: 5)

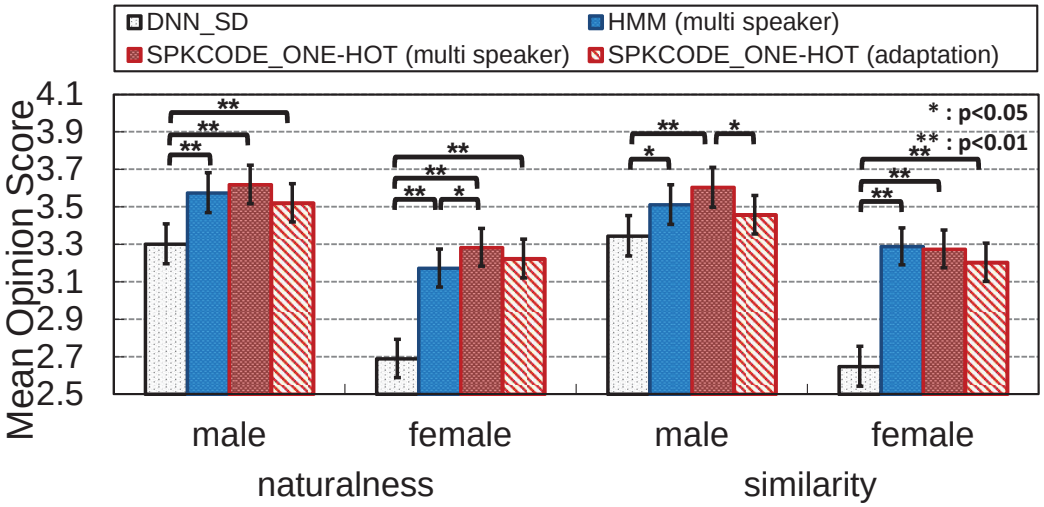


Figure 2.7: Naturalness and similarity test results with their 95% confidence interval. (The number of target speaker utterances: 300)

multi-speaker speech corpus in DNN-based speech synthesis. On the other hand, there were no significant differences between HMM and SPKCODE_ONE-HOT when using 300 target speaker utterances.

Although the speaker adaptation scores obtained for SPKCODE_ONE-HOT were slightly worse than those obtained for multi-speaker modeling under each condition, they were higher than those obtained for HMM when using five target speaker utterances and comparable to those obtained for HMM and higher than

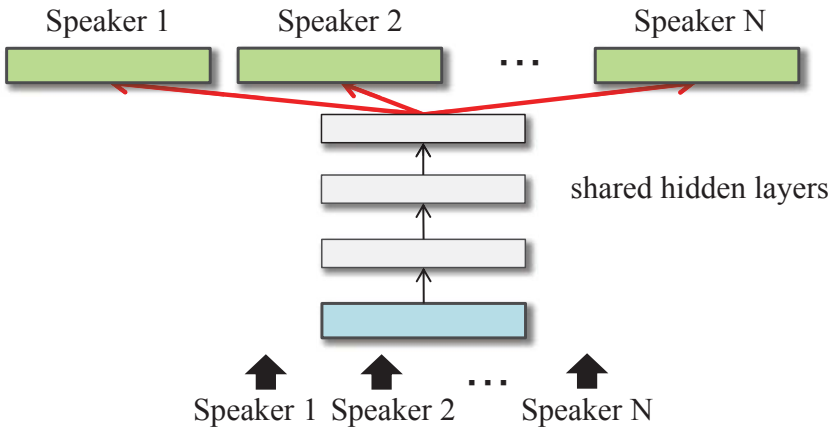


Figure 2.8: The model architecture of the conventional shared hidden layer model [1]. The red bold lines indicate the re-estimated parameters for speaker adaptation.

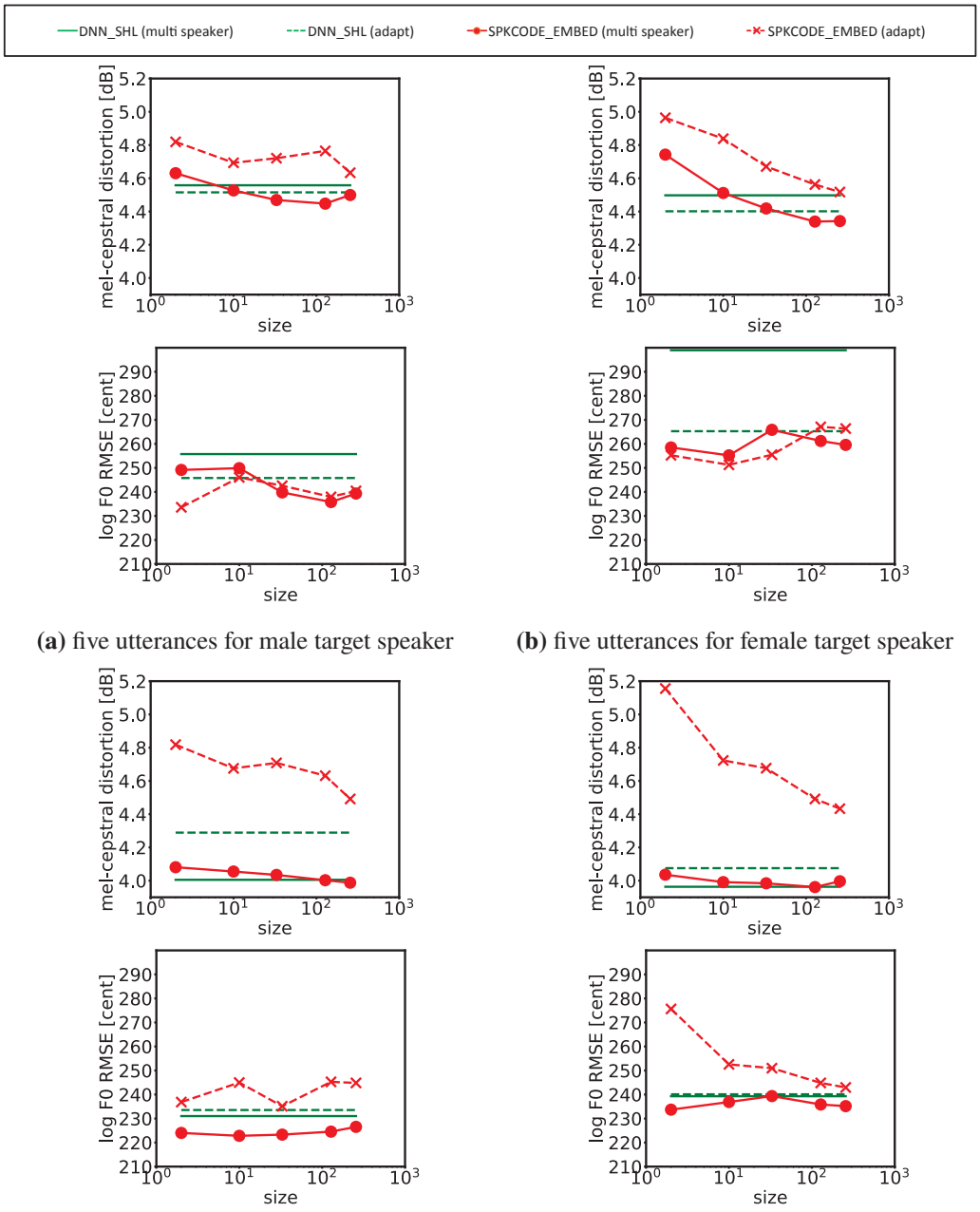
DNN_SD when using 300 target speaker utterances. These results confirmed that the adaptation based on speaker codes generates speech whose quality is comparable to or higher than that obtained with a conventional speaker dependent DNN, or a speaker adapted HMM.

Although we were able to observe performance improvement with the proposed method for the two target speakers in our experiments, the obtained subjective evaluation results were not always consistent between the two target speakers. Therefore, our future work will include further investigating whether or not the improvements by the proposed method provides can be obtained for any speakers included in the training corpus. For HMM-based cluster adaptive training [13], it is known that the performance depends on the speaker similarity between a target speaker and those in the training corpus.

2.6 Comparison with the conventional shared hidden layers (SHL) model [1]

2.6.1 Objective of the experiments

As a method of DNN-based multi-speaker modeling and speaker adaptation for DNN-based speech synthesis, a shared hidden layer (SHL) [1] model is proposed as well as the proposed speaker code-based method. In this section, we compare the



(a) five utterances for male target speaker

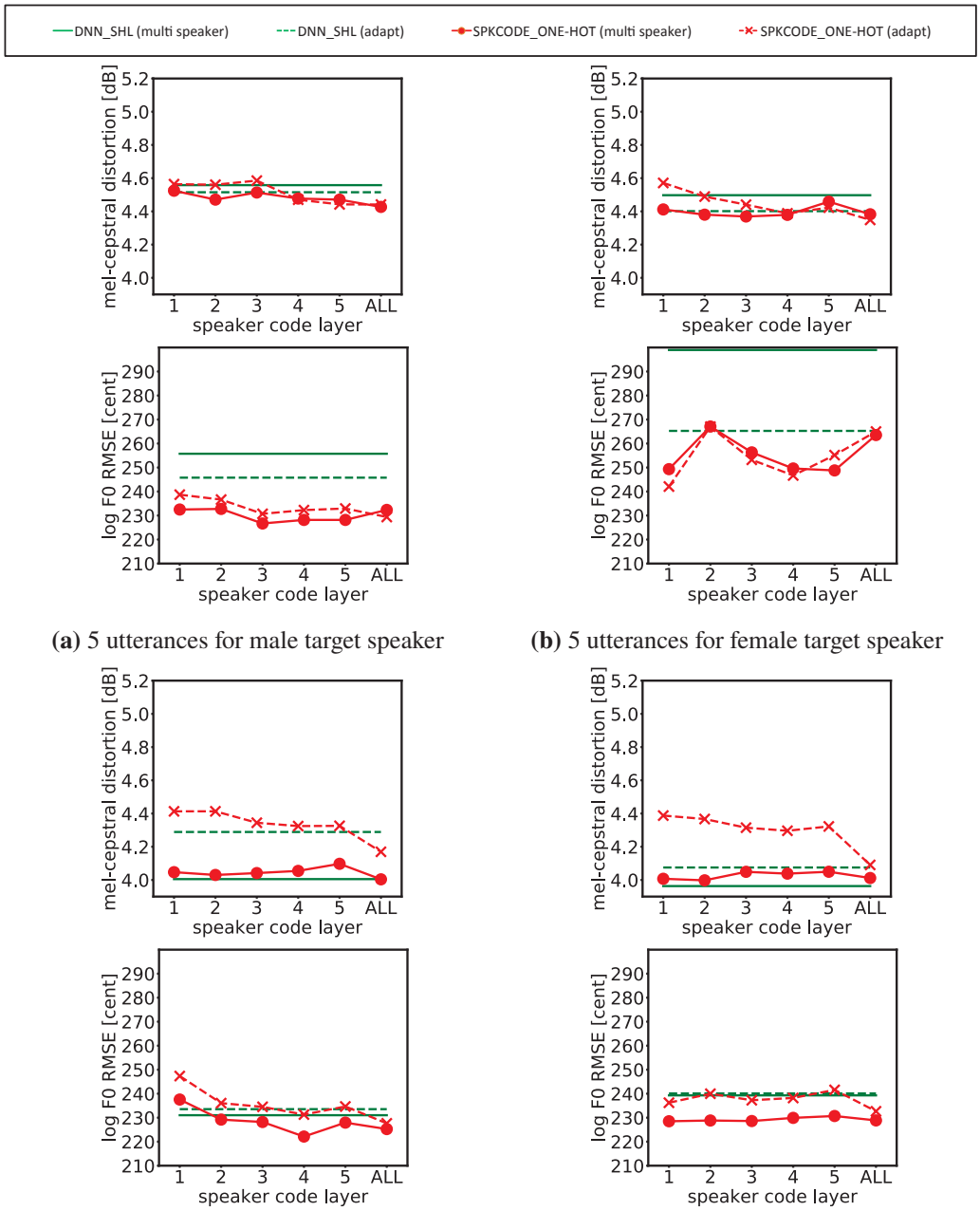
(b) five utterances for female target speaker

(c) 300 utterances for male target speaker

(d) 300 utterances for female target speaker

Figure 2.9: The objective evaluation results for embedded speaker code model. The horizontal axis indicates the size (dimension) of the speaker codes.

performance of the conventional SHL and the proposed model through objective measurements.



(a) 5 utterances for male target speaker

(b) 5 utterances for female target speaker

(c) 300 utterances for male target speaker

(d) 300 utterances for female target speaker

Figure 2.10: The objective evaluation results for one-hot speaker code models. The horizontal axis indicates the speaker code input layer.

Fig. 2.8 shows the architecture of the conventional shared hidden layer structure. In this model, hidden layers are shared across all the speakers in the training corpus, and can be considered as the global linguistic feature transformation shared by all

the speakers. Conversely, all speakers have their own output layers, the so-called regression layer, to model their own specific acoustic spaces. Compared with the speaker dependent DNN, the shared hidden layer model takes the same input linguistic feature, which is converted from text in the same manner, and the same output acoustic feature for each speaker.

For the multi-speaker modeling task performed by the shared hidden layer model, connection weights of the whole DNN are trained using a multi-speaker speech corpus. When synthesizing a speech parameter sequence, a target speaker is chosen from the corpus and the corresponding output layer is selected to generate the speaker's voice.

For the speaker adaptation task performed by the shared hidden layer model, first, the model is trained using multi-speakers' speech corpus. Then, the hidden layers transferred from multiple speakers' data are fixed and only the regression layer is updated using the target speaker's speech data. While this re-estimation can be conducted by a back propagation algorithm, it is also possible to use the least squares method to minimize the squared residuals between prediction and ground-truth since there is only a linear regression between the last hidden layer's output and the target [1]. In our experiments, we used a back-propagation algorithm to re-estimate these values. Specifically, we randomly initialized the connection weights to a new output layer and then optimized them to minimize the mean squared error between the output feature of the adaptation data and predicted values, using the Adam-based back-propagation algorithm [27].

The conventional SHL models are indicated by "DNN_SHL" below. For DNN_SHL, the output layers are composed of speaker-dependent nodes while the architecture of the shared hidden layers (the number of hidden layers and their units) was set to the same as that of other DNN models. The other experimental conditions were set in a way similar to those for the DNN_SPKCODE_ONE-HOT models.

2.6.2 Experimental results

Fig. 2.9 and Fig. 2.10 show the objective evaluation results for SHL. When SPKCODE_EMBED is compared with DNN_SHL in Fig. 2.9, it is found that SPKCODE_EMBED shows performance comparable or superior or comparable

performance to that of DNN_SHL for the multi-speaker modeling task when the size of the speaker code is set large. On the other hand, SPKCODE_EMBED show inferior performance to DNN_SHL for the speaker adaptation task. This is because the DNN_SHL has a larger number of free parameters for adaptation, which is equal to $(1024 \times 139 =)142336$. In comparison, DNN_EMBED has free parameters for adaptation of $2, 10, 33, 128$ or 256 as described in Sect. 2.5.2.1. The larger number of free parameters led to more precise reproducibility for speaker characteristics.

When SPKCODE_ONE-HOT is compared with DNN_SHL models in Fig. 2.5, it is found that MCDs for SPKCODE_ONE-HOT are lower than those for DNN_SHL models when using five target speaker utterances. It is also revealed that MCDs of SPKCODE_ONE-HOT are equivalent to those for DNN_SHL models when using 300 target speakers' utterances. We can also see that F0 RMSEs of SPKCODE_ONE-HOT are consistently lower than those for DNN_SHL. These results confirmed that the proposed SPKCODE_ONE-HOT models are advantageous compared with DNN_SHL in both multi-speaker modeling and speaker adaptation tasks, especially in modeling F0 contour. Since DNN_SHL models utilize the connections to the output layer to model the speaker characteristics, they can represent only the simple characteristics that can be interpreted by a single affine transformation. In contrast, since SPKCODE_ONE-HOT models utilize the augmented input vector to model complex context dependencies, they are advantageous in modeling speaker characteristics more precisely.

2.7 Conclusion

In this chapter, we proposed a DNN-based speech synthesis method using speaker codes to improve speech quality by using a multi-speaker speech corpus. Objective evaluation results showed that the proposed model outperforms the speaker dependent DNNs and multi-speaker DNNs with shared hidden layers for both multi-speaker modeling and speaker adaptation tasks. Subjective evaluation results showed that the proposed model can produce more natural speech than the conventional speaker dependent DNNs and HMM-based speaker adaptation method, especially when using a small number of target speaker utterances. Our

future work will include investigating the relationship between the performance and the speaker similarity between training and target speakers for the proposed DNN-based multi-speaker modeling and speaker adaptation.

Chapter 3

DNN-based speech synthesis using dialogue-act information and its evaluation with respect to illocutionary act naturalness

This chapter aims at improving the naturalness of synthesized speech generated by a TTS synthesis used for a spoken dialogue system with respect to "how natural the system's intention is perceived via the synthesized speech". In this paper, we call this measure "illocutionary act naturalness (IAN)." In order to achieve this aim, we propose to utilize DA information as an auxiliary feature for a DNN-based speech synthesis system. First, we construct a speech database with DA tags. Second, we build the proposed DNN-based speech synthesis system based on the database. Then, we evaluate the proposed method by comparing its performance with two conventional HMM-based speech synthesis systems—namely, the style-mixed modeling method and the style adaptation method. The objective evaluation results show that the proposed method overwhelms the style-mixed modeling method in terms of the accuracy of reproduction of global prosodic characteristics of DAs. The results also reveal that the proposed method overwhelms the style adaptation method in the accuracy of reproduction of sentence-final tone characteristics of DAs. Further, the subjective evaluation results also indicate that the proposed method improves the IAN more as compared with the two conventional

methods.

3.1 Introduction

This chapter proposes DNN-based speech synthesis considering intention in order to improve the IAN of the synthesized speech generated by a spoken dialogue system. In human-to-human communication, participants infer speaker's intention for every utterance [3], [4]. If the expression of the speaker's intention is unnatural, it will increase the cognitive load for inference, which will result in the degradation of the quality of communication. For spoken dialogue systems as well, it is considered necessary to express the system's intention to improve the quality of communication. Since perceived intention differs depending on the prosodic characteristics (fundamental frequencies (F0) and phoneme durations (durations)) [5], [6], it is necessary for speech synthesis to express intentions naturally by controlling prosodic characteristics.

In the speech synthesis field, naturalness of the synthesized speech is often evaluated as the quality of speech samples separated from context. As described above, speech also needs to be natural considering intentions to be expressed. Based on the classification by the speech-act theory [9], [10], the former is rephrased as naturalness with respect to the act of saying something (locutionary act), and the latter is rephrased as naturalness with respect to the act of conveying intentions (illocutionary act). This thesis defines the former as "locutionary act naturalness (LAN)" and the latter as "illocutionary act naturalness (IAN)". This thesis aims at improving the IAN of synthesized speech.

In order to improve the IAN, it is considered necessary to reproduce the prosodic characteristics of intentions in synthesized speech. This problem is related to "emotional" or "speaking style" speech synthesis [11], [31]–[36]. Emotional speech synthesis can express several (typically, 3∼10) emotions, such as "happy" or "sad", in synthesized speech. For acoustic models, HMM-based [11], [31], [32], DNN-based [36], and long short-term memory-recurrent neural network (LSTM-RNN)-based [34], [35] methods are proposed and shown to be effective. These methods can also be effective for expressing intentions, since they all express paralinguistic information [37]. On the other hand, there are certain

differences between emotional speech synthesis and this study. One difference is the importance of sentence-final tone for expressing intentions. Emotional speech has salient characteristics for an entire utterance. For example, it is known that "sad" speech generally has lower F0 and slower speech rate [38]. Similar characteristics are presented for intentions [39]. On the other hand, for Japanese speech, the sentence-final tone characteristics are also important [40]–[42]. In particular, it is shown that the subtle difference in sentence-final tone can change intentions and nuances [42]. Here, sentence-final tone depends on sentence-final morphemes (such as sentence-final particles) as well as intentions [40]–[42]. This is in contrast with global prosodic characteristics, which has less dependency on sentences. The difference above shows the necessity of validating whether emotional speech synthesis methods are also effective for expressing intentions.

There are two previous studies that investigated speech synthesis by considering intentions. One is used for utterance generation for a call center operator [39]. This study is based on concatenative speech synthesis. The other is used for utterance generation for restaurant and hotel guidance [43]. This work is based on HMM-based speech synthesis. While these studies evaluated the LAN and showed its improvement, they do not evaluate the IAN or the DNN-based speech synthesis for expressing intentions. Since DNNs are shown to be effective for emotional speech synthesis [36], they are expected to be effective for expressing intentions as well. This thesis investigates DNN-based speech synthesis considering intentions and compares it with HMM-based conventional methods with respect to the IAN.

Further, this study uses conventional HMM-based methods such as the average-voice-based speech synthesis technique with model adaptation [12], [44] and style-mixed model [31]. The models are trained with regard to each intention as a category of emotions. For the model adaptation methods, average voice models are created from several intensions' speech data and are adapted to a small amount of speech data of a target intention. For average voice model training, a single decision tree is built for all the intentions. Here, for decision tree building, only questions with respect to linguistic information (e.g. phonemes and accents) are used and the intentions are not considered. Therefore, the derived tree structure is not always optimal for modeling prosodic characteristics, which depend on both linguistic and intention information. In particular, as described above, sentence-final tones depend on sentences as well as intentions. Therefore,

there is concern that model adaptation methods do not sufficiently reproduce the sentence-final tone characteristics. On the other hand, the style-mixed modeling utilizes questions with respect to both linguistic and intention information to build a decision tree. Therefore, it is expected that an optimal tree structure is derived for reproducing sentence-final tones. However, questions with respect to intentions are not always used for tree structure because the decision is based on the criterion of minimum description length (MDL) [28]. Then, there is concern that it does not sufficiently reproduce the prosodic characteristics of intentions. Based on the above discussion, HMM-based conventional methods have concerns for reproducibility of the prosodic characteristics of intentions due to the constraints of the tree structure.

The proposed methods model the relationship between input information (linguistic and intention information) and output information (acoustic features) by a neural network. Since neural networks do not have explicit constraints for the input-output relationships, they are expected to solve the problems of the conventional methods and improve the IAN. As described above, the reproducibility of prosodic information is particularly important to improve the IAN. Therefore, this work proposes DNN-based speech synthesis, utilizing intentions as auxiliary features, as described in the previous section. This network architecture is expected to be effective in reproducing sentence-final tone characteristics, since the estimation error of F0 is smaller than the shared hidden layer models [20], as shown in Chap. 2. While sequence models [34]–[36], [45] are also expected to be effective, we focus on feed-forward DNNs due to the restriction of the size of the speech database.

In this chapter, the speech database with intention informations are described in Sect. 3.2. We also describe the analysis on the prosodic characteristics using the speech corpus in Sect. 3.2. The proposed method utilizing intentions as auxiliary features is described in Sect. 3.3. The evaluation experiments are described in Sect. 3.4.

3.2 Speech database with DA information

Number	Speaker	Class	Utterance
	A		I'm not unfamiliar with sports, and I do nothing after I played tennis from elementary school to junior high.
	A		How about you?
	B		I see.
	B		I played basketball.
	B		I also play tennis for enjoyment.
3041	A	Sd_e	**I have no experience in sports that requires team play,**
3041	A	Apro	**I respect that you can play basketball.**
3041	A	Sd_f	**I've also read about slam dunks.**

Figure 3.1: An example of recorded sentences. The last utterances indicated in a red bold font were recorded.

3.2.1 Speech corpus building

In order to investigate speech synthesis considering speech intentions, we need categories (tag sets) of intentions and a speech database annotated with tags. This study utilizes DA sets that were designed for non-task oriented dialogue systems [2] and builds a speech corpus with DA sets. DAs are abstract expression of intentions [46]. The DA sets utilized in this study are designed to describe speech intentions in non-task oriented dialogues. In a usual speech recording for speech synthesis, the speaker reads aloud each sentence, which are separated from their contexts. With regard to the sentences, we often use phoneme-balanced sentences, like ATR 503 sentences [47]. However, in order to express intentions naturally, it is considered desirable to utilize sentences that express intentions that are consistent with contexts. Therefore, in the speech recording of this study, the recorded sentence depicted in Fig. 3.1 is used. While the original tag sets consist of 33 DAs, we used 30 of them by excluding "other", etc. During speech recording, the speaker reads the utterance of the last turn (indicated by the red bold font in Fig. 3.1). The utterances before the recorded utterances are printed to indicate the context. In addition, during speech recording, we instructed the speaker to read aloud sentences naturally and in accordance with the context and DA.

In this thesis, a text chat corpus [48] is used for designing the sentences that are recorded. This corpus contains text dialogues with 12 themes, such as food, travel,

Table 3.1: Statistics of the speech database. [#] indicates the number of utterances, and [sec] indicates the utterance time length (seconds).

DA		amount of speech		DA		amount of speech	
		[#]	[sec]			[#]	[sec]
information	Info	120	293	question_habit	Q_h	108	255
self-disclosure_fact	Sd_f	127	324	question_plan	Q_pl	92	231
self-disclosure_experience	Sd_e	91	298	question_desire	Q_d	75	237
self-disclosure_habit	Sd_h	101	310	question_preference	Q_pr	108	260
self-disclosure_plan	Sd_p	108	275	confirmation	Conf	120	273
self-disclosure_desire	Sd_d	113	298	proposal	Prop	99	234
self-disclosure_preference+	Sd_+	184	530	question_self	Q_s	111	285
self-disclosure_preference-	Sd_-	186	490	repeat	Rept	149	277
self-disclosure_preference0	Sd_0	109	312	paraphrase	Para	121	291
sympathy	Sym	146	321	acknowledgement	Ack	83	165
non-sympathy	N-Sym	176	416	filler	Fill	71	130
approval	Apro	173	400	admiration	Admn	130	193
question_information	Q_i	122	301	greeting	Grt	74	140
question_fact	Q_f	111	267	thanks	Thks	26	81
question_experience	Q_e	100	262	appology	Apol	70	196
					sum	3410	8353

and movies. The corpus is in a dialogue format with two people, and a total of 12 dialogues are collected. In addition, a DA is annotated for each utterance by two experts. From the corpus, the recorded sentences are selected by considering the balance of frequency of each DA and each phoneme on the basis of entropy [49].

In this study, we recorded the utterances of a professional female voice actor. A total of 5177 utterances were recorded. Since the recorded 5177 utterances included those with the same sentence and DA, we excluded such duplicate utterances. As a result, we used 3410 utterances (a total of approximately 140 minutes) as the speech database in this chapter. The statistical information of the speech database is presented in Table 3.1. For all utterances, the phonemes, accent types, phoneme boundaries, accent phrase boundaries, and other accent information were manually annotated in order to use them as training data for speech synthesis models.

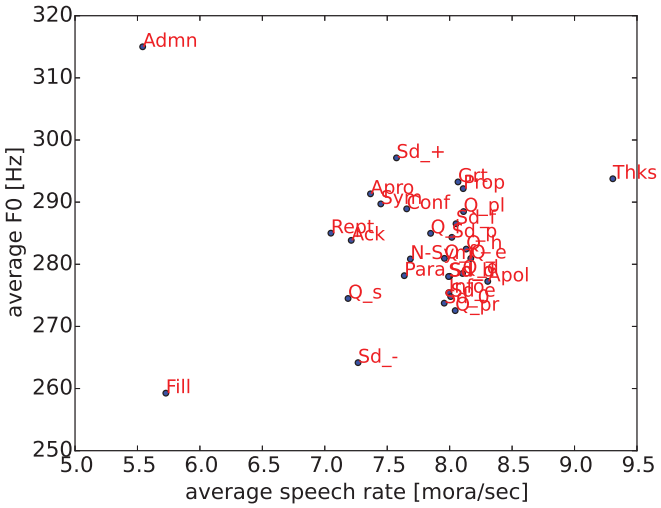


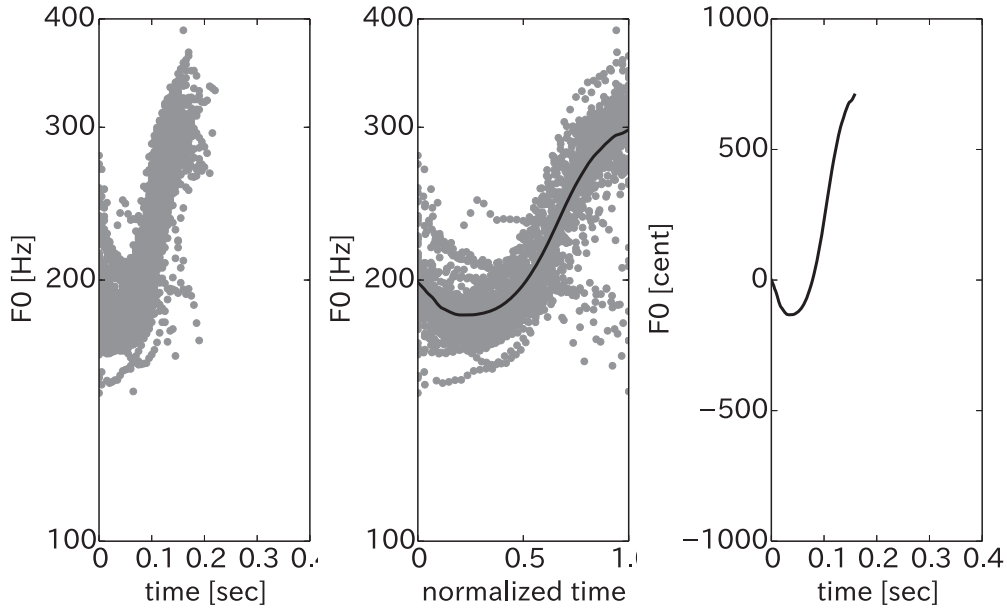
Figure 3.2: Scatter plot of average F0 and speech rate for each DA.

3.2.2 Speech analysis

In order to reveal the prosodic characteristics of each DA, we analyzed prosody based on the speech database that we constructed in Sect. 3.2.1.

3.2.2.1 Analysis of global prosodic characteristics

The average speech rate and average F0 for each DA are depicted in Fig. 3.2. In order to calculate the average speech rate, we used the phoneme boundary information, which was provided manually. For F0 analysis, we used STRAIGHT [26]. It is evident that "admiration (Admn)", "filler (Fil)," and "thanks (Thks)" have significantly different prosodic characteristics from others. In addition, the average F0 tends to be higher for "self-disclosure_preference+ (Sd_+)", "approval (Apr)", and "sympathy (Sym)". On the other hand, it is evident that the average F0 tends to be lower for "self-disclosure_preference- (Sd_-)". It is also evident that the average speech rate tends to be slow for "repeat (Rept)", "question_self (Q_s)", "approval (Apro)", "sympathy (Sym)," and "self-disclosure_preference- (Sd_-)." By considering these DAs for speech synthesis, these global prosodic characteristics are expected to be reproduced and the IAN improved. In contrast, global prosodic characteristics were not significantly different for other DAs.



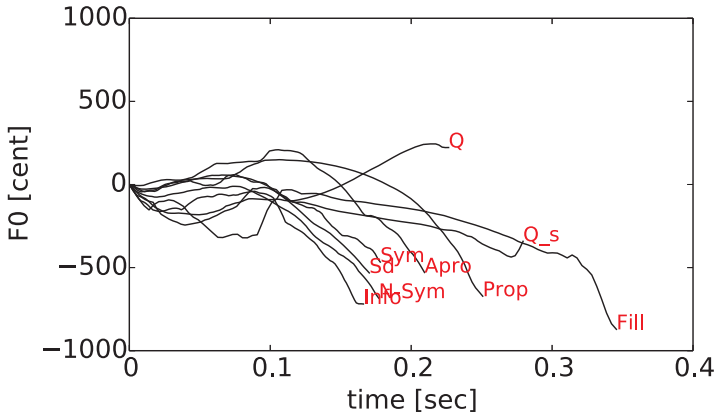
(a) F0 for sentence-final syllable **(b)** Normalization of duration **(c)** Average F0 for sentence-final syllable

Figure 3.3: The procedure to derive average sentence-final F0. The figure is an example for sentence-final particle "ka" and DA "question_experience".

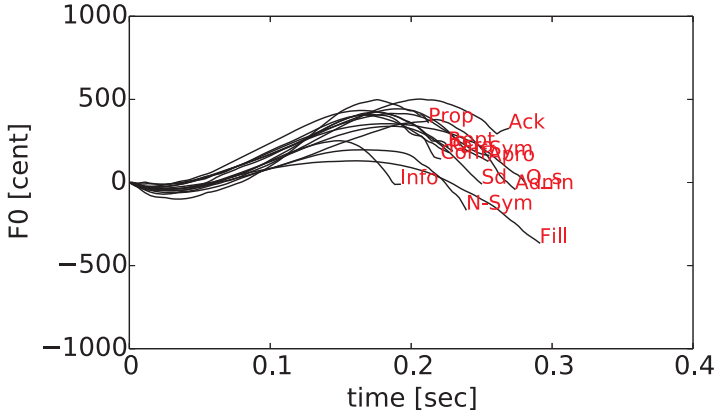
3.2.2.2 Analysis of sentence-final tones

In this section, we analyze the characteristics of the sentence-final tone of DA. It is known that even in utterances with the same sentence, the sentence-final tone can be different depending on intentions [5], [42]. We test whether our speech database has the same tendency. Here, sentence-final tones depend on sentence-final morphemes (e.g. sentence-final particles) as well as intentions. Therefore, the sentence-final tone is analyzed for each sentence-final morpheme. Since there are numerous end-of-sentence morphemes, it is difficult to show the results for all of them. Therefore, the analysis results are presented for three sentence-final morphemes ("No sentence-final particle", "-ne", and "-ka"), which were the top three frequencies in the corpus. In order to describe the sentence-final tones, we used the following classification: (question rising tone↗, insisting rising tone↑, rise-fall tone↷ and falling tone ↘) [50].

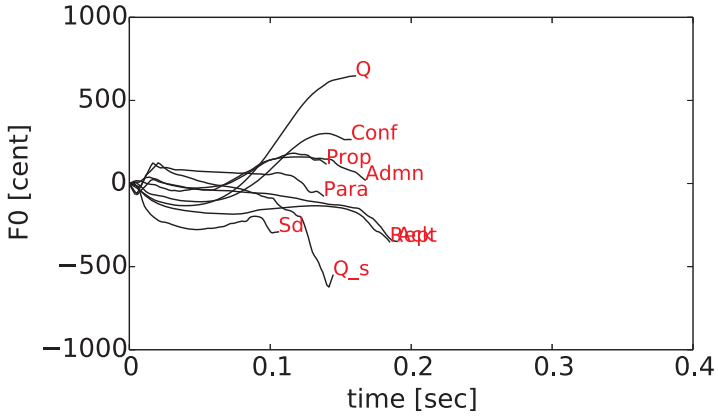
The analysis method is illustrated in Fig. 3.3. First, we extracted the F0 trajectory of the last syllable (sentence-final F0) for the utterances of a certain



(a) no sentence-final particle



(b) sentence-final particle: “ne”



(c) sentence-final particle: “ka”

Figure 3.4: Average F0 of the final syllable for each sentence-final context and DA

sentence-final morpheme and DA (the dotted lines in Fig. 3.3a). Next, for each utterance, we normalized the duration of sentence-final F0 (the dotted lines in Fig. 3.3b). We also calculated the average duration of sentence-final F0. Finally, we adjusted the duration of the average sentence-final F0 so that it becomes the average duration (the solid line in 3.3c). The plot was also adjusted in terms of the frequency axis direction so that the frequency of the starting edge was 0. In order to simplify the plots, the subcategories of "self-disclosure (Sd)" and "question (Q)" were not distinguished, while "question_self (Q_s)" was distinguished from "Q" because its tendency was different.

Fig. 3.4 depicts the average sentence-final F0 for each sentence-final morpheme. Fig. 3.4a depicts that sentence-final F0 descends by approximately 500 cents for "Info" or "Sd", while it rises by approximately 200 cents for "Q" for "No sentence-final particle". This is because falling tone was used for "Info" and "Sd" (e.g. "結構有名です↘/It is quite famous."), while a question rising tone was used for "Q". (e.g. "パクチーの日本名って知ってます↗/Do you know what is the Japanese for coriander?")

Fig. 3.4b indicates that sentence-final F0 of "Fill" descends by approximately 400 cent for "-ne", while other DAs do not have such a tendency. This is because the insisting rising tone was used for other DAs (e.g. "そうなんですね↑/I see."), while the falling tone was used for "Fill" (e.g. "そうですね⌒/Let's see.").

Fig. 3.4c indicates that sentence-final F0 descends for "Ack", "Rept", "Para", "Admin", and "Q_s" for "-ka", while it rises for "Q" and "Conf". This is because the falling tone was used for "Ack", "Rept", "Para", "Admin", and "Q_s" (e.g. "一人旅ですか↘/You're traveling alone."), while the question rising tone was used for "Q" and "Conf" (e.g. "小学校の宿題ですか↗/Is it homework for elementary school?").

These analysis results revealed the dependency of sentence-final F0 on DAs in our speech database. The results also reveal that sentence-final F0 depends on sentence-final morphemes. It is evident that there were several DAs (e.g. "Sd" and "Q") that have similar global prosodic characteristics but different sentence-final tone characteristics. Based on these analysis results, it is considered important to reproduce both global and sentence-final F0 characteristics in order to improve the IAN.

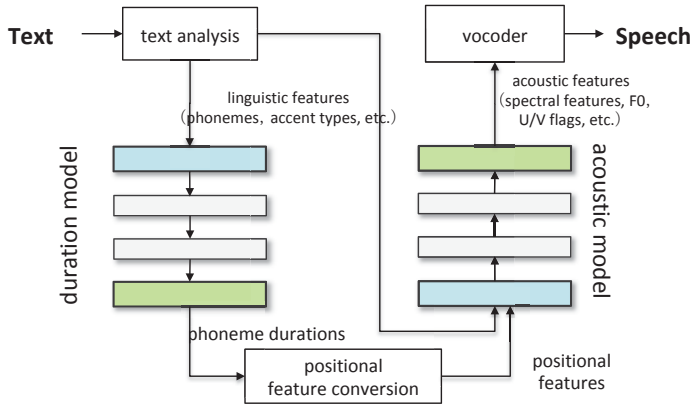


Figure 3.5: A schematic diagram of DNN-based speech synthesis.

3.3 Speech synthesis method

3.3.1 DNN-based speech synthesis (DNN-BASELINE)

Fig. **??** presents a schematic diagram of conventional DNN-based speech synthesis [14]. The conventional methods utilizes DNN as a regression model from input linguistic feature vectors to output duration vectors (duration models) and acoustic features (acoustic models). First, the input text is converted to linguistic feature vectors. Linguistic vectors consist of binary variables that correspond to answers to questions on linguistic features and numerical variables. Next, phoneme durations are estimated from the linguistic feature vectors using duration models. Then, the frame features [51] are calculated on the basis of the estimated durations and they are input to acoustic models along with linguistic feature vectors to derive acoustic feature vectors. The acoustic feature vectors consist of spectral features, F0, and dynamic features. Then, the estimated acoustic feature vectors are converted to output speech waveforms.

3.3.2 DNN-based speech synthesis considering DAs (DNN-DACODE)

The straightforward method to generate speech considering DA is to train a model for each DA and to switch the model depending on the DA when synthesizing. However, this method needs a considerable amount of training database for each DA in order to obtain high-quality speech, which increases the training cost. The

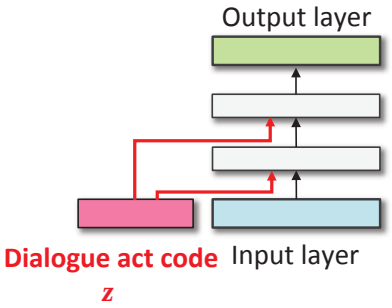


Figure 3.6: Schematic diagram of the proposed method. The same model architectures are used as duration models and acoustic models.

proposed method regards DAs as well as linguistic features as input features, and models the speech of all of the DAs using a single model. It is expected that high-quality speech is derived using a small amount of training data for each DA, because the part of the network that regresses linguistic feature vectors to acoustic feature vectors is trained using the speech data of all DAs. Further, the proposed method does not constrain the relationship between input and output vectors. Therefore, it is expected to model the complex dependency among linguistic feature vectors, DAs, and acoustic features. The conventional HMM-based model adaptation methods [12], [44] have the constraint that the decision tree does not include a question regarding DAs. Hence, the derived tree structure is not always optimal for reproducing sentence-final tones, because they are dependent on both sentence-final morphemes and DAs. Therefore, there is concern that the sentence-final tones are not reproduced. On the other hand, the conventional style-mixed modeling [31] decides whether to use questions for DAs depending on the MDL criterion. Therefore, if the prosodic characteristics of intentions are not salient, there is concern that they are not always reproduced. It is considered that both the conventional methods have the problem of reproducibility of prosodic characteristics due to the constraint of the decision tree structure. It is expected that the proposed method can solve this problem by utilizing neural networks.

Fig. 3.6 presents the schematic diagram of the proposed method. The proposed DNN-based method is basically the same as the DNN-based speech synthesis using speaker codes [52]. The difference is that the proposed method utilizes DAs instead of speaker id. The DA codes $z_c = [z_{1c}, \cdots, z_{Mc}]$ are described below for each DA

$c (c = 1, \cdots, M)$, where the number of DAs is M;

$$z_{mc} = \begin{cases} 1 & (m = c) \\ 0 & (m \neq c). \end{cases} \tag{3.1}$$

The DA code z_c is regressed by a set of weight parameters and input to hidden layers. In the process of training a model, all the parameters of the network are trained using linguistic feature vectors, DA codes, and acoustic feature vectors of all DAs included in the training data. When synthesizing, the DA code is input to the DNNs along with linguistic feature vectors and the output acoustic features are derived by forward propagation.

3.4 Evaluation Experiments

3.4.1 Design of the test sets

This section describes the design of the test set for the evaluation experiments. Previous studies randomly select test sets from the speech database. However, as described in Sect. 3.2.2, the prosodic characteristics of the DAs depend on not only DAs but also on sentence-final morphemes. Therefore, evaluation results depend on randomly selected sentence-final morphemes, which result in degradation of the reproducibility of the evaluation results. Hence, this study designs test sets by controlling DAs and sentence-final morphemes. Here, there are 30 patterns for DAs and numerouspatterns for sentence-final morphemes; thus, it is difficult to use all their combinations due to the evaluation cost entailed. Therefore, we

1. classify sentences and DAs into a small number of classes (Sect. 3.4.1.1 and Sect. 3.4.1.2) and

2. uses the top three sentence classes in terms of frequency for each DA class (Sect. 3.4.1.3).

Table 3.2: Sentence classes and the classification criteria.

Class Name	Classification Criteria				Sub-category	Examples
complete	not a fixed-phrase	have POSs other than interjection	with predicate	with main clause	12 (NONE, ka, ne, ya, yo, kana, kane, ke, nano, kashira, kudasai)	ちょうど良いですよ It's just right.
suspended				without main clause	15 (ka, ga, shi, te, de, to, ni, kamo, kara, kedo, tara, tte, dewa, node, kedomo)	偶然というか Whether it is by chance.
predicate-omitted			without predicate		-	香川が Kagawa?
interjection		only interjection			-	えー Eh
courtesy	fixed-phrase				3 (greeting, thanks, appology)	ありがとうございます Thank you.

Table 3.3: The DA classes used in the evaluation experiments.

Class Name	DA
information	information, self-disclosure_{fact, habit, experience, plan, preference+, preference-, preference0, desire}, sympathy, non-sympathy, approval
question	question_{information, fact, habit, experience, plan, preference, desire}, confirmation, proposal
question_self	question_self
repeat/ paraphrase	repeat, paraphrase
acknowledgement	acknowledgement
filler	filler
admiration	admiration
courtesy	greeting, thanks, appology

3.4.1.1 Classification of sentences

Since the prosodic characteristics of DAs depend on sentence-final morphemes, as described in Sect. 3.2.2, we classify sentences focusing on them. Here, some dialogue sentences have main clause and sentence-final particles such as "ちょうど良いですよ/It's just right" while some do not such as "偶然というか/Coincidence or to say", "香川が/Kagawa?" and "えー/Eh". In order to include such sentences in test sets, we use the sentence classes presented in Table 3.2. We classified sentences with main clause as a "complete" sentence and those without main clause as a "suspended", "predicate-omitted", or "interjection" sentence based on the presence of predicates and parts of speech (POS) other than injections. Moreover, since there are fixed forms of sentences for utterances of "greeting", "thanks", and "apology", they were classified as a "courtesy" sentence. We further classified "complete" sentences into "complete_NONE", "complete_ne", "complete_ka", etc. based on the sentence-final particles. Moreover, "suspended" sentences were also classified into "suspended_kedo", "suspended_noni", "suspended_te" etc based on the final morphemes. Based on this classification, the sentences were classified into 32 classes.

Table 3.4: Frequency of each DA class and sentence class in the text chat corpus [2]

DA class	1st			2nd			3rd			Other		total
	Class name	Number	Ratio	Class name	Number	Ratio	Class name	Number	Ratio	Number	Ratio	
information	complete_NONE	34168	38%	complete_ne	22897	25%	suspended_ga	5745	6%	27368	30%	90178
question	complete_ka	14191	69%	complete_ne	1990	10%	complete_NONE	1531	7%	2860	14%	20572
question_self	complete_NONE	297	21%	complete_kana	261	19%	complete_kane	159	11%	674	48%	1391
repeat/ paraphrase	complete_ne	1111	35%	complete_ka	1001	31%	predicate-omitted	477	15%	606	19%	3195
acknowledgement	interjection	1926	43%	complete_ka	1403	32%	complete_ne	834	19%	269	6%	4432
filler	interjection	596	62%	complete_ne	309	32%	predicate-omitted	27	3%	33	3%	965
admiration	interjection	1624	91%	predicate-omitted	53	3%	complete_ka	49	3%	55	3%	1781
courtesy	greeting	8652	78%	thanks	2091	19%	apology	368	3.3%	0	0%	11111

3.4.1.2 Classification of DAs

First, we assigned a class for each of the DAs with salient global prosodic characteristics ("filler", "admiration", and "question_self") in Sect. 3.2.2.1. The subcategories of "self-disclosure" and "question" were not distinguished in the evaluation. The other DAs were classified on the basis of semantic similarity. Based on the classification, we used the DA classes presented in Table 3.3. Here, we assigned a single class to the DAs with different prosodic characteristics (such as "self-disclosure_preference+" and "self-disclosure_preference-") due to the constraint of the cost of evaluation experiments. Future work can include increasing the size of the test set.

3.4.1.3 Investigation of the utterance frequency of each class

We also investigated frequently used sentence classes. For this investigation, we used the text chat corpus [2]. Table 3.4 presents the investigation results. It is evident that by three most frequently used classes covers utterances of more than 50% of "question_self", 70% of "information" and 80% of the other DAs. Based on the classification above, the number of target DA classes and sentence classes obtained was 24 (8×3). As a test set for objective evaluation, we randomly selected 10 utterances for each of the 24 classes (a total of 240 utterances) from the speech database. As a test set for subjective evaluation, we randomly selected 5 utterances for each of the 24 classes (a total of 120 utterances) from the text chat corpus. For the test set for subjective evaluation, we excluded utterances whose context was difficult to understand from the several preceding utterances.

3.4.2 Experimental conditions

The sampling frequency was 22.05 kHz and the frame shift was 5 ms. We used 40 dimensional mel-cepstral features derived from STRAIGHT [26] analysis. In addition, we used five band aperiodicities.

We compared the following five methods:

- HMM-BASELINE : HMM making use only of conventional linguistic features [29],

- HMM-MIXED : HMM-based style-mixed modeling method [31],
- HMM-ADAPT : HMM-based average model and model adaptation method [12], [44],
- DNN-BASELINE : DNN making use only of conventional linguistic features [14],
- DNN-DACODE : The proposed DNN-based method using DA codes.

In order to compare the proposed method with the conventional methods that consider DAs, we compared HMM-MIXED, HMM-ADAPT, and DNN-DACODE. We also compared DNN-BASELINE and DNN-DACODE in order to validate the effect of utilizing DA information. We also compared HMM-BASELINE and DNN-BASELINE to ascertain the validity of the subjective evaluation method in this study as a method to evaluate the IAN. The evaluation results with respect to naturalness of the dialogue speech can be affected by both the LAN and IAN. In order to focus on evaluating the IAN, it is important to eliminate the possibility of the LAN affecting the evaluation results. Therefore, we instructed the participants not to include voice quality in the evaluation score (Sect. 3.4.4.4). Here, it is considered that the DNN-BASELINE overwhelms the HMM-BASELINE in voice quality due to the model structure. On the other hand, it is considered that there are small differences in terms of the of IAN between these methods, because they are both based on the same speech database and methods and consider only conventional linguistic features. Therefore, by testing whether the evaluation scores for these methods were close, we tested whether the evaluation methods employed in this study were valid for the evaluation of the IAN.

We used 486 dimensional linguistic feature vectors for the duration models DNN-BASELINE and DNN-DACODE. We added frame features to them to be 489-dimensional linguistic feature vectors for acoustic models. The output 139-dimensional vectors comprise mel-cepstra, log F0, and band aperiodicities and their dynamic features and voiced/unvoiced binary variables. The input vectors were normalized to be in the range from 0.01 to 0.99. The output vectors were standard normalized. DNN-based duration models had two hidden layers with 256 units. DNN-based acoustic models had four hidden layers with 256 units. The 30-dimensional DA codes were input to all the hidden layers of DNN-DACODE.

Further, we used a sigmoid activate function for hidden layers and a linear activate function for linear layers.

The weight parameters of DNN were initialized randomly. We used random variables of average 0 and standard deviation $\frac{1}{\sqrt{D}}$, where D is the dimension of input vectors. The initial value of the bias parameters of DNNs were set to 0. The weight and bias parameters were updated to minimize the mean square error between training data and output data using the Adam algorithm [27]-based back propagation. The parameters for the Adam algorithm were set to $\alpha = 0.00015$ for acoustic models, $\alpha = 0.001$ for duration models, and $\beta_1 = 0.9, \beta_2 = 0.999$ and $\epsilon = 1e - 8$ for both models. We used five state left-to-right multispace probability distribution hidden semi-Markov models (MSD-HSMMs) without skip for HMM-BASELINE, HMM-MIXED, and HMM-ADAPT. The 138 dimensional observation vectors comprise 40-dimensional mel-cepstra, log F0, and five band aperiodicities as well as their dynamic features. The state output distribution function was a Gauss function with diagonal covariance. The MDL parameter was set to $\alpha = 1.0$.

For HMM-MIXED, we used questions with respect to DAs as well as linguistic information. We used 30 questions that indicate whether it is a specific DA, as in the conventional method [31]. We added eight questions that indicatewhether it is a specific DA class. The total number of questions with respect to DA was 38.

For HMM-ADAPT, we trained the average voice model using speech data of all DAs. We used speaker adaptive training (SAT) to prevent global prosodic characteristics of DAs from affecting the tree structure. Then, we then adapted the average voice model to each DA using the speech data of a corresponding DA. We used a combination of the CSMAPLR and MAP adaptation [12].

For all the five methods, we used speech of all DAs as training data. In this study, there are numerous utterances (240) in the test set that take into consideration the entire size of the speech database (3410). Therefore, if we exclude the entire test set from the training data, the training data will be substantially reduced. In order to prevent this, this we divided the test set into 10 subsets and trained a model for each subset. We used all utterances other than the concerned subset when training a model. We balanced frequency of DA class and sentence class in each subset.

For DNN-based methods, we used 5% of the training data as the development

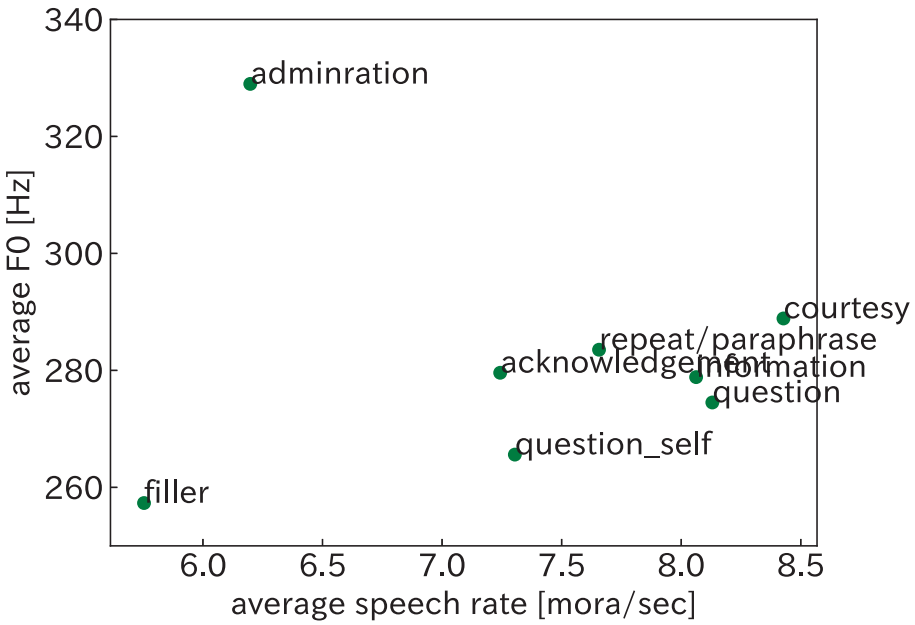


Figure 3.7: Scatter plot of average F0 and average speech rate of natural speech for each DA class.

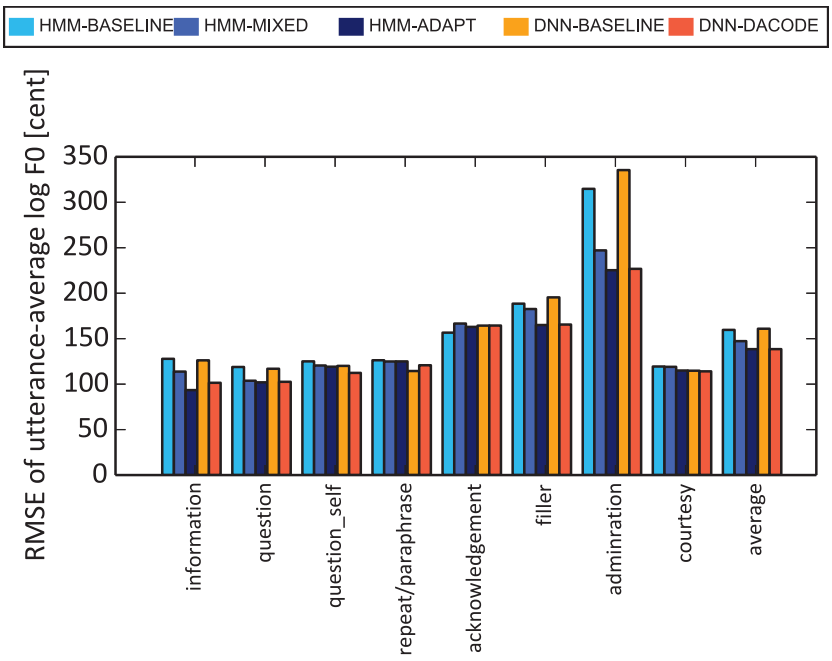
set. For all the five methods, we adopted the maximum likelihood parameter generation (MLPG) [29] to generate acoustic feature sequences from output vectors. For DNN-based methods, we used variances calculated from the training data for MLPG. For all the five methods, we used the Affine transform-based post filter [53] for the mel-cepstra sequence.

3.4.3 Objective evaluation

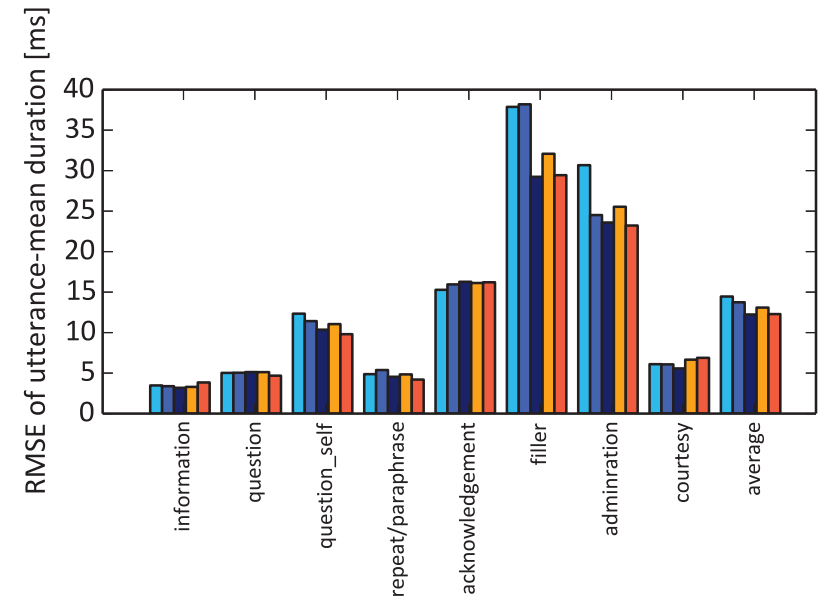
This section conducts objective evaluation experiments to validate the reproducibility of global prosodic characteristics and sentence-final tone characteristics. We used the "RMSE of average log F0" and "RMSE of average duration" as well as conventional measures, "RMSE of log F0" and "RMSE of duration". "RMSEs of average log F0/duration", $D_{utterance}$, was calculated using a value averaged over an utterance,

$$D_{utterance} = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(\bar{x}_n - \bar{y}_n)^2}. \tag{3.2}$$

Here, we denoted the number of utterances in the test set as N, the utterance



(a) RMSE of average log F0



(b) RMSE of average duration

Figure 3.8: RMSE of average F0 and average duration.

id as n $(= 1, \cdots, N)$, average features for each utterance of synthesized speech (average log F0 or average duration) as $\bar{x}_n$, and the average features for each

utterance of natural speech as $\bar{y}_n$. We used $D_{utterance}$ to evaluate reproducibility of global prosodic characteristics. Further, we used correlation of sentence-final F0s (F0s of the last syllable) as a measure of reproducibility of the sentence-final tone. Since the correlation of sentence-final F0s depend not only on DAs but also sentence-final morphemes, the results are presented for each DA and sentence class.

3.4.3.1 RMSE of average log F0 and average duration.

In order to reveal global prosodic characteristics of DA classes, we first show average F0 and average speech rate for each DA class in Fig. 3.7. We can see clear trends that

- question_self, filler : lower F0, lower speech rate
- admiration : higher F0, lower speech rate
- acknowledgement : lower speech rate.

These trends are similar to those presented in Fig. 3.2. This section focuses on these DA classes in order to identify whether or not their tendencies were reproduced.

Fig. 3.8 presents the RMSE of average log F0 and average duration. First, we compared HMM-MIXED, HMM-ADAPT, and DNN-DACODE to confirm whether the proposed method was superior to conventional methods. We found that

- HMM-MIXED > HMM-ADAPT $\fallingdotseq$ DNN-DACODE

for RMSEs of average log F0 and average duration. Here, a difference larger than 7.0 cent for average F0 and 1.0 ms for average duration was denoted by inequality, and a smaller difference than these values was denoted by $\fallingdotseq$. When compared the results for each DA class, it is evident that the RMSE of average F0 was

- HMM-MIXED $\fallingdotseq$ HMM-ADAPT > DNN-DACODE (question_self)
- HMM-MIXED > HMM-ADAPT $\fallingdotseq$ DNN-DACODE (filler, admiration)

and the RMSE of average duration was

- HMM-MIXED > HMM-ADAPT ≒ DNN-DACODE (question_self, filler, admiration) .

The results showed that the proposed method had smaller RMSEs than HMM-MIXED and comparable RMSEs to HMM-ADAPT. It is likely that global prosodic characteristics were not reproduced because there were cases in which the questions with respect to DAs did not appear in the decision trees.

Thereafter, we compared the DNN-BASELINE and DNN-DACODE to validate the effect of considering DAs. It is evident from the RMSE that the average log F0 and average duration for all DA classes was

- DNN-BASELINE > DNN-DACODE
  (question_self, filler, admiration) .

The results show that the proposed method was superior to DNN-BASELINE in terms of reproducing global prosodic characteristics. These results show the effect of taking DAs into consideration.

In summary, the results in this section indicate that the proposed method had better reproducibility of global prosodic characteristics of DAs than HMM-MIXED and was comparable to HMM-ADAPT. It was also shown that taking DAs into consideration, as done by the proposed method, was effective from the perspective of reproducibility of average F0 and average duration.

3.4.3.2 Evaluation results of RMSE of log F0 and duration

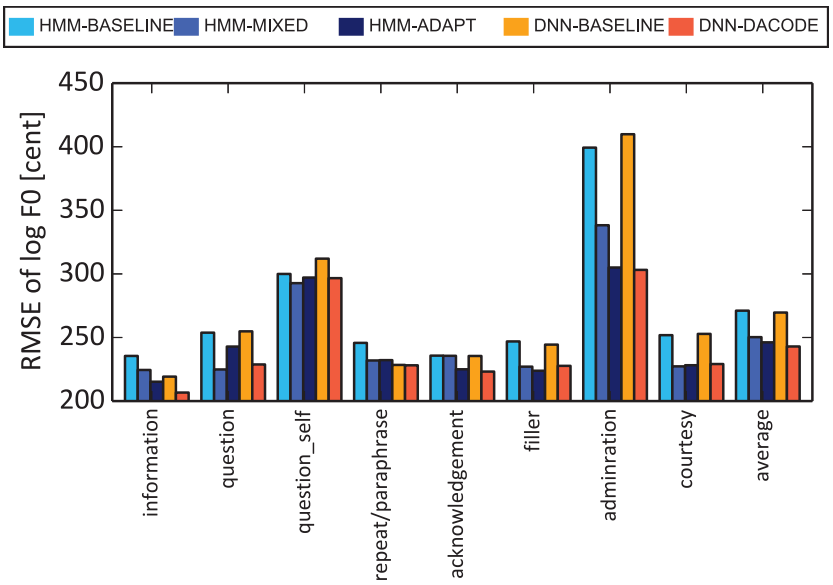
Fig. 3.9 presents the RMSE of log F0 and duration. First, we compare HMM-MIXED, HMM-ADAPT, and DNN-DACODE to confirm whether the proposed method was superior to the conventional methods. It is evident that the RMSE of log F0 averaged over all the DA classes was

- HMM-MIXED > HMM-ADAPT ≒ DNN-DACODE

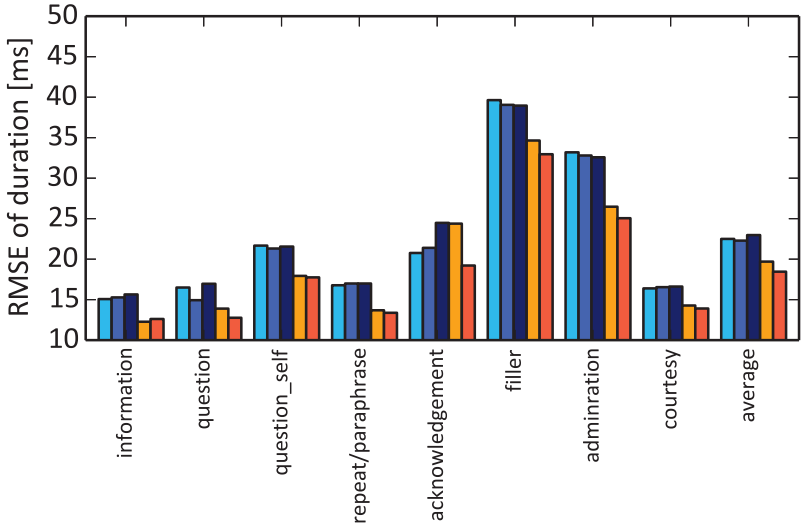
and the RMSE of duration was

- HMM-MIXED ≒ HMM-ADAPT > DNN-DACODE.

Here, a difference of greater than 7.0 cent for average F0 and 1.0 ms for average duration was denoted by inequality, and a smaller difference than these was denoted



(a) RMSE of log F0



(b) RMSE of duration

Figure 3.9: RMSE of log F0 and duration

by $\fallingdotseq$, as done in the previous subsection. The results reveal the superiority of the proposed method as compared to the conventional methods in terms of the RMSE.

Further, we compare DNN-BASELINE and DNN-DACODE to confirm the effect of considering DAs by the proposed method. It is evident that the RMSE of F0 was

- DNN-BAESLINE > DNN-DACODE

except for F0 in "repeat/paraphrase". Thus, the results reveal the effect of considering DAs by the proposed method in terms of the RMSE of F0 and duration.

The results in this section indicate the superiority of the proposed method as compared to HMM-MIXED and HMM-ADAPT with respect to the generation error of F0 and duration. They also show the superiority of the proposed method to DNN-BASELINE in terms of generation error. This demonstrates the effect of considering DAs by the proposed method in terms of generation error.

The results in Fig. 3.4.3.1 and Fig. 3.4.3.2 indicate that the proposed method ensures greater reproducibility of global prosodic characteristics than HMM-MIXED. The proposed method was comparable to HMM-ADAPT in terms of the RMSE of average log F0 and RMSE of log F0. While the RMSEs of duration improved, the RMSEs of average duration were comparable. Therefore, we conclude that the proposed method had comparable reproducibility of global prosodic characteristics to the HMM-ADAPT in terms of duration. Moreover, the proposed method was superior to DNN-BASELINE, which indicates the effect of taking DAs into consideration by the proposed method in terms of the reproducibility of global prosodic characteristics.

3.4.3.3 Evaluation results of correlation of sentence-final F0s

Table 3.5 presents the correlations of sentence-final F0s of synthesized speech and natural speech. First, we compare HMM-MIXED, HMM-ADAPT, and DNN-DACODE to confirm whether the proposed method was superior to conventional methods. It is evident that the averaged correlation over the test set was

- HMM-ADAPT < HMM-MIXED $\fallingdotseq$ DNN-DACODE.

Here, a difference larger than 0.05 was denoted by inequality, and a smaller difference than that was denoted by $\fallingdotseq$. The results reveal that the proposed method has greater reproducibility of sentence-final F0 than HMM-ADAPT and is comparable to HMM-MIXED. Moroever, HMM-ADAPT did not reproduce sentence-final tones because the tree structure was not optimal to reproduce them.

When we compare the results for each DA class and sentence class, the correlation of the proposed method was improved by more than 0.2 compared with that

Table 3.5: Correlation of sentence-final F0s

DA class	sentence class					
	class 1		class 2		class 3	
information	0.56	(+0.06)	0.96	(+0.01)	0.79	(+0.04)
question	0.58	(+0.07)	0.78	(+0.00)	**0.86**	**(+0.74)**
question_self	0.41	(-0.01)	0.74	(+0.09)	0.71	(+0.01)
repeat/ paraphrase	0.95	(+0.05)	**0.72**	**(+0.58)**	**0.77**	**(+0.38)**
acknowledgement	0.79	(-0.06)	**0.81**	**(+0.93)**	0.88	(+0.01)
filler	0.69	(-0.02)	**0.76**	**(+0.46)**	0.80	(+0.29)
admiration	0.62	(-0.18)	0.33	(+0.10)	**0.25**	**(+0.90)**
courtesy	0.70	(+0.09)	0.88	(+0.25)	**0.83**	**(+0.39)**
average			0.72	(+0.22)		

(a) HMM-MIXED (difference from HMM-BASELINE)

DA class	sentence class					
	class 1		class 2		class 3	
information	0.68	(+0.18)	0.96	(+0.01)	0.78	(+0.03)
question	0.46	(-0.06)	0.83	(+0.05)	0.35	(+0.23)
question_self	0.41	(-0.01)	0.58	(-0.06)	0.78	(+0.08)
repeat/ paraphrase	0.95	(+0.04)	**0.87**	**(+0.73)**	0.64	(+0.25)
acknowledgement	0.86	(+0.01)	**0.56**	**(+0.67)**	0.92	(+0.05)
filler	0.67	(-0.04)	**0.68**	**(+0.38)**	0.77	(+0.26)
admiration	0.69	(-0.11)	0.43	(+0.21)	**-0.27**	**(+0.38)**
courtesy	0.73	(+0.13)	0.90	(+0.27)	**0.85**	**(+0.41)**
average			0.67	(+0.17)		

(b) HMM-ADAPT (difference from HMM-BASELINE)

DA class	sentence class					
	class 1		class 2		class 3	
information	0.60	(-0.00)	0.94	(+0.02)	**0.77**	**(+0.32)**
question	0.64	(-0.01)	0.89	(+0.06)	**0.57**	**(+0.40)**
question_self	0.56	(+0.11)	0.61	(+0.16)	0.76	(+0.05)
repeat/ paraphrase	0.96	(+0.04)	**0.93**	**(+0.52)**	0.73	(+0.11)
acknowledgement	0.81	(+0.02)	**0.92**	**(+1.07)**	0.91	(-0.03)
filler	0.65	(-0.12)	**0.61**	**(+0.37)**	0.71	(+0.28)
admiration	0.82	(+0.07)	0.37	(-0.00)	**0.18**	**(+0.79)**
courtesy	0.72	(+0.03)	0.93	(+0.01)	0.72	(+0.16)
average			0.73	(+0.19)		

(c) DNN-DA (difference from DNN-BASELINE)

for HMM-ADAPT for

- question-3 (complete_NONE)
- acknowledgement-2, admiration-3 (complete_ka).

Here, we denote "the third sentence class of question" as "question-3" for brevity. It is expected that the IAN for these classes of the proposed method will be improved in comparison to HMM-ADAPT.

We then compare DNN-DACODE and DNN-BASELINE. The correlation of sentence-final F0s were improved by over 0.3 for

- question-3 (complete_NONE)
- filler-2 (complete_ne)
- repeat/paraphrase-2, acknowledgment-2, admiration-3 (complete_ka) .

It is expected that the proposed method will yield better IAN as compared to that yielded by DNN-BASELINE.

The results in this section showed that the reproducibility of sentence-final F0s by the proposed method was superior to that by HMM-ADAPT and is comparable to HMM-MXIED. The proposed method was also superior to DNN-BASELINE, which shows the effect of considering DAs by the proposed method in terms of reproducibility of sentence-final F0.

The objective evaluation results in Sect. 3.4.3.1, Sect. 3.4.3.2 and Sect. 3.4.3.3 make it evident that the proposed method provides greater reproducibility of global prosodic characteristics and comparable reproducibility of sentence-final tone characteristics as compared to that yielded by HMM-MIXED. On the other hand, the proposed method has comparable reproducibility of global prosodic characteristics and greater reproducibility of sentence-final tone compared to that yielded by HMM-ADAPT. In addition, the proposed method showed greater reproducibility of global prosodic characteristics and sentence-final tones compared with DNN-BASELINE, which shows the effect of considering DAs using the proposed method.

3.4.4 Subjective evaluations

3.4.4.1 Evaluation method of IAN

We also conducted subjective evaluations in order to evaluate the IAN of the synthesized speech. Here, intentions are inferred by considering contexts, sentence,

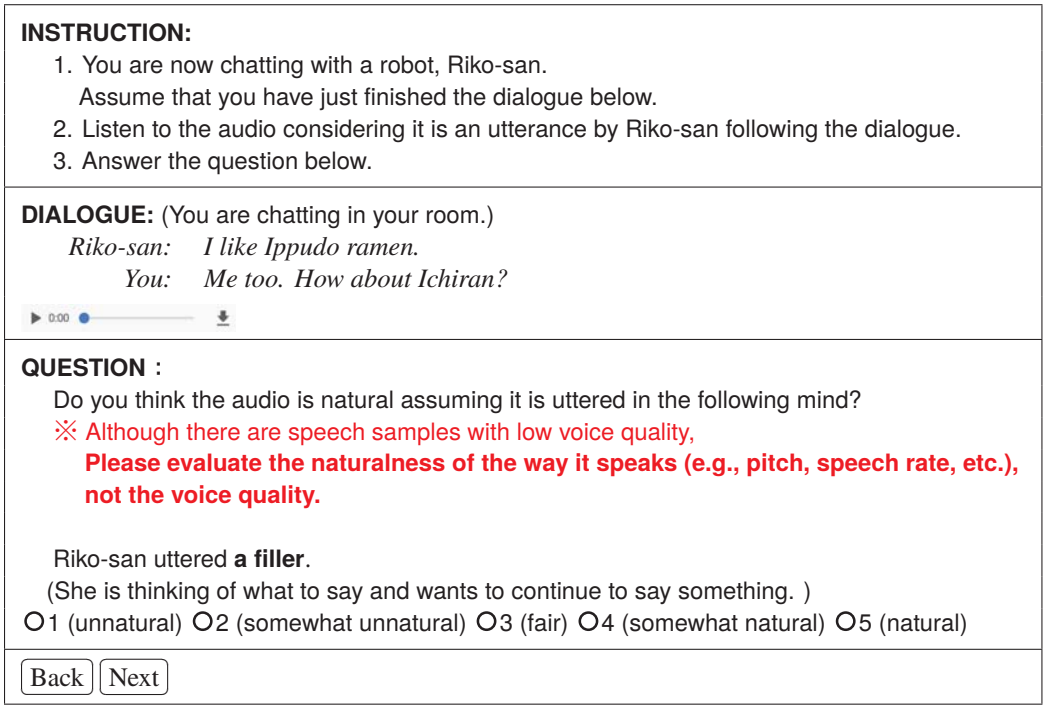


Figure 3.10: An example of GUI used in the subjective evaluation experiments (the sentence of the speech sample "so desune.") .

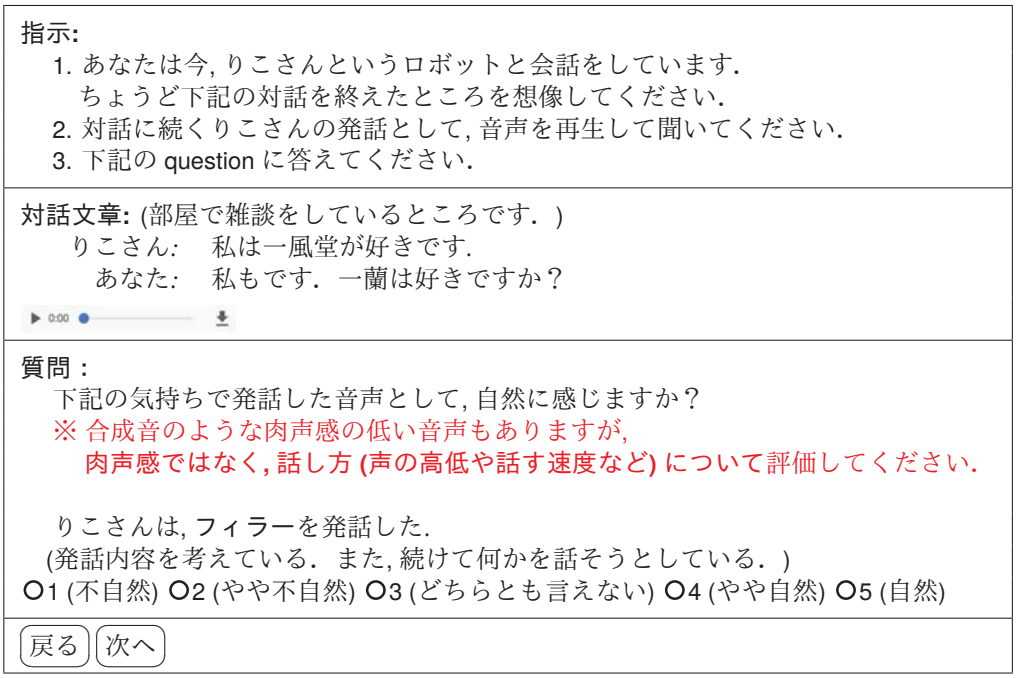


Figure 3.11: An example of GUI used in the subjective evaluations (in Japanese).

Table 3.6: MOS results for the entire test set with 99% confidence interval.

Method	MOS
HMM-BASELINE	3.01 ± 0.07
HMM-MIXED	3.36 ± 0.07
HMM-ADAPT	3.48 ± 0.07
DNN-BASELINE	3.02 ± 0.07
DNN-DACODE	**3.63** ± 0.06

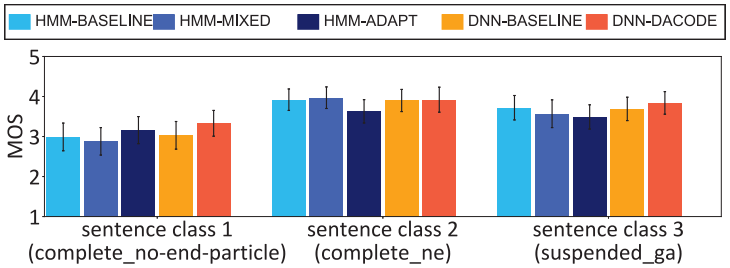
and prosody. Therefore, showing the conversational context is considered necessary for this evaluation. In order to show the context, we used GUI in Fig. 3.11. We first instructed participants to assume that they are chatting with a robot and then read the displayed context. Then, they listened to an audio sample, being provided a context, assuming that it is an utterance by the robot. Finally, they evaluated the samples in terms of whether it was natural considering it was uttered in the instructed intention. If we described the reference intention only by a single word (e.g. "filler"), the reliability of the evaluation may be degraded because participants may not be able to understand its definition [54]. Therefore, we displayed the description of the definition of the reference intention as well. We also displayed the instruction "Please evaluate the naturalness of the way it speaks (e.g. pitch and speech rate etc.), not the voice quality," in order to ensure that the evaluation is not affected by the voice quality. We also instructed them to evaluate the naturalness of the way it speaks, not that of the prompt (sentence).

We used 5 utterances for each DA class and sentence class (a total of 120 utterances in total). We used several preceding utterances as conversational context, which were extracted from the original text chat corpus. A total of 20 Japanese (10 males and 10 females) with normal hearing participated in this evaluation. They listened to the audio using headphones in a soundproof room. There was no time constraint for the evaluation. We conducted a five-point mean opinion score (MOS) test for the IAN. (1 : unnatural, 2 : somewhat unnatural, 3 : fair, 4 : somewhat natural, 5 : natural). Speech samples were displayed in random order.

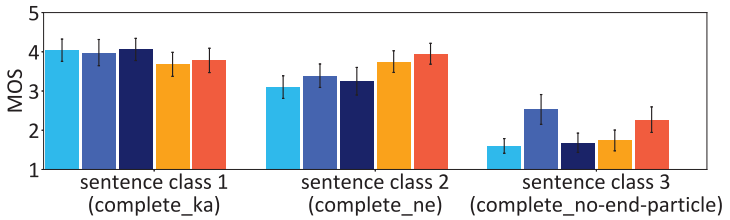
Table 3.6 presents the subjective evaluation results of the entire test set. Fig. 3.13 presents the subjective evaluation results for each DA class and sentence class. Table 3.7 presents t-test results for each DA class and sentence class. Here, statistical tests with 5% significance level are likely to result in false positives because they are multiple tests. Therefore, we conducted statistical tests with

Table 3.7: The results of t-test for MOS value. ($\alpha = 0.01$) . The following abbreviations were used: HMM-B : HMM-BASELINE HMM-M : HMM-MIXED HMM-A : HMM-ADAPT DNN-B : DNN-BASELINE DNN-D : DNN-DACODE a) : Comparison between the baselines (HMM-B and DNN-B). b) : Comparison among methods considering DAs (HMM-M and DNN-D, HMM-A, and DNN-D). c) : Comparison with or without DAs (DNN-B and DNN-D).

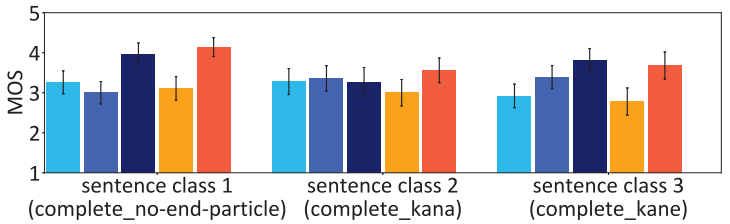
DA	**sentence class**		
class	**class 1**	**class 2**	**class 3**
information	-	-	-
question	-	a) HMM-B < DNN-B b) HMM-M < DNN-D b) HMM-A < DNN-D	b) HMM-A < DNN-D c) DNN-B < DNN-D
question_self	b) HMM-M < DNN-D c) DNN-B < DNN-D	c) DNN-B < DNN-D	c) DNN-B < DNN-D
repeat/paraphrase	-	c) DNN-B < DNN-D	-
acknowledgement	-	b) HMM-A < DNN-D c) DNN-B < DNN-D	-
filler	b) HMM-M < DNN-D c) DNN-B < DNN-D	b) HMM-M < DNN-D c) DNN-B < DNN-D	b) HMM-M < DNN-D c) DNN-B < DNN-D
admiration	c) DNN-B < DNN-D	c) DNN-B < DNN-D	b) HMM-M < DNN-D b) HMM-A < DNN-D c) DNN-B < DNN-D
courtesy	b) HMM-M < DNN-D	b) HMM-M < DNN-D c) DNN-B < DNN-D	b) HMM-M < DNN-D



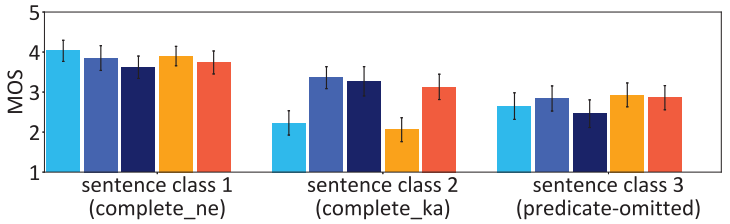
(a) information



(b) question



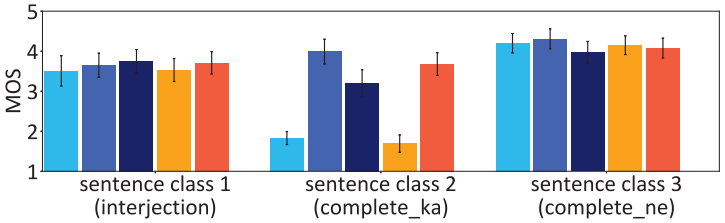
(c) question_self



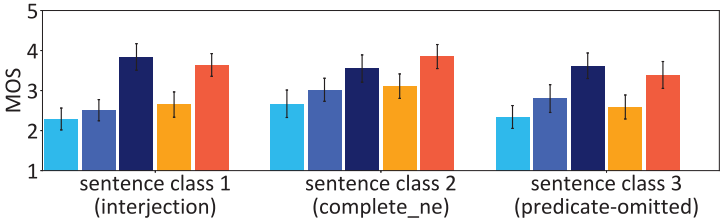
(d) repeat·paraphrase

Figure 3.12: Subjective evaluation results with respect to the IAN. Error bars display 99% confidence intervals.

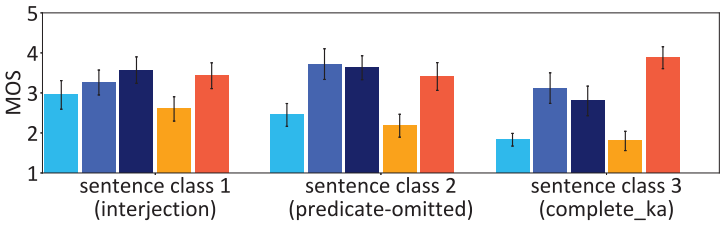
1% significance level and compared the results with objective evaluation results in order to check their validity. The results in Table 3.7 are presented for a) : comparison between the baselines (HMM-B and DNN-B), b) : comparison among methods considering DAs (HMM-M and DNN-D, HMM-A, and DNN-D), c) : comparison with or without DAs (DNN-B and DNN-D).



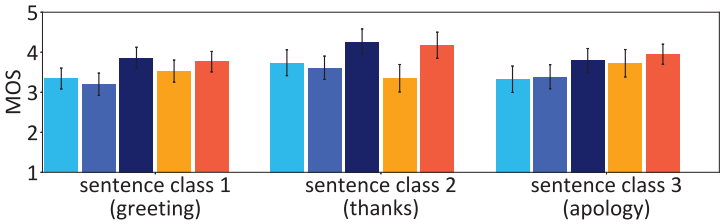
(a) acknowledgement



(b) filler



(c) admiration



(d) courtesy

Figure 3.13: Subjective evaluation results with respect to the IAN. Error bars display 99% confidence intervals.

Table 3.8: Analysis of correlation between MOS scores and number of mora in a sentence for "information" and "question-1" sentences.

Method	Correlation	p-value
HMM-BASELINE	-0.29	0.21
HMM-MIXED	-0.33	0.16
HMM-ADAPT	-0.55	0.01
DNN-BASELINE	0.10	0.69
DNN-DACODE	-0.14	0.55

3.4.4.2 Comparison of the scores averaged over the test set

It is evident from Table 3.6 that the MOS scores averaged over the test set were

- HMM-BASELINE $\fallingdotseq$ DNN-BASELINE < HMM-MIXED < HMM-ADAPT < DNN-DACODE.

There were significant differences at the 1% significance level for all of the combinations, except HMM-BASELINE and DNN-BASELINE. Here, there were no significant differences between the baselines, although their voice qualities likely differed. This shows that the validity of the questions and instructions in this experiment suppresses the effect of voice quality on MOS scores. Therefore, improvement of MOS scores by the proposed method is due to the improvement in the IAN and not in voice quality. Thus, these results indicate that the proposed method improves the IAN more as compared to HMM-MIXED and HMM-ADAPT.

3.4.4.3 Comparison of the baselines

Here, it is evident that the MOS score of DNN-BASELINE is higher than that of HMM-BASELINE in "question-2". It is considered that this difference is not due to the improvement in the IAN but the superiority of the DNNs in generating speech from long sentences. We present the correlation between MOS scores and the number of mora in a sentence in Table 3.8. We used 20 sentences from "information" and "question-1", which were considered as being barely affected by DAs. The results indicate that there are negative correlations for HMM-based methods and no significant correlations for DNN-based methods. Further, it is also evident that there are longer sentences for "question-2", as presented in Table 3.10. From the above results, it is evident that the improvement in "question-2" was not

due to an improvement in the IAN but due to a general improvement in generating speech from longer sentences, which prominently appeared in "question-2". In other words, this improvement originated from the design of the test set. It is considered that this tendency does not degrade the validity of the evaluation results because there were no significant differences between DNN-BASELINE and HMM-BASELINE for any of the classes other than "question-2". Future work can include designing the test set by controlling the length of the sentences as well as the sentence-final particle.

3.4.4.4 Comparison among methods that consider DAs

In order to confirm the superiority of the proposed method over conventional methods, we compared HMM-MIXED, HMM-ADAPT, and DNN-DACODE. First, we compared HMM-MIXED and HMM-DACODE. We found that DNN-DACODE shows comparable or higher naturalness for all the DA classes and sentence classes. This shows the superiority of the proposed method over HMM-MIXED. Further, the proposed method showed higher naturalness for

1. question-2
2. question_self-1, filler-{1, 2, 3}, admiration-3
3. courtesy-{1, 2, 3}.

Here, it is considered that (1) was due to the improvement in generating speech from longer sentences, as described in Sect. 3.4.4.3. It is also considered that (2) was due to the improvement in the reproducibility of global prosodic characteristics for "question_self", "filler", and "admiration", as described in Sect. 3.4.3.1. On the other hand, there was no significant improvement in

(i) question_self-{2, 3}

(ii) admiration-{1, 2}.

For (i), there was a slight tendency that DNN-DACODE improves the scores, although it was not statistically significant. For (ii), there was no consistent tendency between the two methods. When we focused on the difference of sentence

forms, we noticed that intention of "admiration" can be inferred only from the sentence of "admiration-1 (interjection-sentence)" and "admiration-2 (predicate-omitted-sentence)". On the other hand, we require prosodic characteristics to infer "admiration" for sentences of "admiration-3", because they can be inferred as other DAs, like "question". This implies less need of reproducibility of global prosodic characteristics for "admiration-{1,2}". It is also considered that DNN-DACODE can improve naturalness if we specify the degree of emotion expression to be evaluated. Future research can include evaluation that specifies the degree of emotion expression.

For (3), there was no clear correspondence with objective evaluation results. Future research can also include an analysis of the IAN of courtesy sentences.

The results above indicate the superiority of the proposed method to HMM-MIXED in terms of IAN. The correspondence with objective evaluation results show that this is due to the improved reproducibility of global prosodic characteristics.

We then compared DNN-DACODE with HMM-ADAPT. It was evident that DNN-DACODE showed comparable or higher MOS scores for all the DA classes and sentence classes compared with HMM-ADAPT. The results indicate the superiority of the proposed method over HMM-ADAPT. The MOS scores were improved in

1. question-2
2. question-3, acknowledgement-2, admiration-3.

Here, it is considered that (1) was due to the improvement in generating speech from longer sentences, as described in Sect. 3.4.4.3. In addition, (2) corresponds to the classes where correlations of sentence-final F0s were improved by over 0.2 by the proposed method in Sect. 3.4.3.3. It is considered that the IAN of (2) was improved due to the improved correlation of sentence-final F0s.

These results indicate that the IAN obtained by the proposed method was comparable or superior to HMM-ADAPT for all the DA classes and sentence classes. The correspondence with objective evaluation results implies that this is due to the improved reproducibility of sentence-final tones.

The results in this section showed that the IAN obtained by the proposed method was comparable to HMM-MIXED and superior to HMM-ADAPT for all the DA

classes and sentence classes. Further, the correspondence with the objective evaluation results implies that these were due to the improved reproducibility of global prosodic characteristics compared to HMM-MIXED and the improved reproducibility of sentence-final tones compared to HMM-ADAPT.

3.4.4.5 Comparison with or without DAs

We compared DNN-BASELINE and DNN-DACODE to validate the effect of considering DAs using the proposed methods. It is evident from Table 3.7 that MOS scores by DNN-DACODE were comparable or higher than DNN-BASELINE for all the DA classes and sentence classes. These show that DNN-based speech synthesis considering DAs improves the IAN. It is also evident that naturalness was improved in

1. question_self-{1, 2, 3}, filler-{1, 2, 3}, admiration-{1, 2, 3}
2. question-3, repeat/paraphrase-2, acknowledgement-2
3. courtesy-2.

Improvement in (1) is due to reproducing the characteristics of average F0 and average duration by DNN-DACODE, as described in Sect. 3.4.3.1. For these DA classes, it is considered that naturalness was improved due to improved reproducibility of global prosodic characteristics. As shown in Sect. 3.4.3.3, correlations of sentence-final F0s were improved by over 0.3 by DNN-DACODE for "filler-2", "admiration-3" of (1) and all of (2). Further, the IAN for these classes improved due to improved reproducibility of sentence-final tones. For (3), there were no clear correspondence with objective evaluation results. Future research can include an analysis of factors that contribute to improving the IAN of courtesy sentences.

The results in Sect. 3.4.4 reveal that the proposed method improved the IAN more compared to the conventional HMM-based methods. The correspondence of subjective and objective evaluation results implies that improvement from HMM-MIXED was due to improved reproducibility of global prosodic characteristics. It also implies that improvement from HMM-ADAPT was due to improved reproducibility of sentence-final tones. The proposed method also showed higher IAN

compared with DNN-BASELINE. Thus, the IAN improves by considering DAs using the proposed method.

3.5 Conclusion

This chapter investigated the speech synthesis method that can improve the IAN. Based on the DA tag set designed for non-task oriented dialogue systems, we first constructed a speech database with DA tags. Then, we proposed DNN-based speech synthesis considering DAs and compared it with the two HMM-based conventional methods (the style-mixed model and the model adaptation method). The subjective evaluation results reveal that the proposed method improves the IAN more as compared to the HMM-based conventional methods. Further, the correspondence of subjective and objective evaluation results implies that improvement from HMM-MIXED was due to improved reproducibility of global prosodic characteristics. It also implies that improvement from HMM-ADAPT was due to improved reproducibility of sentence-final tones. The proposed method also showed higher IAN compared with the conventional DNN-based speech synthesis without considering DAs. Therefore, there is an improvement in the IAN by considering DAs using the proposed method. Future research can include the evaluation of "admiration" speech by specifying the degree of emotional expressions.

Table 3.9: Sentences and DAs used in the subjective evaluation results.

DA class	sentence class	sentence
information	complete_NONE	けど, 好物なためかうどんはよく作ります (self-disclosure_habit) ただ辛いだけかなという気もしました (self-disclosure_preference0) 掃除機だけ嫌いという感じです (self-disclosure_preference-) 行ったのは一回だけですが, そのサービスは利用しました (self-disclosure_experience) 今度試してみます (self-disclosure_plan)
	complete_ne	飲み会はやりますけど, バイトではやったことありませんね (self-disclosure_experience) 冬なんかはだいたい温かい日本茶ですね (self-disclosure_habit) アクティブですね (approval) 私は海外だったら, 取り合えず英語圏で食べ物がおいしいところに行きたいですね (self-disclosure_desire) それは実用的ですね (self-disclosure_preference+)
	suspended_ga	全然無理でしたが (self-disclosure_preference-) 米沢牛でもいろいろ種類があったんですが (information) 対戦が面白いんですが (self-disclosure_preference+) 特定のアーティストが好きというよりは, その時々で気に入った曲を聴くという感じなんですが (non-sympathy) キャンプ場のアルバイトだったのですが (self-disclosure_fact)
question	complete_ka	紅葉とか見に行く plan はありますか (question_plan) お笑いの話はいかがでしょうか (proposal) 他に気になる水族館や動物園はありますか (question_desire) 魚好きですか (question_preference) どんな絵なんですか (question_information 要求)
	complete_ne	では東京に詳しいのですね (question_fact) じゃあトレーニングも本格的にやって, 体をつくらないといけないですね (question_plan) モンブランとか買って食べたいですね (question_desire) なんで導入しないんでしょうね (question_information 要求) そのうえお子さんがいらっしゃると, 生活がかなり制限されますね (question_habit)
	complete_no sentence-final particle	どこか行きたいところあります (question_desire) ご飯とかどうする plan です (question_plan) 今年の夏は夏らしい事しました (question_experience) カフェはどこによく行きます (question_habit) 他に以前読んだおもしろかったラブコメはあったりします (question_preference)
question_self	complete_no sentence-final particle	5 兆なくなっちゃいそうです (question_self) ルンバでしたっけ (question_self) だれだろう (question_self) どうでしょう (question_self) なんでしょう (question_self)
	complete_kana	今日はこれから降るかな (question_self) それかバターかな (question_self) 次の日だったかな (question_self) あとは何があったかな (question_self) 近くのスポーツクラブは, 確か 22 時とか 23 時くらいまでだったかな (question_self)
	complete_kane	要は相性ですかね (question_self) そうなんですかね (question_self) 子供ってなんであんなずーっと集中して観れるんですかね (question_self) このままだとすぐに冬ですかね (question_self) 設置アンテナの数が電話の数に対応出来るのでしょうかね (question_self)
repeat/paraphrase	complete_ne	スパルタですね (paraphrase) ブラッディマリーですね (repeat) ビタミンですね (repeat) 呑み屋をやってらしたんですね (repeat) 海は見てるだけで涼しくなりそうですね (paraphrase)
	complete_ka	完全におもちゃ扱いではないですか (paraphrase) キングコングでしたか (repeat) 公園の木ですか (repeat) そんなにするんですか (paraphrase) ロックですか (repeat)
	predicate-omitted	妊婦さんはお金かかりそう (paraphrase) 九州 (repeat) 見た目すごい赤そう (paraphrase) ワカメは日本と韓国だけ (repeat) そんな優しいところ (paraphrase)

Table 3.10: Sentences and DAs used in the subjective evaluation results.

DA class	sentence class	sentence
acknowledgment	interjection	あぁなるほど (acknowledgment)
		あー (acknowledgment)
		ねえ (acknowledgment)
		うーん (acknowledgment)
		おっ (acknowledgment)
	complete_ka	本当ですか (acknowledgment)
		そうだったのですか (acknowledgment)
		あそうでしたか (acknowledgment)
		あっそうなんですか (acknowledgment)
		そんなものがあるんですか (acknowledgment)
	complete_ne	どうなんでしょうかね (acknowledgment)
		なるほどそうなんですね (acknowledgment)
		そうですね (acknowledgment)
		そうなのですね (acknowledgment)
		いろいろあるんですね (acknowledgment)
filler	interjection	あー (filler)
		いやー (filler)
		ううん (filler)
		うーむ (filler)
		いやぁ (filler)
	complete_ne	うーん悩みますね (filler)
		うーんどうでしょうね (filler)
		どうなんでしょうね (filler)
		なんかですね (filler)
		んーそうですね (filler)
	predicate-omitted	ランチは (filler)
		普通に知名度があるのは (filler)
		なんか (filler)
		何の (filler)
		その (filler)
admiration	interjection	あら (admiration)
		お (admiration)
		ありゃりゃ (admiration)
		あはは は (admiration)
		まあ (admiration)
	predicate-omitted	びっくり (admiration)
		それそれ (admiration)
		香川が (admiration)
		そんなに (admiration)
		無理ゲー (admiration)
	complete_ka	そうなんですか (admiration)
		買ったんですか (admiration)
		そうですか (admiration)
		駅二つですか (admiration)
		黒いんですか (admiration)
court	greeting	おはようございます (greeting)
		ではここで (greeting)
		まあ後ほど (greeting)
		こちらこそ (greeting)
		今日も宜しくお願いします (greeting)
	thanks	それではありがとうございました (thanks)
		いろいろ情報ありがとうございます (thanks)
		はい 4 日間ありがとうございました (thanks)
		ありがとうございましたまた後ほど (thanks)
		本日はありがとうございました (thanks)
	appology	何か話題変わってすみません (appology)
		すみませんぞろぞろと (appology)
		名前間違えて失礼しました (appology)
		ちょっと語弊があったようで申し訳ないです (appology)
		すいません, わかりにくくて (appology)

Chapter 4

Estimation of sentence-final tone labels using dialogue-act information

(This chapter is not published in this outline because it is scheduled to be published by another journal.)

Chapter 5

Conclusions and future work

5.1 Summary of the thesis

In this thesis, we provided a novel approach to text-to-speech (TTS) synthesis systems used for a spoken dialogue system. The research objective was to improve the naturalness of synthesized speech with respect to "how natural the system's intention is perceived via the synthesized speech". We call this measure for TTS "illocutionary act naturalness (IAN)" in this thesis.

In order to achieve this aim, in Chap. 2, we investigated a technique that can improve quality of synthesized speech with a limited amount of training data for each speaker. The proposed method utilizes speaker codes in the DNN-based speech synthesis framework to improve speech quality by using a multi-speaker speech corpus. Objective and subjective evaluation was conducted to compare the proposed method with the conventional methods. The results of objective evaluation showed that the proposed model outperforms the speaker-dependent DNNs and multi-speaker DNNs with shared hidden layers for both multi-speaker modeling and speaker adaptation tasks. The subjective evaluation results showed that the proposed model can produce more natural-sounding speech than the conventional speaker dependent DNNs and HMM-based speaker adaptation method, particularly when using a small number of target speaker utterances.

In Chap. 3, the proposed model is applied to TTS considering intentions and evaluated with respect to IAN. Based on the DA tag set designed for non-task oriented dialogue systems, we first construct speech database with DA tags. The

proposed method is based on that in Chapter 2, which uses DA codes instead of speaker codes. The proposed method was compared with the two HMM-based conventional methods (the style-mixed model and the model adaptation method). The results of the subjective evaluation revealed that the proposed method improves the IAN compared with the HMM-based conventional methods. In addition, a comparison of the subjective and objective evaluation results revealed that the improvement of the proposed method over the HMM-based style-mixed model method is due to the improved reproducibility of global prosodic characteristics. We also showed that the improvement of the proposed method over the HMM-based model adaptation method is due to the improved reproducibility of sentence-final tones. Further, the proposed method also showed a higher IAN compared with the conventional DNN-based speech synthesis that considers only linguistic information. Thus, this results showed that IAN can be improved by considering DAs information.

In Chap. 4, the sentence-final tone label estimation is focused to precisely investigate the efficacy and difficulty of the approach to utilize DA for speech synthesis. The speech database constructed in Chap. 3 of this thesis was annotated with sentence final tone labels for this investigation. The proposed method utilizes DAs as well as conventional morpheme and parts-of-speech information derived from an utterance to estimate a sentence final tone label. The objective evaluation revealed that the proposed method improves Cohen's κ by 0.12. This result showed the effectiveness to consider dialogue act information in the proposed method. Analysis results showed that the proposed method was effective for expressing three types of intentions of the system; "self-question and consent", "interrogative by rising intonation" and "self-questions/fillers". We also analyzed incorrect estimation results by the proposed method. The results showed that more detailed DA tags such as "self- disclosure with/without expressing a speaker's emotion" or "repeat with/without expressing surprise" were effective to improve estimation accuracy.

5.2 Future work

As for DNN-based speech synthesis using speaker codes, multi-speaker modeling techniques has been progressed after our investigation, such as utilizing sequence modeling [55], "multi-modal architecture" [56], etc. Since our focus was to improve IAN, we didn't further investigated this point. These models are expected to further improve IAN.

As for the evaluation of IAN, the size of the evaluation set in our experiments was limited due to the cost. There are several points to be investigated by future work. Firstly, we didn't investigate dependency on sentences enough. While only sentence-end forms were considered in our study, there can be other dependencies on sentence forms, such as length of the sentence, its person, tense and presence of interrogatives. Secondly, we need to control and evaluate the degree of emotional expression, such as one for "admiration", "question_self" and "filler". Finally, dependency on the contexts should be further investigated. While contexts were designated by only texts in our experiments, there can be other dependencies on contexts, such as prosody of the preceding the user's and the system's utterances.

As for the sentence final tone estimation, the experiments were based on a single speech corpus. However, sentence-end forms can differ depending on the social relationships of the participants and their characters [57]. Experiments should be conducted on different corpus. Moreover, in order to investigate sentence final tone estimation using semantic information, larger size of speech database will be necessary.

Bibliography

[1] Y. Fan, Y. Qian, F. K. Soong, and L. He, "Multi-speaker modeling and speaker adaptation for DNN-based TTS synthesis," in Proc. ICASSP 2015, 2015, pp. 4475–4479.

[2] T. Meguro, R. Higashinaka, Y. Minami, and K. Dohsaka, "Controlling listening-oriented dialogue using partially observable Markov decision processes," in Proc. the 23rd International Conference on Computational Linguistics. Association for Computational Linguistics, 2010, pp. 761–769.

[3] H. P. Grice, Studies in the Way of Words. Harvard University Press, 1991.

[4] D. Wilson and D. Sperber, Meaning and Relevance. Cambridge University Press, 2012.

[5] K. Maekawa, "Phonetic and phonological characteristics of paralinguistic information in spoken Japanese," in the 5th International Conference on Spoken Language Processing, 1998, pp. 635–638.

[6] N. Hellbernd and D. Sammler, "Prosody conveys speaker's intentions: Acoustic cues for speech act perception," Journal of Memory and Language, vol. 88, pp. 70–86, 2016.

[7] ITU-T, A Method for Subjective Performance Assessment of the Quality of Speech Voice Output Devices, International Telecommunication Union Std., 1994.

[8] R. van Bezooijen and V. J. van Heuven, "Assessment of speech synthesis," Handbook of Standards and Resources for Spoken Language Systems, pp. 481–653, 1997.

[9] J. L. Austin, How to Do Things with Words. Oxford University Press, 1975.

[10] J. R. Searle, F. Kiefer, and M. Bierwisch, Speech Act Theory and Pragmatics. Springer, 1980, vol. 10.

[11] J. Yamagishi, T. Kobayashi, M. Tachibana, K. Ogata, and Y. Nakano, "Model adaptation approach to speech synthesis with diverse voices and styles," in Proc. ICASSP 2007, vol. 4, 2007, pp. 1233–1236.

[12] 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/ACM Trans. Audio, Speech, Lang. Process., vol. 17, no. 1, pp. 66–83, 2009.

[13] V. Wan, J. Latorre, K. Chin, L. Chen, M. Gales, H. Zen, K. Knill, and M. Akamine, "Combining multiple high quality corpora for improving HMM-TTS," in Proc. INTERSPEECH 2012, 2012, pp. 1134–1137.

[14] H. Zen, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," in Proc. ICASSP 2013, 2013, pp. 7962–7966.

[15] H. Zen and A. Senior, "Deep mixture density networks for acoustic modeling in statistical parametric speech synthesis," in Proc. ICASSP 2014, 2014, pp. 3844–3848.

[16] Y. Fan, Y. Qian, F. Xie, and F. K. Soong, "TTS synthesis with bidirectional LSTM-based recurrent neural networks," in Proc. INTERSPEECH 2014, 2014, pp. 1964–1968.

[17] G. Saon, H. Soltau, D. Nahamoo, and M. Picheny, "Speaker adaptation of neural network acoustic models using i-vectors," in Proc. ASRU 2013, 2013, pp. 55–59.

[18] O. Abdel-Hamid and H. Jiang, "Fast speaker adaptation of hybrid NN/HMM model for speech recognition based on discriminative learning of speaker code," in Proc. ICASSP 2013, 2013, pp. 7942–7946.

[19] S. Xue, O. Abdel-Hamid, H. Jiang, L. Dai, and Q. Liu, "Fast adaptation of deep neural network based on discriminant codes for speech recognition," IEEE/ACM Trans. Audio, Speech, Lang. Process., vol. 22, no. 12, pp. 1713–1725, 2014.

[20] Z. Wu, P. Swietojanski, C. Veaux, S. Renals, and S. King, "A study of speaker adaptation for DNN-based speech synthesis," in Proc. INTERSPEECH 2015, 2015, pp. 879–883.

[21] N. Hojo, Y. Ijima, and H. Mizuno, "An investigation of DNN-based speech synthesis using speaker codes," in Proc. INTERSEPECH 2016, 2016, pp. 2278–2282.

[22] H.-T. Luong, S. Takaki, G. E. Henter, and J. Yamagishi, "Adapting and controlling DNN-based speech synthesis using input codes," in Proc. ICASSP 2017, 2017, pp. 4905–4909.

[23] B. Potard, P. Motlicek, and D. Imseng, "Preliminary work on speaker adaptation for DNN-based speech synthesis," Idiap, Tech. Rep., 2015.

[24] Y. Fan, Y. Qian, F. K. Soong, and L. He, "Unsupervised speaker adaptation for DNN-based TTS synthesis," in Proc. ICASSP 2016, 2016, pp. 5135–5139.

[25] Y. Fan, Y. Qian, F. K. Soong, and L. He, "Speaker and language factorization in DNN-based TTS synthesis," in Proc. ICASSP 2016, 2016, pp. 5540–5544.

[26] H. Kawahara, I. Masuda-Katsuse, and A. De Cheveigné, "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, no. 3, pp. 187–207, 1999.

[27] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv preprint arXiv:1412.6980, 2014.

[28] K. Shinoda and T. Watanabe, "Acoustic modeling based on the MDL criterion for speech recognition," in Proc. EUROSPEECH 1997, 1997, pp. 99–102.

[29] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in Proc. ICASSP 2000, vol. 3, 2000, pp. 1315–1318.

[30] T. Toda and K. Tokuda, "A speech parameter generation algorithm considering global variance for HMM-based speech synthesis," IEICE TRANSACTIONS on Information and Systems, vol. 90, no. 5, pp. 816–824, 2007.

[31] J. Yamagishi, K. Onishi, T. Masuko, and T. Kobayashi, "Acoustic modeling of speaking styles and emotional expressions in HMM-based speech synthesis," IEICE TRANSACTIONS on Information and Systems, vol. 88, no. 3, pp. 502–509, 2005.

[32] T. Nose, J. Yamagishi, T. Masuko, and T. Kobayashi, "A style control technique for HMM-based expressive speech synthesis," IEICE TRANSACTIONS on Information and Systems, vol. 90, no. 9, pp. 1406–1413, 2007.

[33] R. Barra-Chicote, J. Yamagishi, S. King, J. M. Montero, and J. Macias-Guarasa, "Analysis of statistical parametric and unit selection speech synthesis systems applied to emotional speech," Speech Communication, vol. 52, no. 5, pp. 394–404, 2010.

[34] S. An, Z. Ling, and L. Dai, "Emotional statistical parametric speech synthesis using LSTM-RNNs," in Proc. APSIPA ASC 2017, 2017, pp. 1613–1616.

[35] J. Lorenzo-Trueba, G. E. Henter, S. Takaki, J. Yamagishi, Y. Morino, and Y. Ochiai, "Investigating different representations for modeling and controlling multiple emotions in DNN-based speech synthesis," Speech Communication, vol. 99, pp. 135–143, 2018.

[36] L. Xue, X. Zhu, X. An, and L. Xie, "A comparison of expressive speech synthesis approaches based on neural network," in Proc. the Joint Workshop of the 4th Workshop on Affective Social Multimedia Computing and first

Multi-Modal Affective Computing of Large-Scale Multimedia Data, 2018, pp. 15–20.

[37] H. Fujisaki, "Prosody, models, and spontaneous speech," in Computing Prosody. Springer, 1997, pp. 27–42.

[38] K. R. Scherer and J. S. Oshinsky, "Cue utilization in emotion attribution from auditory stimuli," Motivation and Emotion, vol. 1, no. 4, pp. 331–346, 1977.

[39] A. K. Syrdal and Y.-J. Kim, "Dialog speech acts and prosody: Considerations for TTS," in Proc. Speech Prosody 2008, 2008, pp. 661–665.

[40] E. Ofuka, J. D. McKeown, M. G. Waterman, and P. J. Roach, "Prosodic cues for rated politeness in Japanese speech," Speech Communication, vol. 32, no. 3, pp. 199–217, 2000.

[41] Y. Katagiri, "Dialogue functions of Japanese sentence-final particles 'Yo' and 'Ne'," Journal of Pragmatics, vol. 39, no. 7, pp. 1313–1323, 2007.

[42] K. Iwata and T. Kobayashi, "Expression of speaker's intentions through sentence-final particle/Intonation combinations in Japanese conversational speech synthesis," in Proc. the 8th ISCA Speech Synthesis Workshop, 2013, pp. 235–240.

[43] P. Tsiakoulis, C. Breslin, M. Gasic, M. Henderson, D. Kim, M. Szummer, B. Thomson, and S. Young, "Dialogue context sensitive HMM-based speech synthesis," in Proc. ICASSP 2014, 2014, pp. 2554–2558.

[44] M. Tachibana, J. Yamagishi, T. Masuko, and T. Kobayashi, "A style adaptation technique for speech synthesis using HSMM and suprasegmental features," IEICE TRANSACTIONS on Information and Systems, vol. 89, no. 3, pp. 1092–1099, 2006.

[45] R. Skerry-Ryan, E. Battenberg, Y. Xiao, Y. Wang, D. Stanton, J. Shor, R. J. Weiss, R. Clark, and R. A. Saurous, "Towards end-to-end prosody transfer for expressive speech synthesis with tacotron," arXiv preprint arXiv:1803.09047, 2018.

[46] A. Stolcke, K. Ries, N. Coccaro, E. Shriberg, R. Bates, D. Jurafsky, P. Taylor, R. Martin, C. V. Ess-Dykema, and M. Meteer, "Dialogue act modeling for automatic tagging and recognition of conversational speech," Computational Linguistics, vol. 26, no. 3, pp. 339–373, 2000.

[47] A. Kurematsu, K. Takeda, Y. Sagisaka, S. Katagiri, H. Kuwabara, and K. Shikano, "ATR Japanese speech database as a tool of speech recognition and synthesis," Speech Communication, vol. 9, no. 4, pp. 357–363, 1990.

[48] R. Higashinaka, K. Imamura, T. Meguro, C. Miyazaki, N. Kobayashi, H. Sugiyama, T. Hirano, T. Makino, and Y. Matsuo, "Towards an open-domain conversational system fully based on natural language processing." in Proc. COLING 2014, 2014, pp. 928–939.

[49] T. Nose, Y. Arao, T. Kobayashi, K. Sugiura, and Y. Shiga, "Sentence selection based on extended entropy using phonetic and prosodic contexts for statistical paramaetric speech synthesis," IEEE/ACM Trans. Audio, Speech, Lang. Process., vol. 25, no. 5, pp. 1107–1116, 2017.

[50] J. J. Venditti, "The J_ToBI model of Japanese intonation," Prosodic Typology: The Phonology of Intonation and Phrasing, pp. 172–200, 2005.

[51] Z. Wu, O. Watts, and S. King, "Merlin: An open source neural network speech synthesis system," in Proc. the 9th ISCA Speech Synthesis Workshop, 2016, pp. 202–207.

[52] N. Hojo, Y. Ijima, and H. Mizuno, "DNN-based speech synthesis using speaker codes," IEICE TRANSACTIONS on Information and Systems, vol. 101, no. 2, pp. 462–472, 2018.

[53] H. Silén, E. Helander, J. Nurminen, and M. Gabbouj, "Ways to implement global variance in statistical speech synthesis," in Proc. the 13th Annual Conference of the International Speech Communication Association, 2012, pp. 1436–1439.

[54] N. Hojo and N. Miyazaki, "Evaluating intention communication by TTS using explicit definitions of illocutionary act performance," in Proc. INTERSPEECH 2019, pp. 1536–1540, 2019.

[55] B. Li and H. Zen, "Multi-language multi-speaker acoustic modeling for LSTM-RNN based statistical parametric speech synthesis," in Proc. INTERSPEECH 2016, 2016, pp. 2468–2472.

[56] H.-T. Luong and J. Yamagishi, "Multimodal speech synthesis architecture for unsupervised speaker adaptation," arXiv preprint arXiv:1808.06288, 2018.

[57] C. Miyazaki, T. Hirano, R. Higashinaka, and Y. Matsuo, "Towards an entertaining natural language generation system: Linguistic peculiarities of Japanese fictional characters," in Proc. the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue, 2016, pp. 319–328.

List of Publications

Publications Related to This Thesis

Journal

1. Nobukatsu Hojo, Yusuke Ijima, Hideyuki Mizuno,
   "DNN-based speech synthesis using speaker codes,"
   IEICE TRANSACTIONS on Information and Systems, Vol. 101, No. 2, pp. 462–472, 2018.

2. Nobukatsu Hojo, Yusuke Ijima, Hiroaki Sugiyama, Noboru Miyazaki, Takahito Kawanishi, Kunio Kashino,
   "DNN-based speech synthesis using dialogue act information and its evaluation with respect to illocutionary act naturalness,"
   Transactions of the Japanese Society for Artificial Intelligence (in press) (in Japanese).

3. Nobukatsu Hojo, Yusuke Ijima, Hiroaki Sugiyama,
   "Sentence final tone estimation using dialogue act information for speech synthesis within a spoken dialogue system,"
   (under review) (in Japanese).

International Conference

1. Nobukatsu Hojo, Yusuke Ijima, Hideyuki Mizuno,
   "An investigation of DNN-based speech synthesis using speaker codes," in Proc. INTERSEPECH 2016, pp.2278–2282, 2016.

2. Nobukatsu Hojo, and Noboru Miyazaki,

"Evaluating intention communication by TTS using explicit definitions of illocutionary act performance," in Proc. INTERSPEECH 2019, pp. 1536–1540, 2019.