PAPER Special Section on Human Communication V

Simultaneous Adaptation of Acoustic and Language Models for Emotional Speech Recognition Using Tweet Data***

Tetsuo KOSAKA[†a)], ***Member***, **Kazuya SAEKI**[†*], **Yoshitaka AIZAWA**[†**], **Masaharu KATO**[†], ***Nonmembers***, ***and*** **Takashi NOSE**[††], ***Member***

SUMMARY Emotional speech recognition is generally considered more difficult than non-emotional speech recognition. The acoustic characteristics of emotional speech differ from those of non-emotional speech. Additionally, acoustic characteristics vary significantly depending on the type and intensity of emotions. Regarding linguistic features, emotional and colloquial expressions are also observed in their utterances. To solve these problems, we aim to improve recognition performance by adapting acoustic and language models to emotional speech. We used Japanese Twitter-based Emotional Speech (JTES) as an emotional speech corpus. This corpus consisted of tweets and had an emotional label assigned to each utterance. Corpus adaptation is possible using the utterances contained in this corpus. However, regarding the language model, the amount of adaptation data is insufficient. To solve this problem, we propose an adaptation of the language model by using online tweet data downloaded from the internet. The sentences used for adaptation were extracted from the tweet data based on certain rules. We extracted the data of 25.86 M words and used them for adaptation. In the recognition experiments, the baseline word error rate was 36.11%, whereas that with the acoustic and language model adaptation was 17.77%. The results demonstrated the effectiveness of the proposed method.

key words: *speech recognition, emotional speech, acoustic model adaptation, language model adaptation, deep neural networks*

1. Introduction

Speech dialogue systems have been widely researched in recent years [1]–[5]. Some of these systems have been used for practical purposes such as information retrieval. Research is also being conducted to realize human-machine chat conversations [6]–[8]. Building a system that considers emotions is important for applications that want to provide a more enjoyable dialogue.

Manuscript received May 9, 2023.
Manuscript revised August 28, 2023.
Manuscript publicized December 5, 2023.

[†]The authors are with Yamagata University, Yonezawa-shi, 992–8510 Japan.
[††]The author is with Tohoku University, Sendai-shi, 980–8579 Japan.
[*]Presently, the author is with Tohoku Electric Power Co.
[**]Presently, the author is with Daiwabo Information system co. Ltd.

***The original version of this paper was published at conferences [23], [24], and [25]. In addition, new analysis results are added and summarized. Reference [23] proposed an acoustic model adaptation, Ref. [24] proposed simultaneous adaptation, and Ref. [25] proposed language model adaptation using a large amount of tweed data. © 2018 APSIPA. Reprinted, with permission, from [23]. © 2019 IEEE. Reprinted, with permission, from [24]. © 2020 APSIPA. Reprinted, with permission, from [25].

a) E-mail: tkosaka@yz.yamagata-u.ac.jp
DOI: 10.1587/transinf.2023HCP0010

The system must accurately recognize utterances and emotions to build a speech dialogue system that considers emotions [9]. This study focused on emotional speech recognition. Several speech corpora have been used to study emotional speech recognition [10], [11]. A few Japanese emotional speech corpora are available [12], [13]. In recent years, the Japanese Twitter-based emotional speech (JTES) has been proposed to be used for the study of emotional speech [14]. Here, emotion and emotional speech recognition were conducted. Emotional speech recognition was performed using standard Gaussian mixture model-based hidden Markov model (GMM-HMM) as the acoustic model. Sufficient recognition performance was not obtained.

Here, we built a speech recognition system using a deep neural network-based hidden Markov model (DNN-HMM). DNN-based speech recognition performed better than GMM-based recognition in large-vocabulary continuous speech recognition tasks. Similarly, DNN-based models show higher performance in emotional speech recognition [34]. However, even with DNN-based models, the recognition performance for emotional speech is still lower than that for non-emotional speech. For example, the word error rate (WER) was 15.12% for *testset1* in the corpus of spontaneous Japanese (CSJ) [15] using a well-trained acoustic model (AM) and language model (LM) (see Sect. 3.2 for training conditions). However, when the recognition for JTES was performed using the same models described above, the performance dropped significantly to 38.1%.

In this study, we used a DNN-HMM as an AM and considered further model adaptation to improve performance. Methods for improving emotional speech-recognition performance can be broadly classified into two categories. One is the improvement by acoustic features, and the other is model adaptation.

First, here are some examples of research on performance improvement using acoustic features. In [19] and [20], an improvement of acoustic features for emotional speech recognition was proposed. Robust emotional speech recognition in noisy environments was proposed by [21]. They used two types of masks, a binary mask and an emotion-specific mask, were used to reduce the effects of noise and emotions. In [22], frequency warping was applied to the feature extraction process of an automatic speech-recognition system.

Next, we introduce examples of model adaptation studies. AM and LM adaptation techniques have been widely

used to improve speech-recognition performance. In particular, many methods have been proposed for speaker adaptation (e.g. [16]). Adaptation to emotional speech is believed to be possible using the same methods; however, research examples are limited. This is thought to be because of the small amount of data contained in the emotional corpus. The text corpus is insufficient and LM adaptation is rarely performed. In [17], AM adaptation using MLLR and bottleneck features was performed and its effectiveness was demonstrated. However, because the LM was trained with the closed data, the recognition performance for open data was unknown. In addition, the AM was based on the GMM-HMM. AM adaptation with multiple-regression HMM was proposed in [18]. This method could be applied with a small amount of adaptation data, but the AM was based on the GMM-HMM. In [33], a DNN-based model adaptation for emotional speech recognition was proposed, in which a new model adaptation technique based on knowledge distillation was described[†].

As mentioned above, there are few studies on DNN-based AM and LM adaptations for emotional speech recognition. In this study, we aimed to improve the performance of emotional speech recognition using the simultaneous adaptation of the AM and LM.

Training texts are required for LM adaptation in emotional speech recognition. However, the emotional speech corpus is small, and LM adaptation cannot be performed sufficiently. In this study, we used tweet data instead of an emotional text corpus. Tweet data are written language; however, because there are many sentences with emotions, they are considered useful for LM adaptation for emotional speech recognition.

Another problem is that only a few types of emotional corpora exist. In many studies, the same corpus was used for training and evaluation, and experiments were conducted under task-closed conditions. However, when considering practical use, it is necessary to recognize speech outside of the task. In this study, we conducted evaluation experiments using an online gaming voice chat corpus with emotional labels (OGVC) [12], which was not used for training or adaptation (Sect. 5). For comparison, we conducted experiments with various adaptations of the OGVC (Sect. 6), and clarified the difference in performance between closed and open tasks.

The contributions of this study are as follows:

(i) Simultaneous adaptation Currently, there are not enough emotional speech corpora; therefore, there are very few studies on adaptation techniques for emotional speech recognition. There are almost no examples of LM adaptations in this field. Here, we clarified the effects of AM and LM adaptations on emotional speech recognition. Furthermore, the effects of simultaneous adaptation differed depending on the type of emotion.

(ii) LM adaptation using tweet data One of the main linguistic features of utterances in emotional speech is colloquial expression at the end of utterances. Because online tweet data contain linguistic features similar to the expressions, they are expected to be useful as training data for LMs. We performed LM adaptation using tweet data and demonstrated the effectiveness of using tweet data for emotional speech recognition.

(iii) System generalization Owing to the small number of emotion corpora, there are few recognition experiments using different corpora for training and adaptation. Therefore, the generalizability of emotion recognition methods is unclear. Here, we performed recognition experiments targeting the game chat emotion corpus, which differs from the style of tweets in JTES. Through experiments, we clarified that the proposed method has the possibility of general use, rather than simply adapting to the writing style of tweets.

The remainder of this paper is organized as follows. Section 2 describes adaptation and recognition methods. Section 3 describes the experimental conditions. Section 4 discusses the results of the speech-recognition experiments. Section 5 describes the results of recognition experiments conducted in OGVC. Section 6 describes the results of adaptation experiments in the OGVC. Finally, conclusions are presented in Sect. 7.

2. Adaptation and Recognition Methods

This section describes acoustic model adaptation [23], language model adaptation [24], [25], and speech recognition methods.

2.1 Acoustic Model Adaptation

In the past, the mainstream acoustic model in speech recognition was the GMM-HMM, which uses a Gaussian mixture distribution for the output probability of the HMM. However, since the advent of deep learning, deep neural networks (DNNs) have been used to calculate output probabilities. The associated model is called the DNN-HMM method because it replaces a part of the HMM framework with DNNs. The DNN-HMM was used as the acoustic model in this study.

Fine-tuning or transfer learning is often used as an adaptation method for deep-learning models. In either case, the model was trained using a large amount of general-purpose data, and this was used as the initial model. Subsequently, retraining was performed using a small amount of domain data to be adapted to improve the recognition accuracy for that domain. In fine-tuning, all parameters are retrained. Moreover, training only the parameters of the output layer is generally referred to as transfer learning. Here, we used the former as the adaptation method. Lecture speech was used for training the initial model. The lecture speech is unsuitable for emotional speech recognition because few speeches contain emotions. However, because the amount of lecture speech is large, it is considered suitable for the initial model training.

[†]In [33], our APSIPA2018 paper [23] is listed as a reference, and it is stated that other than this, there is almost no research on model adaptation in DNN-based emotional speech recognition.

Emotional speech was used as the adaptation data, and the initial model was adapted for emotional speech. The emotional speech data used in this study were small compared to the lecture speech data. A back-propagation algorithm was used for fine-tuning, where early stopping was introduced to automatically determine the number of epochs.

Early stopping is a technique for automatically determining the number of epochs for training parameters. In this method, the number is determined using part of the training data as the evaluation data and by performing cross-validation. In the iteration step of training, the iteration was stopped when the improvement rate of frame recognition was lower than the threshold value. In this study, the division ratio between the training and evaluation data was set at 9:1.

Conventionally, the adaptation of DNN-HMMs to emotional speech has not been sufficiently studied. To clarify the type and quantity of effective adaptation data, we conducted several types of adaptation experiments.

Corpus adaptation We used lecture speeches to train our initial model. The acoustic environment differs greatly between lecture speech and emotional speech used for evaluation. To eliminate this difference, the model is adapted for emotional speech. The adapted model is independent of speaker and emotion type. This method is generally effective for emotional speech, and a relatively large amount of adaptive data can be used. However, the characteristics of each speaker cannot be considered.

Emotion adaptation Adaptation was conducted using specific emotion data. The adapted model depends on a specific emotion and is independent of speakers. Acoustic characteristics are thought to differ depending on the emotions. This method considered these emotional differences. However, the characteristics of each speaker cannot be considered. A matched model is used in the recognition experiments.

Speaker adaptation Adaptation was conducted using data from the same speaker as the evaluated speaker. The adapted model is independent of emotion. This method reflects the emotions of a specific speaker. However, preparing utterances with transcription labels is necessary for a specific speaker in advance, which poses a practical problem. A matched model is used in the recognition experiments.

2.2 Language Model Adaptation

Linguistic features are also important for emotional-speech recognition. The paper [32] presents the results of an analysis of the linguistic features of angry speeches. In the case of anger, colloquial expressions that appear at the end of the utterance are said to be characteristic. From the above, it is possible that LM adaptation can model not only the appearance of words that express emotions directly but also colloquial expressions peculiar to emotions.

The speech recognition system used in this study employs an n-gram as the language model. We use a mixed n-gram as the LM adaptation method [26]. In this method, adaptation is performed by calculating the linear sum of the n-gram counts of the data for the initial model training and adaptation data. The probability of occurrence of word w_i in the adapted LM is calculated as follows:

$$p(w_i) = \frac{\alpha \cdot n_i^{adapt} + n_i^{base}}{\alpha \cdot N^{adapt} + N^{base}}, \tag{1}$$

where n_i^{adapt} and n_i^{base} are n-gram counts of the adaptation and training data for the initial model, respectively. N^{adapt} and N^{base} are the total number of $n$gram counts. α is the weight for adjusting the imbalance between the amount of adaptation and initial data. This value was experimentally set.

JTES was used as the adaptation data for LM adaptation; however, the amount of text data contained in JTES is limited. Unfortunately, no Japanese text corpus contains large-scale emotional expressions that are useful for LM adaptations. To solve this problem, we propose an LM adaptation method that uses online tweets. As expected, the tweet data contained several emotional and colloquial expressions. In the adaptation step, the sentences used for adaptation are extracted from the collected tweet data based on certain rules. After filtering based on these specified rules, a large amount of tweet data can be obtained.

We used the *Twitter* API to collect the online tweet data. The data used in this experiment were tweets posted on *Twitter* over 51 days in May, June, and October, 2019. Japanese tweet data, excluding retweets and tweets from bots, were collected randomly. Additionally, the collected tweet data included symbols such as pictograms, emoticons, and typographical errors; therefore, using these in the state in which they were collected was difficult. Accordingly, they were converted into an appropriate data format using the following process.

- URLs, hash tags, line feeds, and reply destination's "@account name" were replaced with blanks.
- Assuming that punctuation marks were sentence breaks, a text split was performed at these points.
- Texts with more than three words and less than 20 words were extracted from the results of the *MeCab* segmentation. *MeCab* is a morphological analysis tool for Japanese [27].

The reason for limiting the number of words in each sentence was to match the characteristics of JTES, which consisted of independent short utterances with an average of 17 words. Furthermore, during the extraction of each text, the value of bigram perplexity was calculated using the initial LM to select natural sentences as Japanese. Through this process, a large amount of tweet data consisting of 25.86 million words, was obtained.

2.3 Speech Recognition System

A two-pass decoder was used in this study as the speech

recognition system, where a bigram and a trigram were used for the first and second passes, respectively. These are types of n-gram LM. In the first pass, a word graph was generated using the DNN-HMM and bigram LM. Decoding was performed using a one-pass algorithm that involves a frame-synchronous beam search and tree-structured lexicon. In the second pass, the trigram LM was applied to rescore the word graph and the recognition result was obtained.

In the output probability calculation step, we use the probability compensation method [28]. In this step, the occurrence probability of the state becomes extremely high with some phonemes such as silence. To solve this problem, the output probability is compensated. The output probability of the DNN-HMM was calculated as

$$p(x|s_i) = \frac{p(s_i|x)p(x)}{p(s_i)}, \tag{2}$$

where $p(x)$ is the occurrence probability of input feature x, is omitted because it does not affect the recognition results. $p(s_i)$ is the probability of occurrence of state s_i. This value depends on the frequency of the appearance of the phoneme in the training data. Because phonemes, such as silence, frequently appear in the training data, $p(s_i)$ increases. By limiting this value, an extreme decrease in output probability can be prevented. The specific methods and effects are presented in [23]. This method is particularly effective when the amount of adaptation data is small.

3. Experimental Conditions

3.1 Emotional Speech Corpora

Here, we used two emotional speech corpora, JTES and an online gaming voice chat corpus with emotional labels (OGVC) [12].

JTES is based on tweets on *Twitter*. As tweets contain many emotional expressions, collecting speech utterances with various emotions is possible by emotionally reading the content. In the emotional speech corpus, it is important to consider the balance of the type of phonemes and prosody. JTES considers this point. Tweets containing emotions were extracted by referring to emotional keywords. Next, the extracted tweets were categorized by matching them with emotional expression words and classified into four types: joy, anger, sadness, and neutral. In addition, the number of characters per sentence was limited, sentences that ended with nouns were removed, and other special phrases or proper nouns were removed or edited manually.

Based on the above, a list of 2000 sentences (500 sentences × 4 emotions) was created and used as text for LM adaptation. When recording utterance data, it is necessary to consider phoneme and prosody balance. Therefore, balanced sentences were selected from the above 2000 sentences using the entropy-based method described in [14]. Finally, 200 sentences (50 sentences × 4 emotions) were extracted and used as the text for AM adaptation (refer Table 1).

The OGVC consists of the voices of game players, who play a massive multiplayer online role-playing game (MMOPRG). In an MMOPRG, players play online games and talk to one another. While concentrating on the game, they uttered speech containing various emotions. The OGVC comprises two types of speech: spontaneous and acted. The former involves recording conversations during the games. In the latter, professional actors read the transcripts of 17 dialogues extracted from gameplay conversations. We used the acted speech set in the experiments. When they read the transcripts, emotion type and intensity were specified in advance. There are eight types of emotions in the OGVC, and we used them in the recognition experiments.

Here, the experiments were performed under the assumption that the emotion labels and intensity labels contained in the corpora are correct.

Table 1 Number of sentences, speakers, and utterances used for adaptation and evaluation on JTES.

Title	total	adaptation	evaluation
# Sents for LM	2000	1960	40
# Sents for AM	200*	160	40
# Speakers	100	90	10
# Utterances**	20000	14400***	400

* Part of 2000 sentences for LM.
** # Sents for AM × # Speakers.
*** Same as *corpus adaptation* in Table 2.

3.2 Recognition Experiments

This section describes the conditions of the recognition experiment. First, we describe the proposed recognition system. In the speech analysis module, the speech signal was digitized at a sampling frequency of 16 kHz. The length of the analysis frame was 25 ms, and the frame period was set to 8 ms. A 25-dimensional feature, which comprised the log mel-filter bank features and log power, was derived from the digitized samples for each frame. Moreover, the delta and delta-delta features were calculated from the 25-dimensional features; hence, the total number of dimensions was 75 per frame. The input layer of the DNN used 75 coefficients with a temporal context of 11 frames, totaling 825 input features. The DNN had seven hidden layers with 2048 hidden units in each layer. The final output layer had 3003 units, corresponding to the total number of states for the shared-state triphone.

The AM and LM used in the baseline speech recognition system were trained using the CSJ by considering the number of data [15]. CSJ is Japan's largest spontaneous speech corpus. However, because the CSJ consists of lecture speech, it is unsuitable for recognizing emotional utterances. Speech data from 963 lectures in the CSJ were used for DNN-HMM training. The total speech length was approximately 203 hrs. The training method of the DNN is as follows: In the pre-training step, the restricted Boltzmann machine was used as the training method in the unsupervised mode. In the fine-tuning step, a class label was assigned to each frame, and a back-propagation algorithm with stochastic gradient descent was used. Cross-entropy was used as the loss function. The

Table 2 Number of adaptation samples and adapted models for each experiment.

Title	#adaptation samples	#adapted models
Corpus adaptation	14,400 (40 sentences × 4 emotions × 90 speakers)	1
Emotion adaptation	3,600 (40 sentences × 1 emotion × 90 speakers)	4 (= #emotions)
Speaker adaptation	160 (40 sentences × 4 emotions × 1 speaker)	10 (= #evaluation speakers)

Kaldi speech recognition toolkit was used to train DNN parameters [29]. Bigram and trigram models were used as the LMs. They were trained on textual data containing 2,668 lectures from CSJ, and the total number of words was 6.68 million. CSJ consists of lecture speech, whereas JTES consists of tweets. This causes problems with the unknown words. The proportion of unknown words in the evaluation data is 3.15%. To avoid this problem, unknown words were added to the word lexicon.

The DNN was adapted using a back-propagation algorithm. The mini-batch size was set to 2,048, and the L2 normalization factor was 2.0×10^{-4}. No momentum was used. The early stopping method was used to determine the number of epochs. The process terminates when the frame accuracy increases by less than 0.5%. Table 2 lists the number of adaptation samples for each adaptation. 10% of the adaptation samples was used for cross-validation. Bigram and trigram models were used as the LMs.

The LM adaptation using JTES was performed as follows. The 2,000 sentences for the LM listed in Table 1 were divided into 1,960 sentences for adaptation and 40 sentences for evaluation. This type of LM is refferred to as a *small-scale LM*. The adaptation data consisted of 17,711 words. To compare the effectiveness of the amount of training data, we also prepared an LM adapted with tweet data containing 25.86 million words. This LM is referred to as *large-scale LM*.

Speech samples for evaluation were designed such that the utterance content and speaker did not overlap with the adaptation data. The data of ten speakers uttering the above 40 evaluation sentences (400 utterances in total) were used as evaluation data (refer Table 1).

4. Results of Recognition Experiments

This section describes the effectiveness of simultaneous adaptation using the JTES. Section 4.1 describes the results of the AM adaptation alone, and Sect. 4.2 discusses the results of the LM adaptation alone and simultaneous adaptation.

4.1 Acoustic Model Adaptation

Table 3 lists the WER results of the AM adaptation experiments. The notations used in this table are as follows.

EPOCH5 The number of epochs was fixed to five.
ESTOP Early stopping was used in the experiments.
ESTOP+UNK Combination method of early stopping and unknown word addition.

Column *ESTOP+UNK* for the baseline is unrelated to *ESTOP*, because it has not been adapted. Only unknown-word additions were performed. The results demonstrate that the recognition performance without adaptation is low, whereas every adaptation method is effective. *ESTOP+UNK* yields the best results.

Table 3 Results of AM adaptation experiments (WER(%)). †Only unknown word addition was performed.

Type	EPOCK5	ESTOP	ESTOP+UNK
Baseline	38.10		36.11 †
Corpus	32.37	29.76	26.91
Emotion	29.50	29.42	26.98
Speaker	27.86	25.50	23.05

When comparing *EPOCH5* and *ESTOP*, the early stopping method was found to be effective. The number of epochs in the *ESTOP* experiments was large when speaker adaptation was conducted, and it depended on the speaker. The numbers ranged from 11 to 23. By contrast, the number is small when the corpus adaptation or emotion adaptation is conducted. As mentioned above, the number varies depending on the adaptation method, and fixing the number degrades recognition performance. Appendix A describes a further detailed study on early stopping.

In a comparison of the various adaptation methods in Table 3, speaker adaptation showed the best performance. This indicates that the influence of speaker characteristics is stronger than that of emotions. However, because speaker adaptation is performed in a supervised manner, labeled speech is required in advance for the speaker to be recognized. Therefore, speaker-adaptation is impractical. However, corpus adaptation and emotion adaptation do not require the speaker's speech; thus, they are highly practical. The corpus and emotion adaptation exhibited similar performance. This means that the emotion-dependent model was not as effective under these experimental conditions. In this experiment, four emotions, anger, joy, sadness, and neutral, were used. However, since various emotional strengths are possible even for the same emotion, it is considered inappropriate to simply classify emotion types. To create emotion-dependent models, further consideration of the classification type is necessary.

Figure 1 shows the WER for each emotion in the *ESTOP+UNK* condition. A performance improvement was observed in the results for all emotions. However, the performance for each emotion was lower than that for the neutral emotion. This agrees with the general finding that the recognition of emotional speech is difficult.

To analyze these results, we conducted two types of phoneme recognition experiments, as shown in Table 4. In this table, the phoneme recognition results obtained using the phoneme bigrams are shown in the upper row. The

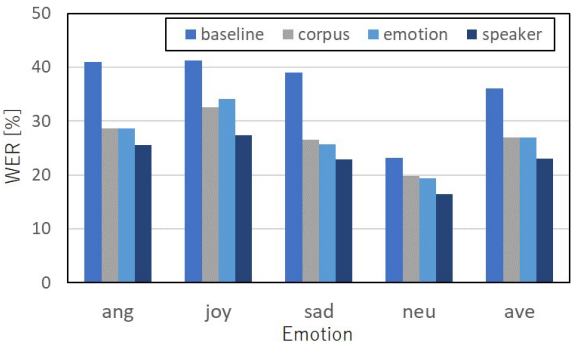

Fig. 1 Word error rate for each emotion using various AMs without LM adaptation [%].

Table 4 Phoneme error rate for each emotion using the corpus adaptation model with the phoneme bigram as the language model (%).

n-gram	ang	joy	sad	neu
Phoneme	18.38	20.27	12.94	14.99
Word	9.1	14.14	8.62	9.32

corpus adaptation model was used as the acoustic model, and the phoneme bigram was trained using the CSJ corpus. Using a different method, the word recognition results were converted into phoneme sequences to calculate the phoneme error rate (PER), and the results are shown in the lower row. These experiments also used the corpus adaptation model. Note that the results in the upper row were not affected by the word *n*-grams. The results in the upper row show that *ang* and *joy* are acoustically difficult to recognize. These emotions are characterized by large variations in their acoustic features. The standard deviation of the cepstrum from c_1 to c_{12} in the JTES evaluation set was calculated. The results were 7.17 for *ang*, 6.91 for *joy*, 6.19 for *sad*, and 6.14 for *neu*. As shown, *ang* and *joy* exhibit large variation in their cepstral features. By contrast, the lower results show different tendencies. The PER of *joy* is high as shown in the upper row, but that of *ang* is greatly reduced. This shows that the use of word LM is effective for the recognition of emotion *ang*. However, the WER of *ang* in Fig. 1 is not low compared to the PER shown at the bottom of this table. It is believed that there are some homonyms or morpheme segmentation errors.

As described above, acoustic variations and linguistic tendencies differ depending on emotion type. This suggests the importance of the simultaneous adaptation of the AM and LM.

4.2 Language Model Adaptation and Simultaneous Adaptation

We calculated the perplexity of the test set to examine the effects of LM adaptation. Table 5 shows the values of the test set perplexity in each LM when the weight α in Eq. (1) was set to the optimum value. The notations used in this table are as follows:

Baseline We used the LM trained using the CSJ. The CSJ consists of lecture speech.

Small-scale We used the LM adapted from the data in JTES. The JTES consists of tweet data; however, the amount of data is limited.

Large-scale We used the LM adapted by tweet data obtained online.

Table 5 Test set perplexity for JTES with various LMs.

Domain	CSJ	Twitter	
Model	baseline	small-scale	large-scale
Bigram	1053	849	272
Trigram	996	907	224

Table 6 Word error rate for JTES with various LMs (%).

LM adaptation			
Domain	CSJ	Twitter	
Model	baseline	small-scale	large-scale
Anger	40.97	36.26	28.61
Joy	41.23	31.19	30.31
Sadness	39.09	35.77	27.42
Neutral	23.13	20.60	16.39
Average	36.11	30.95	25.68
Simultaneous adaptation of AM and LM			
Model	baseline	small-scale	large-scale
Anger	28.61	28.29	20.70
Joy	32.62	24.24	21.85
Sadness	26.52	25.82	15.91
Neutral	19.88	18.26	12.65
Average	26.91	24.15	17.77

From the results, the *small-scale LM* has some effect, whereas the *large-scale* LM is the most effective. We experimentally determined the weight α that yields the lowest perplexity. Specifically, we set the weight to 100 for the bigram, 30 for the trigram for the *small-scale LM*, and two for the *large-scale LM* in each case. In the *small-scale LM*, the adaptation data are small compared to the data of the baseline model; thus, the optimum value of the weight is large.

The following describes LM adaptation and the simultaneous adaptation of LM and AM. AM adaptation in the latter was examined using corpus adaptation. As shown in Table 3, the performance of emotion adaptation is the same as that of corpus adaptation, although its calculation cost is high. Therefore, a corpus adaptation model was used.

Recognition experiments were conducted using only LM adaptation and simultaneous adaptation of LM and AM. The former performed only LM adaptation, and the AM used the baseline model, whereas the latter used the corpus adaptation model as the AM. Table 6 presents the results. *large-scale LM* yielded the best results. Based on the hypothesis that tweet data contains colloquial expressions that express emotions, we constructed *large-scale LM*. The results suggest the importance of using a large number of colloquial expressions.

The results demonstrate that the performance of the simultaneous adaptation of LM and AM was better than that of LM adaptation alone. Compared to the results of the AM adaptation alone, the performance of the large-scale LM

Table 7 Example of differences in recognition results by the *small-scale* and *large-scale LMs*. System used adapted AM. /Q/ is a geminate stop consonant.

Type	recognition results
Correct	Kore baQkari
Meaning	Only this
Small-scale LM	Kore bakari
Large-scale LM	Kore baQkari
Correct	Zikan aru toki
Meaning	When you have time
Small-scale LM	Zikan ga aru toki
Large-scale LM	Zikan aru toki
Correct	Heizitsu na no ni komi sugi
Meaning	It's too crowded on weekdays
Small-scale LM	Heizitsu no nikomi sugi
Large-scale LM	Heizitsu na no ni komi sugi

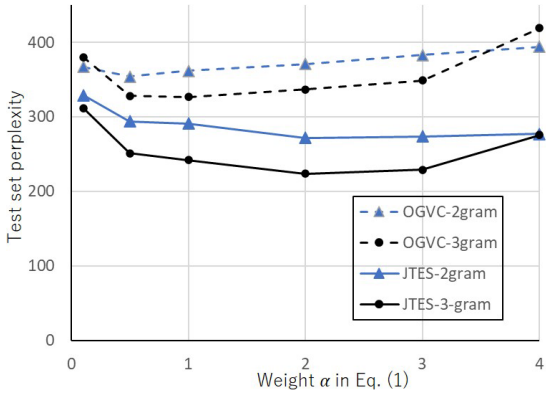

Fig. 2 Test set perplexity for JTES and OGVC with *large-scale LMs*.

adaptation alone was slightly higher. The WER of the former was **26.91%** and that of the latter was **25.68%**. Finally, the simultaneous adaptations yielded the best results of **17.77%**. Simultaneous adaptation of the AM and LM significantly improved performance, suggesting that the adaptations of the AM and LM had different characteristics.

Next, we analyzed the recognition results. Table 7 lists the differences between the recognition results of the *small-scale* and *large-scale LMs*. In the first example, the consonant /Q/ is missing in the *small-scale LM* condition. /Q/ is a geminate stop consonant expressed by a one-mora pause, which is a consonant unique to Japanese. In this case, although this consonant does not appear in typical vocalizations, it tends to appear when the emotions are emphasized. In the next example, the postpositional particle *ga* is inserted in the case of *small-scale LM*. Postpositional particles are parts of speech unique to the Japanese. Generally, *ga* occupies this position in Japanese grammar. However, postpositional particles often drop when colloquial expressions are used under the influence of emotions. Based on this, *large-scale LM* is believed to effectively express emotions and colloquial styles. In the last example, the auxiliary verb *na* is missing in the *small-scale LM* condition. Additionally, because of the elimination of *na*, a morphological analysis error occurred. The words *ni* and *komi* were merged into one word. We also investigated cases in which the recognition performance did not improve, even when *large-scale LM* was used. We found that recognizing fillers in colloquial style was difficult. In Japanese, fillers are often articulated as long vowels. These long vowels are often devocalized in emotional speech. The frequency of this error depended on the speaker. These errors are thought to be caused by the AM rather than the LM.

Next, we investigated why the *large-scale LM* performed better than the LM trained using only tweet data (*tweet LM*). In our experiments, tweet data were used for evaluation; thus, the *tweet LM* is expected to have higher performance. However, the results differ: when the *tweet LM* was used instead of the *large-scale LM* under the simultaneous adaptation conditions shown in Table 6, the WER was 25.97% for *ang*, 33.69% for *joy*, 34.55% for *sad*, and 37.23% for *neu*. The fact that the recognition performance using *tweet LM* degrades is consistent with the results of perplexity in Fig. 2, which will be described in the next section.

Table 8 Example of recognition results for each LM. Both results are the same as the phoneme sequence, but have different meanings or different morpheme segments.

Phoneme seq.	Large-scale LM	Tweet LM
saQki	さっき	殺気
desune	です/ね	で/拗ね
desu	です	で/ス
desunode	です/の/で	で/スノ/で

The rate of increase in WER was much higher for *sad* and *neu* than for *ang* and *joy*. Tweets contain many colloquial expressions; therefore, it is believed that whether the LM is suitable depends on the type of emotion. For *sad* and *neu*, which have many relatively formal utterances, the *tweet LM* can cause poor performance. By contrast, the performance on these emotions can be expected to be improved by using the CSJ and the tweet data together in training, because the CSJ contains many formal-style utterances. Looking at the specific errors, words and phrases unique to tweets tend to appear in the recognition results of the *tweet LM*. Examples are listed in Table 8. In the results of the *tweet LM*, many homonym and morphological segmentation errors are observed. Thus, by mixing tweet data and CSJ data, an effective LM can be created regardless of the degree of colloquialism.

5. Recognition Experiments in the OGVC

The experiments described in the previous section were task-closed experiments using the JTES for both adaptation and evaluation. To verify the usability of the adapted AM and LM for other tasks, we conducted recognition experiments using evaluation data from different tasks. In this section, we use the OGVC described in Sect. 3.1 as evaluation data. The OGVC consists of game players' online chats, and their acoustic and linguistic characteristics are significantly different from those of the JTES.

The corpus-adapted model for the JTES was used as the AM, and the *small-scale* and *large-scale LMs* were used as LMs. The evaluation data were the acted speech of the

OGVC, and four types of emotion-intensity data were used. A total of 448 sentences (112 sentences × 4 speakers) were used for each emotional intensity. The four levels of intensity range from 0 to 3, where 0 represents no emotion and 3 represents a strong emotion. Note that although the utterance with level 0 was uttered without emotion, the content of the utterance contained emotional expressions. For the LM, 64 unknown words appearing in the evaluation data were added to avoid the influence of unknown words, as in the experiments with the JTES.

We examined test set perplexity to confirm the validity of the adapted LMs. Figure 2 shows the results for *large-scale LM*. Although not shown here the test set perplexity of the baseline LM before adaptation (i.e., $\alpha = 0$) for the OGVC is 1279 for the bigram and 1245 for the trigram. On the other hand, the test set perplexity of the LM trained by tweet data only (i.e., $\alpha = \infty$) was 413 for bigram, and 378 for the trigram. From these results, it can be observed that the performance of *large-scale LM* generated by LM adaptation is higher than that of the baseline LM and the LM trained using tweet data only. The utterance contents were significantly different between the JTES and OGVC. However, because the test set perplexity dropped sufficiently for the OGVC, the *large-scale LM* adapted to tweet data was considered versatile. Compared to the results of JTES, the values were higher, but the tendencies were similar. The reason why the best weight for the JTES is shifted to a higher value than that for the OGVC is that the domain of the JTES matches adaptation data.

Table 9 presents the recognition results for the OGVC. From the comparison of LM adaptation and simultaneous adaptation, AM adaptation was effective, although it was adapted in another task. When comparing emotional intensity, recognition becomes difficult if emotional intensity is high. This suggests that emotional intensity has a significant influence on the acoustic characteristics. From the comparison between the LMs, the performance improvement in the *small-scale LM* was limited; however, a sufficient improvement was observed in the *large-scale LM*. Although the OGVC consists of in-game utterances and is dissimilar to the JTES, the results show significant improvements. This suggests that the proposed LM is versatile for emotion recognition applications. In Table 9, the *large-scale LM* used the weights determined using the JTES, whereas the optimal weights for the OGVC (0.5 for the bigram and 1.0 for the trigram) were used in *large-scale LM1*. The difference in performance between the two methods was small. This indicated that the weights determined by the other tasks were versatile. From this, the AM and LM adaptations were shown to be effective, even in open tasks.

6. Adaptation Experiments in the OGVC

In Sect. 5, we described the recognition results of the OGVC using AMs and LMs adapted to the JTES. In this section, we present the results of recognition experiments with models adapted to the OGVC. Because acted speech in the OGVC has a limited amount of data, some ingenuity was required. Only four speakers were included in the acted-speech dataset. Therefore, the utterances of three speakers were used for adaptation, while the remaining one was used for evaluation. The AM adaptation experiments were performed by repeating this process four times. By contrast, the utterance content of the acted-speech dataset is part of the spontaneous-speech dataset. Therefore, the utterance content of the spontaneous speech can be used as the adaptation texts for the LM adaptation.

Based on the above, 1,334 sentences (112 sentences × 4 emotion strengths × 3 speakers) from the acted speech were used for the AM adaptation, and 17,711 words from the spontaneous speech were used for the LM adaptation. We used the same evaluation data as described in Sect. 5. For the AM adaptation, the corpus adaptation model was used as the initial model. For the LM adaptation, we compared the results when the LM trained by the CSJ was used as the initial model (*OGVC LM*) and when the *large-scale LM* was used as the initial model (*OGVC LM base on large-scale*). As in the previous section, the weights α for adaptation of bigrams and trigrams were experimentally set so that perplexity was the lowest.

Table 10 lists the obtained experimental results. As for the AM adaptation, the effect is small compared with Table 9, but some effect is seen for evaluation data with high emotional intensity. As for the LM adaptation, *OGVC LM*

Table 9 Word Error rate for OGVC with various LMs (%). *Large-scale LM1* shows the results using the optimal weight for OGVC.

LM adaptation				
Domain	CSJ	Twitter		
Emotional intensity	baseline LM	small-scale LM	large-scale LM	large-scale LM1
0	32.61	31.52	24.48	24.30
1	41.59	39.79	33.03	33.37
2	47.30	45.89	40.16	39.81
3	54.29	52.94	47.33	47.14
Simultaneous adaptation of AM and LM				
0	28.02	26.75	21.01	20.85
1	33.37	31.64	25.99	25.65
2	37.94	36.46	31.17	30.77
3	43.12	42.66	37.73	37.42

Table 10 Word Error rate for OGVC with various LMs adapted to OGVC (%). The corpus adaptation model was used as the baseline AM.

LM adaptation by OGVC				
Domain	CSJ	Twitter	OGVC	
Emotional intensity	baseline LM	large-scale LM	OGVC LM	OGVC LM based on large-scale
0	28.02	21.01	24.36	20.63
1	33.37	25.99	28.77	25.16
2	37.94	31.17	34.22	29.72
3	43.12	37.73	39.14	35.94
Simultaneous adaptation of AM and LM by OGVC				
0	27.15	20.38	23.64	19.82
1	32.23	26.15	28.46	25.07
2	35.70	28.83	31.82	27.44
3	39.58	35.73	36.07	33.73

has lower performance than the *large-scale LM*, but when the *large-scale LM* is used as the initial model, some effect is seen for evaluation data with high emotional intensity. As mentioned in Sect. 5, it is clear that the corpus adaptation AM and the *large-scale LM* are effective for different tasks, such as OGVC. Moreover, when AM and LM models are adapted to the OGVC, some effect is seen for data with strong emotional intensity.

7. Conclusions

We investigated the improvement in emotional speech recognition performance through the simultaneous adaptation of AMs and LMs. Both LM and AM adaptations yielded good results in recognition experiments. The baseline WER was 36.11%, whereas that of simultaneous AM and LM adaptation it was 17.77%. This establishes the effectiveness of the proposed method. In addition, the proposed simultaneous adaptation demonstrated performance improvement even in dissimilar emotional environments such as the OGVC, thereby confirming its potential for general use.

In these experiments, the proposed adaptation method was found to be effective for OGVC; however, the recognition performance was still low. In the future, we plan to examine emotion adaptation by investigating AM adaptation in relation to emotion intensity rather than simply creating emotion-dependent models. Such class models can be used in ASR systems by automatically selecting them based on a likelihood criterion. Additionally, we plan to improve the emotion recognition system developed in our laboratory by using the speech recognition outputs examined in this study [30]. Furthermore, we plan to introduce the proposed recognition techniques to our multimodal dialogue system [31].

Acknowledgments

This study was supported in part by a grant-in-aid for scientific research (KAKENHI 19K12014 and 22K12087) from the Japan Society for Promotion of Science.

References

[1] L. Smidl, A. Chylek, and J. Svec, "A Multimodal dialogue system for air traffic control trainees based on discrete-event simulation," Proc. Interspeech2016, San Francisco, USA, pp.379–380, 2016.
[2] A. Maier, J. Hough, and D. Schlangen, "Towards deep end-of-turn prediction for situated spoken dialogue systems," Proc. Interspeech2017, Stockholm, Sweden, pp.1676–1680, 2017.
[3] M. Li, Z. He, and J. Wu, "Target-based state and tracking algorithm for spoken dialogue system," Proc. Interspeech2016, San Francisco, USA, pp.2711–2715, 2016.
[4] C. Liu, P. Xu, and R. Sarikaya, "Deep contextual language understanding in spoken dialogue systems," Proc. Interspeech2015, Dresden, Germany, pp.120–124, 2015.
[5] P.-H. Su, D. Vandyke, M. Gasic, D. Kim, N. Mrksic, T.-H. Wen, and S. Young, "Learning from real users: rating dialogue success with neural networks for reinforcement learning in spoken dialogue systems," Proc. Interspeech2015, Dresden, Germany, pp.2007–2011, 2015.
[6] A. Lee, K. Oura, and K. Tokuda, "MMDAgent – a fully open-source toolkit for voice interaction systems," Proc. ICASSP2013, Vancouver, Canada, pp.8382–8385, 2013.
[7] K. Ohta, R. Marumoto, R. Nishimura, and N. Kitaoka, "Selecting type of response for chat-like spoken dialogue systems based on acoustic features of user utterances," Proc. APSIPA-ASC2017, Kuala Lumpur, Malaysia, pp.1248–1252, 2017.
[8] T. Kawahara, "Spoken dialogue system for a human-like conversational robot ERICA," Proc. IWSDS2018, Singapore, pp.65–75, 2018.
[9] R. Zhang, A. Atsushi, S. Kobashikawa, and Y. Aono, "Interaction and transition model for speech emotion recognition in dialogue," Proc. Interspeech2017, Stockholm, Sweden, pp.1094–1097, 2017.
[10] F. Burkhardt, A. Paeschke, M. Rolfes, W.F. Sendlmeier, and B. Weiss, "A database of German emotional speech," Proc. Interspeech2005, Lisbon, Portugal, pp.3–6, 2005.
[11] A. Batliner, C. Hacker, S. Steidl, E. Noth, S. D'Arcy, M. Russell, and M. Wong, "You stupid tin box – children interacting with the AIBO robot: A cross-linguistic emotional speech corpus," Proc. of LREC2004, Lisbon, Portugal, pp.171–174, 2004.
[12] Y. Arimoto, H. Kawatsu, S. Ohno, and H. Iida, "Naturalistic emotional speech collection paradigm with online game and its psychological and acoustical assessment," Acoust. Sci. Technol., vol.33, no.6, pp.359–369, 2012.
[13] H. Mori, T. Satake, M. Nakamura, and H. Kasuya, "Constructing a spoken dialogue corpus for studying paralinguistic information in expressive conversation and analyzing its statistical / acoustic characteristics," Speech Communication, vol.53, no.1, pp.36–50, 2011.
[14] E. Takeishi, T. Nose, Y. Chiba, and A. Ito, "Construction and analysis of phonetically and prosodically balanced emotional speech database," Proc. O-COCOSDA2016, Bali, Indonesia, pp.16–21, 2016.
[15] K. Maekawa, "Corpus of spontaneous Japanese: Its design and evaluation," Proc. ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition, Tokyo, Japan, pp.1–6, 2003.
[16] K. Shinoda, "Transfer learning in speech recognition: Speaker adaptation," Journal of the Japanese Society for Artificial Intelligence, vol.27, no.4, pp.359–364, 2012 (in Japanese).
[17] K. Mukaihara, S. Sakti, G. Neubig, T. Toda, and S. Nakamura, "Bottleneck features for emotional speech recognition," IPSJ SIG Tech. Report, vol.2015-SLP-107, no.2, Nagano, Japan, pp.1–6, 2015 (in Japanese).
[18] Y. Ijima, T. Nose, M. Tachibana, and T. Kobayashi, "A Rapid Model Adaptation Technique for Emotional Speech Recognition with Style Estimation Based on Multiple-Regression HMM," IEICE Trans. Inf. & Syst., vol.E93-D, no.1, pp.107–115, 2010.
[19] M. Sheikhan, D. Gharavian, and F. Ashoftedel, "Using DTW neural-based MFCC warping to improve emotional speech recognition," Neural Computing and Applications, vol.21, no.7, pp.1–9, 2021.
[20] Y. Sun, Y. Zhou, Q. Zhao, and Y. Yan, "Acoustic feature optimization for emotion affected speech recognition," Proc. ICIECS2009, Wuhan, China, pp.1–4, 2009.
[21] M. Bashirpour and M. Geravanchizadeh, "Robust emotional speech recognition based on binaural model and emotional auditory mask in noisy environments," EURASIP Journal on Audio, Speech, and Music Processing, vol.2018, no.1, 2018.
[22] M. Geravanchizadeh, E. Forouhandeh, and M. Bashirpour, "Feature compensation based on the normalization of vocal tract length for the improvement of emotion-affected speech recognition," EURASIP Journal on Audio, Speech, and Music Processing, vol.2021, no.1, 2021.
[23] T. Kosaka, Y. Aizawa, M. Kato, and T. Nose, "Acoustic model adaptation for emotional speech recognition using Twitter-based emotional speech corpus," Proc. APSIPA-ASC2018, Honolulu, Hawaii, pp.1747–1751, 2018.
[24] K. Saeki, M. Kato, and T. Kosaka, "Performance Improvement of Prosody-Controlled Voice Conversion by Language Model Adapta-

tion," Proc. IEEE GCCE2019, Osaka, Japan, pp.854–856, 2019.
[25] K. Saeki, M. Kato, and T. Kosaka, "Language model adaptation for emotional speech recognition using tweet data," Proc. APSIPA-ASC2020, Auckland, New Zealand, pp.371–375, 2020.
[26] A. Ito and M. Kohda, "Evaluation of task adaptation using N-gram count mixture," IEICE Trans., vol.J83-D-II, no.11, pp.2418–2427, 2000 (in Japanese).
[27] T. Kudo, K. Yamamoto, and Y. Matsumoto, "Applying conditional random fields to Japanese morphological analysis," Proc. EMNLP2004, Barcelona, Spain, pp.230–237, 2004.
[28] K. Tomita, A. Takagi, M. Kato, and T. Kosaka, "Evaluation of unsupervised cross adaptation using highly accurate models," Proc. ASJ2016 Autumn Meeting, Toyama, Japan, pp.95–96, 2016 (in Japanese).
[29] P. Daniel et al., "The Kaldi speech recognition toolkit," Proc. 2011 IEEE Workshop on Automatic Speech Recognition and Understanding, Hawaii, USA, pp.1–4, 2011.
[30] M. Sakurai and T. Kosaka, "Emotion recognition combining acoustic and linguistic features based on speech recognition results," Proc. IEEE GCCE2021, Kyoto, Japan, pp.889–892, 2021.
[31] T. Koseki and T. Kosaka, "Multimodal spoken dialog system using state estimation by body motion," Proc. IEEE GCCE2017, Nagoya, Japan, pp.348–351, 2017.
[32] H. Iida and Y. Arimoto, "A method for estimating the degree of emotional expressions by parameterizing acoustic and linguistic features and treating them integratedly," Studies in Pragmatics, no.8, pp.33–46, 2006 (in Japanese).
[33] H.-I. Yun and J.-S. Park, "End-to-end emotional speech recognition using acoustic model adaptation based on knowledge distillation," Multimedia Tools and Applications, vol.82, no.15, pp.22759–22776, 2023.
[34] V. Raju, K. Gurugubelli, M.S. Ganesh, and A.K. Vuppala, "Towards feature-space emotional speech adaptation for TDNN based Telugu ASR systems," Proc. SMM19, Vienna, Austria, pp.16–20, 2019.

Appendix: Detailed Examination of Early Stopping

Table A· 1 shows the relationship between the number of epochs and recognition performance in the corpus and emotion adaptation experiments. In this table, ORACLE indicates that the number was adjusted to the optimum value. In corpus adaptation, the auto-determined number is equal to the optimum number. This indicates that the method was successful in this case. For emotion adaptation, the WERs in the ESTOP experiments were similar to those in the ORACLE except for the neutral emotion. Based on these results, this method is considered effective, particularly when recognition performance is low.

Table A· 1 Relationship between number of epochs and recognition results (WER[%]).

Type	class	EPOCH5	ESTOP (#epochs)	ORACLE (#epochs)
Corpus		32.37	29.76 (1)	29.76 (1)
Emotion	Average	29.50	29.42	28.84
	Anger	35.83	34.58 (2)	34.17 (3)
	Joy	37.23	37.23 (3)	37.08 (1)
	Sadness	25.30	25.76 (1)	25.30 (5)
	Neutral	19.64	20.12 (2)	18.80 (10)

Tetsuo Kosaka received the B.E., M.E. and Ph.D. degrees from Tohoku University in 1984, 1986 and 1997, respectively. In 1986, he joined CANON, Inc. From 1991 to 1995, he was a researcher at the ATR Interpreting Telephony Research Laboratories. He also joined the Computer Science Laboratory at MIT in 1994. Since 2002, he has worked at Yamagata University, where he is currently a professor. His research interests include speech processing and its applications. He received the Paper Award from IEICE, Japan, in 1996. He is a senior member of IEEE and a member of the ASJ and IPSJ.

Kazuya Saeki received the B.E. and M.E. degrees from Yamagata University, Yamagata, Japan, in 2019 and 2021, respectively. His research interests include speech recognition and voice conversions. He is currently working at Tohoku Electric Power Co.

Yoshitaka Aizawa received the B.E. and M.E. degrees from Yamagata University, Yamagata, Japan, in 2016 and 2018, respectively. His research interests include speech recognition and voice conversions. He is currently working at Daiwabo Information System Co. Ltd.

Masaharu Kato received the B.E., and M.E. degrees from Yamagata University, Yamagata, Japan, in 1991 and 1993, respectively. He also received the Ph.D. from Tohoku University in 2010. Since 1993, he had been working at Yamagata University, and was an assistant professor at the Graduate School of Science and Engineering, at the University. He passed away in 2022.

Takashi Nose received the B.E. degree in electronic information processing, from Kyoto Institute of Technology, Kyoto, Japan, in 2001. He received the Dr. Eng. degree in Information Processing from the Tokyo Institute of Technology, Tokyo, Japan, in 2009. He was a Ph.D. researcher of the 21st Century Center of Excellence (COE) program and Global COE program in 2006 and 2007, respectively. He was an Intern Researcher at ATR Spoken Language Communication Research Laboratories (ATR-SLC) from July 2008 to January 2009. He became an assistant professor of the Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology, in 2009. He became a lecturer of the Graduate School of Engineering, Tohoku University, Sendai, Japan in 2013. He is currently an associate professor at the Graduate School of Engineering, Tohoku University, Japan. He is a member of the IEEE, ISCA, IEICE, IPSJ, and ASJ. His research interests include speech synthesis, speech recognition, spoken dialogue systems, image generation, and music information processing.