Acoustic Model Adaptation for Emotional Speech Recognition Using Twitter-Based Emotional Speech Corpus

Tetsuo Kosaka*, Yoshitaka Aizawa*, Masaharu Kato* and Takashi Nose†
* Graduate School of Science and Engineering, Yamagata University, Yonezawa, Japan
E-mail: tkosaka@yz.yamagata-u.ac.jp Tel: +81-238-263369
† Graduate School of Engineering, Tohoku University, Sendai, Japan
E-mail: tnose@m.tohoku.ac.jp Tel: +81-227-957112

***Abstract*—In recent years, Japanese Twitter-based emotional speech (JTES) was constructed as an emotional speech corpus. This corpus is based on tweets, and has features wherein an emotional label is assigned to each sentence, and sentences are selected considering the balance of both phoneme and prosody. Compared to speech recognition without emotion, emotional speech recognition is a difficult task. In this study, we aim to improve the performance of emotional speech recognition on the JTES corpus using acoustic model adaptation. For recognition, a deep neural network-based hidden Markov model (DNN-HMM) is used as the acoustic model. As a baseline, a word error rate (WER) of 38.0% was obtained when the DNN-HMM was trained by the corpus of spontaneous Japanese. This model was used as an initial model for adaptation. In this study, various types of adaptation were examined, and substantial performance improvement was achieved. Finally, a WER of 23.05% was obtained using speaker adaptation.**

I. Introduction

In recent years, spoken dialogue systems have received attention [1]–[5]. In such systems, interaction in a rote routine is sufficiently practical in the case of a specific purpose such as information retrieval. However, the application is being employed to conduct dialogue not only for task achievement but also in a chat-like manner [6]. For applications that enjoy the dialogue itself, it is important to build a system considering emotion. To realize speech dialogue system considering emotion, it is necessary for the system to accurately recognize emotions and utterance contents [7]. In this study, we focus on the latter; emotional speech recognition.

Several emotional speech corpora that can be used for such researches have been constructed [8]–[11]. For Japanese Twitter-based emotional speech (JTES), use of an utterance set that is phonetically and prosodically balanced was proposed in [12]. In this study, emotion recognition and emotional speech recognition were also conducted. For emotion recognition, an accuracy of approximately 68% to 76% was achieved. On the contrary, although emotional speech recognition was conducted using a standard Gaussian mixture model-based hidden Markov model (GMM-HMM) as the acoustic model, sufficient recognition performance could not be obtained.

In our study, we develop a speech recognition system using the deep neural network-based hidden Markov model (DNN-HMM) for recognizing utterances in JTES. DNN-based speech recognition is well known and has received considerable attention for its performance in large-vocabulary continuous speech recognition. However, emotional speech cannot be sufficiently recognized even if the DNN-HMM is used as the acoustic model.

The prosody is different between emotional speech and neutral speech; it appears as a difference in duration, speech strength, and pitch. Prosody also affects spectral patterns. To solve the problem, we investigate the adaptation methods of the acoustic model. Study on the model adaptation of emotional speech recognition using the DNN-HMM is lacking. In this work, we investigate the following three points:

- Examination of effect by difference of adaptation data.
- Investigation of the number of epochs in the adaptation step.
- Output probability compensation in the recognition step.

To confirm the effectiveness of the above, we conducted various speech recognition experiments using 400 emotional utterances in JTES.

The remainder of this paper is organized as follows: Section II introduces the emotional speech corpus JTES. Section III describes the proposed adaptation and recognition methods. Adaptation types are described in Section IV. Section V describes the conditions of the speech recognition experiments. Section VI describes the results of the speech recognition experiments. Section VII provides our conclusions.

II. Emotional Speech Corpus: JTES

JTES is based on tweets on Twitter and comprises speech utterances by 50 males and 50 females [12]. As Twitter contains many colloquial expressions, it is possible to collect speech utterances with various emotions by emotionally reading out the contents. Tweets were classified into four emotion classes, namely joy, anger, sadness and neutral using emotional expression words. Phonetically and prosodically balanced sentences were selected using the sentence selection algorithm based on entropy. Finally, 50 sentences for each emotion were selected. Emotional utterances were recorded

using those sentences. The total number of utterances in JTES is 20,000.

III. Adaptation and Recognition Methods

In this section, we describe the adaptation and recognition methods used in our system. We use the DNN-HMM as an acoustic model in this work. As another method, long short term memory recurrent neural network can also be used [13]. We would like to compare these two methods in the future. In the experiments, we conduct supervised adaptation on the premise that a correct label is given to each utterance for adaptation. The back-propagation algorithm is used for adaptation where early stopping is introduced for automatically determining the number of epochs [14]. In the recognition step, we use the output probability compensation method [15]. In addition, we describe the correspondence of language models to unknown words.

A. Early Stopping

Early stopping is a technique for automatically determining the number of epochs during adaptation or training DNN parameters. In this method, the number is determined using a part of adaptation (or training) data as evaluation data and performing cross-validation. In the iteration step of adaptation or training, the iteration is stopped when the improvement rate of frame recognition becomes lower than the threshold value. This can be expected to avoid over-fitting of parameters. In this study, the division ratio between adaptation data and evaluation data is set to 9:1.

B. Compensation of Output Probability

In output probability calculation, there is a problem that the occurrence probability of state becomes extremely high with some phonemes such as silence. To solve this problem, output probability is compensated in the recognition step. The output probability of the DNN-HMM is calculated as

$$p(x|s_i) = \frac{p(s_i|x)p(x)}{p(s_i)}, \tag{1}$$

where $p(x)$, the occurrence probability of an input feature x, is omitted because it does not affect the recognition result. $p(s_i)$ is the occurrence probability of state s_i. This value depends on the appearance frequency of a phoneme in training data. Since phonemes such as silence frequently appear in training data, $p(s_i)$ becomes high. By limiting this value, an extreme decrease in the output probability can be prevented. The specific method is as follows. When $p(s_i)$ exceeds the upper limit θ, it is replaced with θ. The value θ is determined by setting the limiting rate α in (2).

$$\alpha = \frac{\sum_{i \in D}\{p(s_i) - \theta\}}{\sum_{i=1}^{I} p(s_i)}, \tag{2}$$

where I is the total number of states, and D is the set of i that satisfies $p(s_i) > \theta$. The explanatory diagram is shown in Fig. 1 where i is rearranged in descending order of $p(s_i)$. α represents the ratio of the hatched portion to the total area surrounded by the curve. This method is effective especially when the amount of adaptation data is small.

TABLE I
Training conditions for DNN

Pre-training	
#epochs	10 (20 only for the first layer)
Mimi-batch size	1024
Momentum	0.9
L2 regularization factor	0.0002
Fine-tuning	
#epochs	The process terminates when the frame accuracy increases by less than 0.1%.
Mini-batch size	512

C. Correspondence of Unknown Words

The language models used in speech recognition herein are trained using the corpus of spontaneous Japanese (CSJ) by considering the amount of data. The CSJ is the largest spontaneous speech corpus in Japanese [16]. However, as the CSJ consists of lecture speech, it is not suitable for recognizing utterances in JTES. Further, it is difficult to collect training data sufficiently to create a language model for tweets. We addressed this problem by adding unknown words to the word lexicon. The proportion of unknown words in evaluation data was 3.15%. We added these unknown words in the experiments and conducted them excluding the influence of unknown words.

IV. Adaptation Types

Conventionally, the adaptation of the DNN-HMM for emotional speech has not been sufficiently studied. Therefore, to clarify what type of adaptation data is effective, we compare various adaptation conditions in recognition experiments. The experimental conditions of each adaptation are shown below. Each adaptation is conducted in the supervised mode.

Speaker adaptation Adaptation is conducted using the data of the same speaker as the evaluated speaker. The adapted model is independent of emotion.

Corpus adaptation The acoustic environment differs greatly between the CSJ used for training of the pre-adaptation model and JTES used for evaluation. To handle this difference, the model is adapted to the environment of JTES. The adapted model is independent of the speaker and emotion.

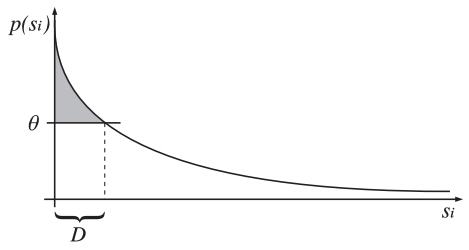

Fig. 1. Explanatory diagram of output probability compensation.

TABLE II
NUMBER OF ADAPTATION SAMPLES AND ADAPTED MODELS FOR EACH EXPERIMENT

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

Emotion adaptation Adaptation is conducted using specific emotion data. The adapted model depends on specific emotion and is independent of the speaker.

Speaker and emotion adaptation The model is adapted to both speaker and emotion. The best matching can be expected between the acoustic model and the acoustic environment. However the amount of adaptation data is small.

V. EXPERIMENTAL CONDITIONS

Experimental conditions are described in this section. First, we describe our recognition system. In the speech analysis module, a speech signal is digitized at a sampling frequency of 16 kHz with a quantization size of 16 bits. The length of the analysis frame is 25 ms, and the frame period is set to 8 ms. A 25-dimensional feature, which comprises the log mel-filter bank features and the log power, is derived from the digitized samples for each frame. Moreover, the delta and delta–delta features are calculated from the 25-dimensional feature, and hence the total number of dimensions is 75 per frame. The input layer of the DNN uses 75 coefficients with a temporal context of 11 frames, summing to a total of 825 input features. The DNN has seven hidden layers with 2048 hidden units in each layer. The total number of states for shared-state triphone is 3003. The final output layer has 3003 units, corresponding to the total number of states.

Speech data of 963 lectures in the CSJ are used for DNN-HMM training. The total speech length is approximately 203 h. The training method of the DNN is as follows. In the pre-training step, the restricted Boltzmann machine was used as the method of training in the unsupervised mode. In the fine-tuning step, a class label was given for each frame, and the back-propagation algorithm with stochastic gradient descent was used. Cross entropy was used as the loss function. Other conditions of DNN training are shown in Table I.

The bigram and trigram models were used as language models. They were trained on textual data containing 2668 lectures from the CSJ, and the total number of words was 6.68M. For experiments of adding unknown words, we added 44 words appearing only in the evaluation data as unknown words to the word lexicon.

The configuration of the recognition system is as follows: A two-pass search decoder with a bigram and trigram was used for recognition. In the first pass, a word graph was generated using the DNN-HMM and the bigram language model. Decoding was performed using a one-pass algorithm that involves a frame-synchronous beam search and a tree-structured lexicon. In the second pass, the trigram language model was applied to re-score the word graph, and the recognition result was obtained.

TABLE III
ADAPTATION CONDITIONS FOR DNN

Mini-batch size	2048
Momentum	0.0
L2 regularization factor	0.0002
#epochs	The process terminates when the frame accuracy increases by less than 0.005% for early stopping experiments.

TABLE IV
RECOGNITION RESULTS FOR LIMITING RATE α

α	0.00	0.05	0.10	0.15
WER(%)	39.33	37.79	36.12	36.75

Adaptation of the DNN was conducted using a back-propagation algorithm like fine-tuning. The number of adaptation samples for each adaptation is shown in Table II. Four hundred sentences (10 sentences × 4 emotions × 10 speakers) from JTES different from the adaptation data were used as evaluation data. The detailed conditions of the DNN adaptation are shown in Table III.

VI. RECOGNITION EXPERIMENTS

A. Effectiveness of Compensation of Output Probability

To clarify the effectiveness of the compensation method described in Section III-B, the preliminary experiment was conducted. Eighty sentences (10 sentences × 4 emotions × 2 speakers) uttered by two of the evaluated speakers were used for evaluation. As $p(s_i)$ in equation (2) can not be calculated accurately with a small amount of data, it was calculated by using the baseline training data. The recognition results are indicated by the word error rate (WER) in Table IV. Performance improvement can be found with a limiting rate compared to the performance without it. The best performance can be obtained at $\alpha = 0.1$. Therefore, the following experiments were conducted using this value.

B. Adaptation Experiments

The results of adaptation experiments are shown in Table V. In this table, *epoch5* means that the number of epochs is fixed to five. *Estop* means early stopping is used for adaptation, and *estop+unk* is a combination method of early stopping and unknown word countermeasure. In speaker and emotion adaptation experiments, since the number of adaptation utterances is only 40, cross-validation for early stopping could not be performed. Hence, the results related to early stopping are omitted. The results demonstrate that the recognition

TABLE V
RESULTS OF ADAPTATION EXPERIMENTS (WER[%]).

	Baseline	Speaker	Corpus	Emotion	Speaker and emotion
Epoch5	38.10	27.86	32.37	29.50	31.81
Estop		25.50	29.76	29.42	-
Estop+unk		23.05	26.91	27.01	-

TABLE VI
RELATIONSHIP BETWEEN NUMBER OF EPOCHS AND RECOGNITION PERFORMANCE (WER[%]).

Adaptation	Type	Epoch5	Estop (#epochs)	Oracle (#epochs)
Corpus	-	32.37	29.76 (1)	29.76 (1)
Emotion	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)
	average	29.50	29.42	28.84

performance without adaptation is very low (see *baseline*), whereas every adaptation method is effective.

When comparing *epoch5* and *estop*, it is found that the early stopping method is effective. The number of epochs in the *estop* experiments is large when speaker adaptation is conducted, and it depends on the speaker. The number ranges from 11 to 23. On the contrary, the number is small when corpus or emotion adaptation are conducted. Thus, since the number varies depending on adaptation method, fixing the number degrades recognition performance. Table VI shows the relationship between the number of epochs and recognition performance on corpus and emotion adaptation experiments. In this table, we describe three cases, the number is fixed to 5 (*epoch5*), early stopping is conducted (*estop*), and the number is adjusted to the optimum value (*oracle*). In the corpus adaptation, the number automatically determined is equal to the optimum number. This means that this method was very successful in this case. For emotion adaptation, WERs in *estop* experiments are similar to those in *oracle* except for *neutral*. From these results, this method is considered to be effective especially when the performance is low.

The results of *stop+unk* in Table V show that the addition of unknown words is effective. In the comparison of various adaptation methods, speaker adaptation shows the best performance. This shows that the influence of speaker characteristics is stronger than that of emotion. However, as speech recording of a specific speaker is required in advance for speaker adaptation, there is a limit in actual use. Since corpus or emotion adaptation do not depend on the speaker, the system can be used when the speaker is unknown. Thus, those are suitable for wide use. Corpus adaptation and emotion adaptation exhibit similar performance. This means that emotion-dependent model is not so effective in this experimental condition. In this experiment, four emotions, namely anger, joy, sadness, and neutral were used. However, it is difficult to classify emotions simply. There are various emotional

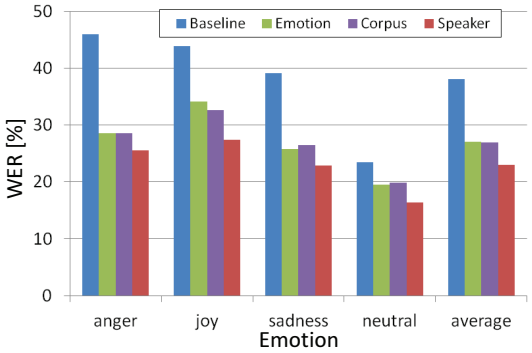

Fig. 2. Word error rate for each emotion [%].

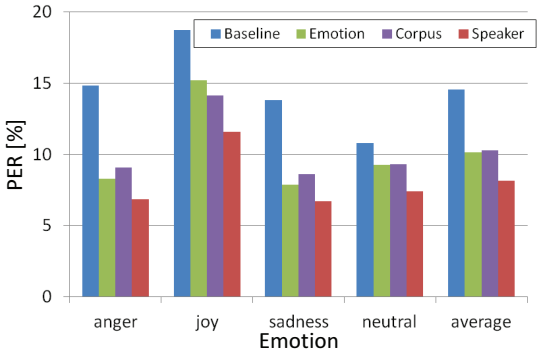

Fig. 3. Phoneme error rate for each emotion [%].

strengths even with the same emotion. To create emotion-dependent models, investigating what type of classification is effective is necessary. Performance is worse with speaker and emotion adaptation where we expected higher performance. This is due to the small amount of adaptation data (see Table II). To solve this problem, it is necessary to adopt a method that can cope with less data.

Figure 2 shows the WER for each emotion in the *estop+unk* condition. The word recognition results are converted into phoneme sequences to calculate a phoneme error rate (PER) (see Fig. 3). From the results of Fig 2, the recognition performance is apparently worse for emotional speech. However, the PER for *anger* and *sadness* indicates better performance than that for *neutral* in the speaker adaptation experiment. Thus, speaker adaptation works well for emotional speech, except for *joy*. Additionally, the mismatch between the WER and PER suggests that language models are not suitable for this task. The language models used in the system are created by lecture speech, and language model adaptation is considered necessary to improve the recognition performance for emotional speech.

VII. CONCLUSIONS

In this study, we examined emotional speech recognition using the emotional speech corpus JTES for adaptation and evaluation. In the experiments of acoustic model adaptation, the speaker, corpus, and emotion adaptation were examined. In each case, performance improvement could be obtained.

Among these methods, the best performance could be obtained with speaker adaptation. In addition, both the early stopping method and unknown word countermeasure were effective for adaptation.

As a future task, we will examine emotion adaptation further by investigating acoustic model adaptation in consideration of emotion strength rather than by simply creating emotion-dependent models. We used a simple retraining algorithm for model adaptation in this work. Because various adaptation methods exist that can be executed with a small amount of adaptation data, we will attempt to use them [17]–[20]. In this study, the language models were created from a lecture speech corpus. However, word occurrence frequency differs between lecture speech and tweets on Twitter. We intend to solve this problem in the future with language model adaptation. In addition, we would like to introduce the emotional speech recognition into the multi-modal dialog system developed by the authors [21].

Acknowledgment

This work was supported in part by a Grant-in-Aid for Scientific Research (KAKENHI 16K00227, JP15H02720) from the Japan Society for the 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. of Interspeech2016*, 2016, pp. 379–380.

[2] A. Maier, J. Hough, and D. Schlangen, "Towrrds deep end-of-turn prediction for situated spoken dialogue systems," *Proc. of Interspeech2017*, 2017, pp. 1676–1680.

[3] M. Li, Z. He, and J. Wu, "Target-based state and tracking algorithm for spoken dialogue system," *Proc. of Interspeech2016*, 2016, pp. 2711–2715.

[4] C. Liu, P. Xu, R. Sarikaya, "Deep contextual language understanding in spoken dialogue systems," *Proc. of Interspeech2015*, 2015, pp. 120–124.

[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. of Interspeech2015*, 2015, pp. 2007–2011.

[6] A. Lee, K. Oura, and K. Tokuda, "MMDAgent - a fully open-source toolkit for voice interaction systems," *Proc. of ICASSP2013*, 2017, pp. 8382–8385.

[7] R. Zhang, A. Atsushi, S. Kobashikawa, and Y. Aono, "Interaction and transition model for speech emotion recognition in dialogue," *Proc. of Interspeech2017*, 2017, pp. 1094–1097.

[8] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of German emotional speech," *Proc. of Interspeech2005*, 2005, pp.3–6.

[9] 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*, 2004, pp. 171–174.

[10] 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.

[11] 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, pp. 36–50, 2011.

[12] E. Takeishi, T. Nose, Y. Chiba, and A. Ito, "Construction and analysis of phonetically and prosodically balanced emotional speech database," *Proc. of O-COCOSDA2016*, 2016, pp. 16–21.

[13] H. Sak, A.W. Senior, and F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling," *Proc. of INTERSPEECH2014*, 2014, pp. 338–342.

[14] C. M. Bishop, *Pattern recognition and machine learning*, Springer, 2006.

[15] K. Tomita, A. Takagi, M. Kato, and T. Kosaka, "Evaluation of unsupervised cross adaptation using highly accurate models," *Proc. of ASJ2016 Autumn Meeting*, 2016, pp. 95–96, in Japanese.

[16] K.Maekawa, "Corpus of spontaneous Japanese: Its design and evaluation," *Proc. of ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, 2003, pp. 1–6.

[17] J. Xue, J. Li, D. Yu, M. Seltzer, and Y. Gong, "Singular value decomposition based low-footprint speaker adaptation and personalization for deep neural network," *Proc. ICASSP2014*, 2014, pp. 6409–6413.

[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," *Proc. ICASSP2013*, 2013, pp. 7942–7946.

[19] A. Senior and I. Lopez-Moreno, "Improving DNN speaker independence with i-vector inputs," *Proc. ICASSP2014*, 2014, pp. 225–229.

[20] L. Samarakoon and K. C. Sim, "Factorized hidden layer adaptation for deep neural network based acoustic modeling," IEEE Trans. Audio, Speech, Lang. Process., vol. 24, no. 12, pp. 2241–2250, 2016

[21] T. Koseki, and T. Kosaka, "Multimodal spoken dialog system using state estimation by body motion," *Proc. IEEE GCCE2017*, 2017, pp. 348–351.