Language Model Adaptation for Emotional Speech Recognition using Tweet data

Kazuya Saeki*, Masaharu Kato* and Tetsuo Kosaka*
* Graduate School of Science and Engineering, Yamagata University, Yonezawa, Japan
E-mail: ten36870@st.yamagata-u.ac.jp Tel: +81-238-263368

***Abstract*—Generally, emotional speech recognition is considered more difficult than non-emotional speech recognition. This is because the acoustic features of emotional speech are different from those of non-emotional speech, and these features vary greatly depending on the emotion type and intensity. In addition, it is difficult to recognize colloquial expressions included in emotional utterances using a language model trained on a corpus such as lecture speech. We have been studying emotional speech recognition for an emotional speech corpus, Japanese Twitter-based emotional speech (JTES). This corpus consists of tweets on *Twitter* with an emotional label assigned to each sentence. In this study, we aim to improve the performance of emotional speech recognition for the JTES through language model adaptation, which will require a text corpus containing emotional expressions and colloquial expressions. However, there is no such large-scale Japanese corpus. To solve this problem, we propose a language model adaptation using tweet data. Expectedly, tweet data contains many emotional and colloquial expressions. The sentences used for adaptation were extracted from the collected tweet data based on some rules. Following filtering based on these specified rules, a large amount of tweet data of 25.86M words could be obtained. In the recognition experiments, the baseline word error rate was 36.11%, whereas that of the language model adaptation was 25.68%. In addition, that of the combined use of the acoustic model adaptation and language model adaptation was 17.77%. These results established the effectiveness of the proposed method.**

I. Introduction

In recent years, speech recognition technology has improved significantly, and high performance has been reported, especially in the recognition of natural speech such as conference lectures and conversational speech. However, it is generally difficult to recognize emotional speech, because its acoustic features differ from those of ordinary speech. Further, the acoustic characteristics and duration vary greatly, depending on the emotion type and intensity. Several emotional speech corpora that can be used for such research have been constructed (e.g. [1–5]).

Recently, a Japanese Twitter-based emotional speech (JTES), which consists of phonetically and prosodically balanced utterance sets, was proposed in [6]. We developed an emotional speech recognition system for the JTES in [7]. First, we used a deep neural network-hidden Markov model (DNN-HMM) acoustic model trained on speech data in the Corpus of Spontaneous Japanese (CSJ) [8]; however, thus far, we have achieved limited recognition accuracy. This is because although the CSJ is the largest speech corpus in Japan, it contains almost no emotional speech. To solve this problem, we used an acoustic model adaptation.

Another problem is the mismatch of the language model (LM) between the CSJ and JTES. The CSJ consists of lecture speeches. There are two types of lecture speech, the speech of lectures at academic conferences and simulated lectures on various topics. These speech data do not contain much emotional expression. On the other hand, the JTES, a corpus based on tweets on *Twitter*, contains many emotional expressions, which were extracted to create the emotion corpus. To solve this problem, we examined a LM adaptation using a small amount of emotional texts, and evaluated it on emotional speech recognition tasks [9]. A total of 1,960 sentences, consisting of 490 sentences for each emotion, were used to adapt the LM. As a result, the performance of the system only improved minimally. This is because there were few adaptation data. It is necessary to examine and select the content of the data to equalize the number of data for each emotion; however, these steps are time-consuming.

The purpose of this work is to improve the performance of the emotional speech recognition on the JTES by adapting a LM using a large amount of tweet data. In constructing the LM, it is important to use a large amount of training data. Because tweet data is believed to be suffused with emotional expression, we conducted the LM adaptation using a large amount of tweet data without considering the types of emotions. Although many LM adaptations have been investigated so far, they have been mainly concerned with adaptation to specific topics [10–12]. On the other hand, LM adaptation for emotional speech recognition is rarely performed. This is because there is not much emotional data transcribed as text. In this work, we try to solve the problem by using tweet data instead of the transcription data.

Like the LM, the acoustic model (AM) is an important factor in speech recognition. This paper also examines the combined use of the LM adaptation and the AM adaptation similar to [7].

The remainder of this paper is organized as follows: Section II introduces the emotional speech corpus, the JTES. Section III describes the process of collecting tweet data and the details of the subsequent processing. Section IV describes the LM adaptation method. Section V describes the acoustic model adaptation. Section VI describes the set-up of the speech recognition experiments. Section VII describes the results of the speech recognition experiments. Section VIII presents our conclusions.

II. Emotional Speech Corpora

In this study, we used two emotional speech corpora: JTES and online gaming voice chat corpus with emotional label (OGVC) [4]. The JTES is based on tweets on *Twitter*, and comprises speech utterances by 50 males and 50 females [6]. As tweets contain many emotional expressions, it is possible to collect speech utterances with various emotions by reading out the contents emotionally. The tweets were classified into four emotion classes: joy, anger, sadness, and neutral. Phonetically and prosodically balanced sentences were selected using a sentence selection algorithm based on entropy. Finally, 50 sentences corresponding to each emotion were selected, and the emotional utterances were recorded using these sentences. The total number of utterances in the JTES is 20,000. Furthermore, 500 sentences corresponding to each of the four emotions were prepared before extracting the sentences using entropy. In our previous work, we used 1,960 of these sentences as the LM adaptation data.

In the evaluation experiments, to examine the versatility of the proposed LM, experiments were also conducted with a corpus other than JTES. The OGVC is one of the typical Japanese emotional corpora. The OGVC uses game players' voices as they play a massive multiplayer online role-playing game (MMOPRG). In an MMOPRG, players play online games and talk to each other. While concentrating on the game they utter speech containing various emotions. The OGVC is comprised of two types of speech: spontaneous and acted. The former involves recording conversations during games. In the latter, professional actors read out the transcripts of 17 dialogues extracted from gameplay conversations. When they read them, the emotion type and intensity are specified. There are eight types of emotion in the OGVC; we used the four types corresponding to those in the JTES in the experiments. There are four levels of emotional intensity from 0 to 3, wherein Level 3 is the strongest emotional expression and Level 0 indicates non-emotional expression. We used the acted speech set in the experiments.

III. Collecting Tweet Data

In this section, we will explain the outline of tweet data collection using *Twitter* API. The data used in this experiment were tweets posted to *Twitter* over the course of 51 days in May, June, and October, 2019; Japanese tweet data, excluding retweets and tweets from bots, were randomly collected. In addition, the collected tweet data included symbols such as pictograms and emoticons, and typographical errors; therefore, it was difficult to use them as they were. Accordingly, they were converted to an appropriate data format through the following process.

- URLs, hash tags, line feeds, and reply destination's "@account name" were replaced with blank.
- Assuming that punctuation marks were sentence breaks, text split was performed at those points.
- Texts with more than three words and less than 20 words were extracted from the results of the MeCab segmentation. MeCab is a Japanese morphological analysis tool [13].

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

IV. Language Model Adaptation

A two-pass decoder was used in this study as the speech recognition system, where a bigram and trigram were used for the first and second passes, respectively. These are a type of n-gram LM. In recent years, many highly expressive LMs using deep learning, such as the long short-term memory [10] and recurrent neural network [11], have been proposed. Because the purpose of this work is to verify the effectiveness of emotional utterance data for LM adaptation, the use of these LMs is a subject for further study. We used a mixed n-gram as the LM adaptation method [14]. In this method, the mixed n-gram was created through the addition of an n-gram count calculated from the baseline data (n_i^{base}) and another, calculated from an adaptation data (n_i^{adapt}). The occurrence probability of the word w_i in the adapted LM was calculated as follows:

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

where w is the weight for adjusting the imbalance between the amount of the baseline and adaptation data. N^{base} and N^{adapt} are the total number of the n-gram counts of the baseline and adaptation data, respectively.

V. Acoustic Model Adaptation

In the recognition experiments, we also attempted to combine the AM and LM adaptations. For the AM adaptation, we used the method described in our previous work [7]. In the experiments, we conducted a supervised adaptation, on the premise that each utterance was correctly labeled. The backpropagation algorithm was used for adaptation, and early stopping was introduced to automatically determine the number of epochs [15].

In the recognition step, we used the output probability compensation method [7]. In the output probability calculation, there is a problem that the occurrence probability of the state becomes extremely high with some phonemes such as silence. To solve this problem, the output probability was compensated in the recognition step. 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 $p(x)$, the occurrence probability of an input feature x, was omitted, because it did not affect the recognition result. $p(s_i)$ was the occurrence probability of the state s_i. This

value depended on the appearance frequency of a phoneme in the training data. Because phonemes, such as silence, frequently appeared in the training data, $p(s_i)$ became high. By limiting this value, the output probability can be prevented from decreasing drastically. The specific method is as follows. When $p(s_i)$ exceeded the upper limit θ, it was replaced with θ. The value θ was determined by setting the limiting rate α in (3).

$$\alpha = \frac{\sum_{i \in D}\{p(s_i) - \theta\}}{\sum_{i=1}^{I} p(s_i)}, \tag{3}$$

where I is the total number of the states, and D is the set of i that satisfies $p(s_i) > \theta$. This method is effective, especially when the adaptation data is small.

VI. Experimental condition

The experimental conditions are described in this section, in which we first describe our recognition system. In the speech analysis module, a speech signal was digitized at a sampling frequency of 16 kHz with a quantization size of 16 bits. The length of the analysis frame was 25 ms, and the frame period was set to 8 ms. A 25-dimensional feature, which comprises the log mel-filter bank features and the log power, was derived from the digitized samples for each frame. Moreover, the delta and delta-delta features were calculated from the 25-dimensional feature; hence, the total number of dimensions was 75 per frame. 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 LM. Decoding was performed using a one-pass algorithm that incorporated a frame-synchronous beam search and a tree-structured lexicon. In the second pass, the trigram LM was deployed to re-score the word graph, and the recognition result was obtained.

The input layer of the DNN used 75 coefficients, with a temporal context of 11 frames, making a total of 825 input features. The DNN had seven hidden layers, each containing 2048 hidden units. The total number of states for the shared-state triphone is 3003. The final output layer had 3003 units, corresponding to the total number of states. The speech data of 963 lectures in the CSJ were used to train the DNN-HMM. The total length of speech was approximately 203 h. The DNN was trained 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 assigned to each frame, and the backpropagation algorithm with stochastic gradient descent was used. The cross entropy was used as the loss function. Other conditions for the DNN training are shown in Table I.

The bigram and trigram models were used as the LMs. The baseline models were trained on a textual data containing 2668 lectures from the CSJ, and the total number of words was 6.68 million. We used two adaptation LMs, to compare the effectiveness of the amount of training data. One was a LM adapted using a large amount of tweet data (25.86 million words), and the other LM was adapted using the small amount of data (17,711 words) introduced in the previous work [9]. In the remainder of this paper, we will refer to the former as *large-scale LM* and the later as *small-scale LM*. Using the adaptation method described in Section IV, w was set to 100 for the bigram, and 30 for the trigram for *small-scale LM*, and two for each for the *large-scale LM*. These values were determined based on the test set perplexity criterion. Table II and Table III show the test set perplexity of the evaluation data using the *small-scale LM* and *the large-scale LM*, respectively. The test set perplexity is greatly improved by using the *large-scale LM*, but it is still high.

A lexicon was created based on the words that appeared in the CSJ corpus. To eliminate unknown words, 44 words that appeared only in the evaluation data were added to the lexicon.

The AM adaptation was conducted to adapt the corpus. Specifically, the AM was adapted to the acoustic environment of the JTES corpus. The DNN was adapted using a backpropagation algorithm such as fine-tuning. The detailed conditions of the DNN adaptation are shown in Table IV. We used 14,400 (40 utterances $\times$ 4 emotions $\times$ 90 speakers) data to adapt the AM to the JTES.

The evaluation data, which were different from the adaptation data, comprised 400 utterances (10 utterances $\times$ 4 emotions $\times$ 10 speakers) from the JTES. All the experiments on the JTES were performed using these evaluation data. In the evaluation of the OGVC, utterances of 0 and 3 intensities were used. For each intensity, 448 utterances (112 utterances $\times$ 4 speakers) were used.

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 increase by less than 0.1%.
Mini-batch size	512

TABLE II
Test set perplexity with small-scale LM

Weight	10	30	50	100	150	200
Bigram	995.69	901.67	869.84	849.22	854.74	869.12
Trigram	931.60	907.14	915.17	960.64	1014.71	1070.22

TABLE III
Test set perplexity with large-scale LM

Weighgt	0.1	0.5	1	2	3	4
Bigram	328.76	293.81	291.19	271.77	274.07	276.95
Trigram	312.17	250.83	242.24	224.10	228.96	276.23

TABLE IV
ADAPTATION CONDITIONS FOR DNN

Mini-batch size	2048
Momentum	0.0
L2 regularization factor	0.0002
#epochs	The process terminates when the frame accuracy increase by less than 0.5% for early stopping experiments.

VII. RECOGNITION EXPERIMENTS

The recognition experiments were conducted using only the LM adaptation and the simultaneous adaptation of the LM and AM. The results for the LM adaptation with an unadapted AM is shown in Table V. In the experiments, we compared the baseline LM trained on the CSJ with the *small-scale LM* and *large-scale LM*. It can be seen that the *large-scale LM* yielded the best results. For the *large-scale LM*, the types and presence/absence of emotions were not considered in collecting the adaptation data. Nevertheless, the *large-scale LM* outperformed the *small-scale LM* in which emotions were considered. This suggested that the amount of data was important.

The results for the simultaneous adaptation of the LM and AM are shown in Table VI. Overall, we can see that the results, as shown in Table VI, were better than those shown in Table V. This demonstrates the effectiveness of the simultaneous adaptation. Compared with the results of the AM adaptation alone, the performance of the *large-scale LM* adaptation was slightly higher. The word error rate (WER) of the former was 26.91%, and that of the latter was 25.68%. Finally, the simultaneous adaptations yielded the best result of 17.77%. The fact that simultaneously adapting the AM and LM improved the performance significantly suggested that the adaptations of the AM and LM had different characteristics.

Next, we describe the analysis of the recognition results. Table VII shows the difference between the recognition results of the *small-scale LM* and *large-scale LM*. In the first example, a consonant /Q/ was missing in the *small-scale LM* condition. /Q/ is a geminate stop consonant expressed by a one-mora pause; it is a consonant peculiar to Japanese. In this case, although this consonant does not appear in typical vocalization, it tends to appear when emotions are being emphasized. In the next example, a postpositional particle *ga* was inserted, in the case of the *small-scale LM*. Postpositional particles are parts of speech unique to Japanese. In Japanese grammar, *ga* occupies this position. However, postpositional particles are often dropped when colloquial expressions are used with emotion. Based on the above points, the *large-scale LM* is thought to express emotions and the colloquial style effectively. In the last example, an auxiliary verb *na* was missing in the *small-scale LM* condition. In addition, due to the elimination of *na*, a morphological analysis error occurred. The word *ni* and *komi* are merged into one word.

Specifically, it is difficult to recognize fillers in the colloquial style. In Japanese, fillers are often articulated as long vowels. Recognition errors related to them were often not improved by the *large-scale LM*. The frequency of this kind of error depended on the speaker. The factors that caused differences in the recognition results, depending on the speakers, were considered to be mainly the individual's acoustical characteristics. Therefore, these errors were thought to be caused by the AM, rather than the LM. In emotional speech, these long vowels as fillers are often devocalized. To recognize these pronunciations successfully, it is necessary to correctly recognize the unvoiced information. As a solution, for example, a recognition method that prepares multiple AMs based on the emotion type and intensity can be considered.

In the experiments described above, both AM and LM were adapted by the JTES data. Although the environment for each evaluation utterance is open, we were concerned that these models may have adapted to the wording of the tweet data rather than to emotions. In consideration of this possibility, an additional evaluation was performed using the OGVC to vary the emotional conditions entirely. The results of the LM only adaptation are shown in Table VIII, and those of the LM and AM's simultaneous adaptation are shown in Table IX. From the results it appears that the performance improvement in the *small-scale LM* is limited; however, sufficient improvement is seen 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 in identifying emotions. From the comparison of Table VIII and Table IX, it is confirmed that the AM adapted by the JTES is effective for the OGVC.

TABLE V
WORD ERROR RATE FOR EACH EMOTION USING BASELINE AM AND VARIOUS LMS (%)

	Baseline LM	Small-scale LM	Large-scale LM
Ang	40.97	36.26	28.61
Joy	41.23	31.19	30.31
Neu	23.13	20.60	16.39
Sad	39.09	35.77	27.42
Ave	36.11	30.95	25.68

TABLE VI
WORD ERROR RATE FOR EACH EMOTION, USING CORPUS-ADAPTED AM AND VARIOUS LMS (%)

	Baseline LM	Small-scale LM	Large-scale LM
Ang	28.21	28.29	20.70
Joy	32.62	24.24	21.85
Neu	19.88	18.26	12.65
Sad	26.52	25.82	15.91
Ave	26.91	24.15	17.77

TABLE VII
EXAMPLE OF DIFFERENCES IN RECOGNITION RESULTS BY VARIOUS LMS. SYSTEM USED ADAPTED AM. /Q/ IS A GEMINATE STOP CONSONANT.

Correct	kore baQkari
Meaning	Only this
Small-sacle 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 Adapt	heizitsu no nikomi sugi
Tweet Adapt	heizitsu na no ni komi sugi

TABLE VIII
WORD ERROR RATE FOR OGVC USING BASELINE AM AND VARIOUS LMS (%)

Intensity level	Baseline LM	Small-scale LM	Large-scale LM
0	32.61	31.99	24.39
3	52.97	51.73	46.81

TABLE IX
WORD ERROR RATE FOR OGVC USING CORPUS-ADAPTED AM AND VARIOUS LMS (%)

Intensity level	Baseline LM	Small-scale LM	Large-scale LM
0	28.02	27.56	21.10
3	43.12	43.33	37.30

VIII. CONCLUSIONS

In this study, we investigated the possibility of improving speech recognition accuracy for the JTES (an emotional speech corpus) using LM adaptation. To improve the recognition performance through LM adaptation, a large-scale textual data containing emotional expressions and colloquial expression was required. Accordingly, we proposed an LM adaptation using tweet data. From the recognition experiments, both the LM and AM adaptations yielded good results. Finally, the baseline WER was 36.11%, whereas that of the simultaneous AM and LM adaptation was 17.77%. Based on these results, the effectiveness of the proposed method was established. In addition, the proposed LM demonstrated a performance improvement even in a dissimilar emotional environment thereby confirming its versatility.

As a future task, we will examine emotion adaptation further by investigating the AM adaptation in relation to emotion intensity, rather than by simply creating emotion-dependent models. For the LM adaptation, we will attempt to use a neural network (NN)-based LM, instead of the n-gram model, because the use of the NN-based LMs is expected to improve the performance. In addition, we plan to improve the emotion recognition system that is being developed in our laboratory using the speech recognition outputs examined in this work. Furthermore, we are considering introducing this recognition system into the multimodal dialogue system [17].

ACKNOWLEDGMENT

This study was supported in part by a grant-in-aid for scientific research (KAKENHI 19K12014) from the Japan Society for the Promotion of Science. We thank Dr. Takashi Nose of Tohoku University for providing us with the Japanese Twitter-based emotional speech (JTES).

REFERENCES

[1] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J.N. Chang, S. Lee, and S.S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, pp. 335–359, Dec. 2008.

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

[3] 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 LREC*, 2004, pp. 171–174.

[4] 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, Jun. 2012.

[5] 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. 2, pp. 36–50, Jan. 2012.

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

[7] T. Kosaka, Y. Aizawa, M. Kato and Takashi Nose, "Acoustic model adaptation for emotional speech recognition using Twitter-based emotional speech corpus," *Proc. of APSIPA ASC*, 2018, pp. 1747–1751.

[8] S. Furui, M. Nakamura, T. Ichiba, K. Iwano, "Analysis and recognition of spontaneous speech using corpus of spontaneous Japanese," *Speech Communication*, vol. 47, no. 1–2, pp. 208–219, Sept.–Oct. 2005.

[9] K.Saeki, M.Kato and T. Kosaka, "Performance improvement of prosody-controlled voice conversion by language model adaptation ," *Proc. of IEEE GCCE*, 2019, pp. 854–856.

[10] M. Sundermeyer, R. Schluter, and H.Ney, "LSTM neural networks for language modeling," *Proc. of Interspeech*, 2012, pp. 194–197.

[11] T. Mikolov, M. Karafiat, L. Burget, J.Cernock, and S. Khudan, "Recurrent neural network based language model," *Proc. of Interspeech*, 2010, pp. 1045–1048.

[12] M.Ma, M.Nirschl, F.Biadsy and S.Kumar, "Approaches for neural-network language model adaptation," *Proc. of Interspeech*, 2017, pp. 259–263.

[13] Taku Kudo, Kaoru Yamamoto and Yuji Matsumoto, "Applying conditional random fields to Japanese morphological analysis," *Proc. of EMNLP*, 2004, pp.230–237.

[14] 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, Nov. 2000 (in Japanese).

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

[16] Tetsuo Kosaka, Yuka Haneda, Daisuke Makabe and Masaharu Kato, "Investigation of acoustic models for emotion recognition using a spontaneous speech corpus," *Proc. of 23nd International Congress on Acoustics*, 2019, pp. 1–6.

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