Effect of Training Data Selection for Speech Recognition of Emotional Speech

Yusuke Yamada, Yuya Chiba, Takashi Nose, and Akinori Ito

***Abstract*—In this paper, we describe the speech recognition from emotional speech. The task treated in this paper is not an emotion recognition from speech but a speech recognition (speech to text) from a speech that contains distinct emotion. First, we compare two acoustic models trained from neutral speech and emotional speech. We expected that the acoustic model trained from emotional speech improves the recognition performance of emotional speech, but the result showed that a larger amount of neutral speech was more effective than a small amount of emotional speech. Next, we applied the data selection method to enhance the phonetic balance of the training data. As a result, the entropy-based selection from the training data enhanced the recognition performance when there is some domain mismatch between the training data and the evaluation data.**

***Index Terms*—Deep neural network, data selection, emotional speech, speech recognition, training data reduction.**

I. Introduction

With the development of deep learning, speech recognition has become more and more accurate in this decade [1], and is widely used in various fields [2]. Along with the expansion of the application field of speech recognition, systems that use speech recognition accept utterances with variations such as gender, age, and emotion.

Emotion recognition from speech [3]-[5] is one of the emergent research fields in speech processing studies. From the viewpoint of speech-based application, it is known that changing the system's response according to a user's emotion estimated from speech can improve the user's impression of the system [6].

When responding to an utterance that contains distinct emotion, not only emotion recognition but also the speech recognition is needed. However, only a few studies treat how the emotional variation contained in an utterance affects the performance of speech recognition [7], [8].

In this paper, we investigate how emotion in an utterance affects speech recognition performance, and investigate how we can improve the recognition accuracy of emotional speech.

Besides, we investigated the performance of training data selection. The training data is not necessarily phone-balanced, which may affect the recognition performance. Therefore, we conducted experiments on selecting the training data considering the variation of phone distribution.

Manuscript received April 21, 2020; revised January 11, 2021. Part of this work is supported by JSPS KAKENHI 17H00823.

Yusuke Yamada, Yuya Chiba, Takashi Nose are with Graduate School of Engineering, Tohoku University, Sendai, 980-8579 (e-mail: y.yamada@spcom.ecei.tohoku.ac.jp, yuya@spcom.ecei.tohoku.ac.jp, nose@tohoku.ac.jp).

A. Ito is with Graduate School of Engineering, Tohoku University, and Tough Cyberphysical AI Research Center, Tohoku University, Sendai, 980-8579 (e-mail: aito@spcom.ecei.tohoku.ac.jp).

II. Speech Corpora

Before explaining the method, we first explain the corpora we used. In this study, we use the following corpora.

A. JTES

The Japanese Twitter-based Emotional Speech (JTES) [9] is a database that contains acted emotional speech. It has 200 distinct Japanese sentences, 50 for each of the emotions (neutral, joy, anger, sad), taken from Twitter's tweets. One hundred speakers (50 males and 50 females) read the prepared sentences in the designated emotion.

B. CSJ

The Corpus of Spontaneous Japanese (CSJ) [10] is a large-scale spontaneous Japanese speech corpus. The corpus contains 300-hour academic presentation speech and 330 public speech; both of them have manual transcriptions. We used the academic presentation speech as a source of acoustic model training.

C. OGVC

The Online Game Voice Chat (OGVC) corpus [11] is a Japanese speech corpus collected by voice chat of an online game. It contains many natural emotional utterances with corresponding emotional labels given by experts. The utterances are a part of dialogue or trialogue among online game players. It contains 9114 utterances.

D. JNAS

The Japanese Newspaper Article Sentences (JNAS) [12] is a reading speech database. The sentences were selected from articles of Mainichi Shimbun newspaper articles and a phoneme-balanced sentence set. 306 speakers (153 males and 153 females) read the sentences, and the total length of the speech is about 60 hours.

III. Baseline Experiment of Emotional Speech Recognition

In the first experiment, we investigated that ordinary spontaneous speech (CSJ) can recognize the speech with emotion (JTES). At the same time, we tested whether using an emotional speech database as the training data contributed to the performance of speech recognition.

A. Experimental Condition

We used CSJ and JTES as training data. Table I shows the

doi: 10.18178/ijmlc.2021.11.5.1062

conditions and amount of training data of the acoustic model. The amount of CSJ is almost 16 times as large as JTES. The test data was taken from JTES, where the speakers and sentences in the test set were different from those in the training set. 800 utterances (40 sentences uttered by 20 speakers, 10 females and 10 males) were involved in the test set.

We used the Kaldi toolkit [13] to train the acoustic model and recognize the speech.

Table II shows the conditions for training the DNN-HMMs for both training data sets. The input vector is calculated from the 13-dimensional Mel-Frequency Cepstral Coefficients (MFCC) with its first and second derivatives. The 39-dimensional vector is combined with nine adjacent frames (117 dimensions) and compressed into 40 dimensions using the linear discriminant analysis (LDA). After that, the maximum likelihood linear transform (MLLT) and frame-based maximum likelihood linear regression (fMLLR) were applied to the features, and finally, the model was trained using the speaker adaptive training (SAT). After calculating the features, 35 contiguous features are input to the network (1400 dimensions).

The DNN was pre-trained using the restricted Boltzmann machine (RBM) training and then trained using the backpropagation training with cross-entropy loss.

The language model was trained from 72.2M tweets from Twitter data, which did not contain sentences in JTES. The vocabulary size was 65k, and the low-frequency words were substituted with an UNK symbol. Then a bigram and trigram model with the Good-Turing discounting were trained.

TABLE I: TRAINING DATA

	ALL_CSJ		ALL_JTES	
	Female	Male	Female	Male
# speakers	809	177	40	40
# utterances	132k	30k	6.4k	6.4k
Length [h]	189.57	50.57	7.87	7.28

TABLE II: DNN-HMM CONDITIONS

	ALL_CSJ	ALL_JTES
# input nodes	1400	
# hidden layers	6	
# hidden nodes	1905	
# output nodes	9270	6292
Activation function	Sigmoid(hidden)/Softmax(output)	

B. Results

Fig. 1 shows the experimental results. The performance of speech recognition was measured using the character error rate (CER). Here, ANG, JOY, SAD, and NEU denote the emotions Angry, Joy, Sad, and Neutral, respectively. The label ALL shows the average of all the four emotion results. From this result, it is evident that the CSJ-based acoustic model gave better performance against the JTES-based acoustic model. The reason may be that variation of sentences in JTES is very limited, and thus it may be difficult to cover all the phoneme environment (a phoneme and the preceding and following phonemes) of the test data. Another observation is that NEU (neutral emotion) utterances were easier to recognize than other emotions. It is natural for the CSJ acoustic model because all of the utterances in CSJ have neutral emotion. As for the JTES acoustic model, this result can be understood that the utterance with neutral emotion may be near to the average of all utterances with various emotions, and thus the acoustic model trained with various emotions will become similar to the neutral emotion.

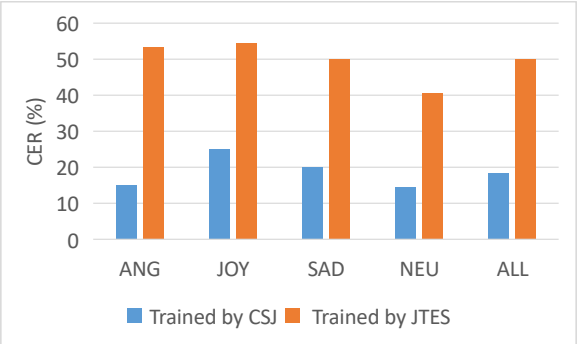


Fig. 1. Experimental results for JTES.

IV. SELECTION OF TRAINING DATA

A. Entropy-Based Sentence Selection

From the previous result, we found that a large corpus is better than a small corpus even if the contents of the training corpus do not match with the evaluation speech. However, it is also known that choosing samples from the training data considering the distribution of phonemes sometimes improves the quality of the trained model even if the selection process reduces the size of the training data [14], [15]. Thus, we tested whether the selection of the training data could improve the recognition performance for emotional speech.

We employed the data selection algorithm proposed by Nose *et al*. [16]. This algorithm selects sentences from a given sentence set so that the entropy of the selected sentences becomes nearly maximum. The selection is based on the entropy of diphones (pairs of contiguous phonemes).

Let Ω be the set of all sentences, $s \in \Omega$ be a sentence, Σ be a set of all diphones, $n(x,s)$ be the frequency of diphone $x \in \Sigma$ in sentence s. Then the maximum likelihood estimate of occurence of x in s is

$$p(x \mid s) = \frac{n(x,s)}{|s|} \tag{1}$$

where $|s|$ is the length of sentence s. Then the entropy of sentence s is calculated as follows.

$$H(s) = -\sum_{x \in \Sigma} p(x \mid s) \log p(x \mid s) \tag{2}$$

Selection of sentences is performed using a greedy algorithm since the optimum selection is not computationally feasible. The selection procedure is simple. First, we calculate entropy values for all of the sentences in the sentence set. Next, we sort the sentences in the descending order of the entropy value. Finally, we select the desired number of sentences from the sentence with the largest entropy. Although this algorithm is not optimum, this method is fast enough to apply to a large corpus.

B. Experiment

We conducted an experiment to confirm the effect of sentence selection. In this experiment, we first determine the

fraction of sentences (e.g. 25%), and the designated fraction of sentences were selected. Then the speech data that correspond to the selected sentences are collected as the training data. We chose 5%, 25%, 50%, and 75% of all data using the entropy-based method. The training data is the same as the ALL_CSJ data shown in Table I.

In this experiment, we used all of JTES corpus as the test set. As a result, 200 sentences uttered by 100 speakers (50 males and 50 females) were used (20k utterances in total).

Fig.2 shows the result of all emotions. We can see that the entropy-based selection slightly improved the CER.

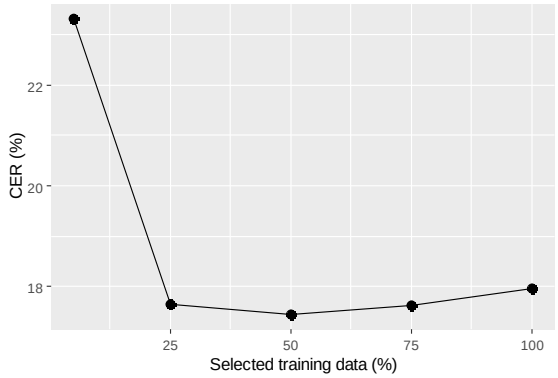


Fig. 2. Selection result (All emotions).

Fig. 3 shows the same result for specific emotions. We can see that the effect of training data selection differs from emotion to emotion. The emotions other than ANG showed some improvement when the training data was properly selected.

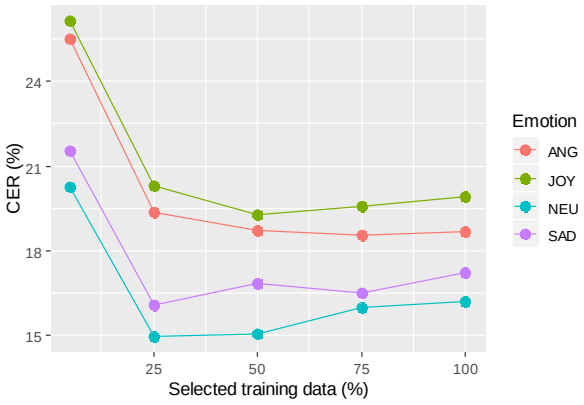


Fig. 3. Selection result (Emotion by Emotion).

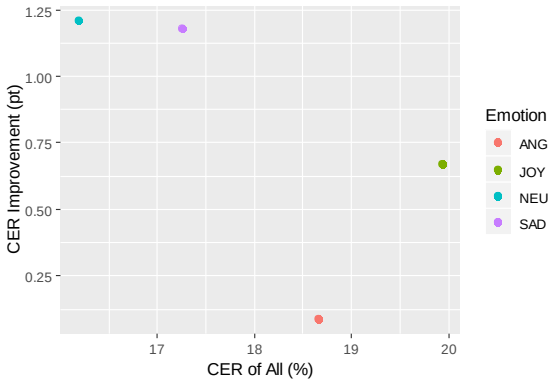


Fig. 4. CER by all training data vs. maximum absolute improvement.

Fig. 4 shows the CER when using all training data and maximum absolute improvement (in points). The effect of training data selection was higher for NEU and SAD, while that for ANG and JOY were limited.

V. Application to Other Emotional Speech

The result obtained in the previous section is not what we expected. Usually, using more data leads to better recognition performance. However, the obtained result is contrary to the expectations. Therefore, we compared the result to the recognition result of another emotional speech taken from the OGVC corpus.

We used 7298 utterances from the corpus (5393 male utterances and 1905 female utterances, about 2 h in total). The acoustic model and the language model were the same as those used in the previous chapter.

Fig. 5 shows the experimental result. Since the utterances in this corpus are spontaneous utterances, the error rate is much higher than that of JTES. However, the entropy-based method could achieve a slight improvement over the full use of the training data.

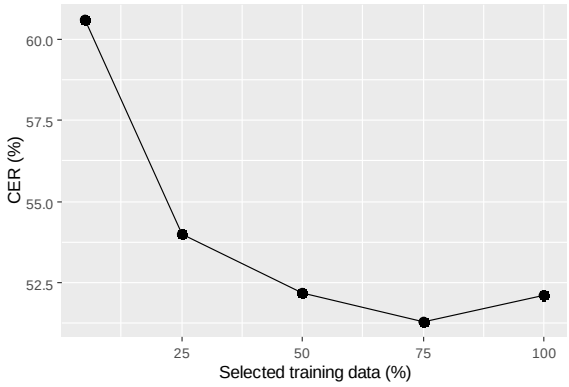


Fig. 5. Result for OGVC.

VI. Comparison with Non-emotional Speech Recognition

Finally, we tried to apply the same training data selection method to a non-emotional speech to investigate whether the improvement by the training data selection is related to emotional speech.

We used JNAS as the test data. Ten male speakers and ten female speakers were selected as the test speakers. We chose 2098 utterances (4.23 h) from newspaper article sentences. The acoustic model and the language model were the same as those used in the previous experiment.

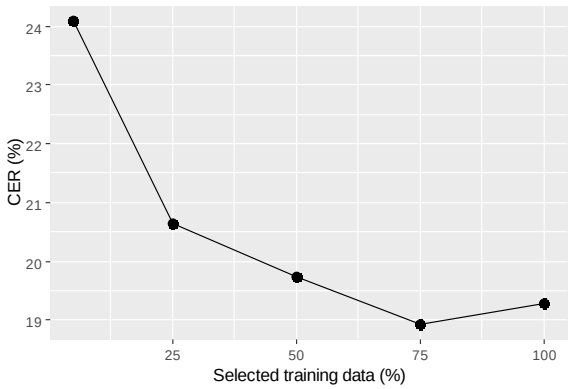


Fig. 6. Result for JNAS.

Fig. 6 shows the experimental result. The entropy-based selection improved the CER, which is as expected. This result shows that the framework of the experiment is universal.

VII. Discussion

From the experimental results shown in the previous sections, it is confirmed that training data selection improved the recognition performance. However, as Fig. 3 shows, the effect of the data selection differed from emotion to emotion. Especially, it was shown that ANG and JOY, the angry and joy emotions belong to "high arousal" emotion [17], which have distinct features from "low arousal" emotional speech. The present results suggest that a large amount of training data can be used for recognizing speech with similar emotional features, but we need to add other training data or make an adaptation for recognizing speech with different features.

The results for emotional speech OGVC (Fig. 5) and JNAS (Fig. 6) look similar and are different from those for JTES (Fig. 2). OGVC is a corpus of emotional speech just like JTES, but the difference is that OGVC contains spontaneous speech, and JTES contains acted speech. The intensity of emotional utterances contained in OGVC is marginal [11], which might be one reason why the result of OGVC was different from JTES.

One possible way to improve the recognition performance of emotional speech is to perform adaptation to emotions with high arousal [7], [8]. We should investigate how the training data selection and emotion adaptation benefits speech recognition accuracy.

VIII. Conclusion

In this paper, we investigated whether the non-emotional speech was useful for recognizing emotional speech. As a result, a sufficient amount of spontaneous speech was more beneficial than a small amount of emotional speech. Then we investigated the effect of training data selection on emotional and non-emotional speech. As a result, the training data selection was effective for both kinds of speech, but it was also revealed that the effect of training data selection was limited for emotions with high arousal, which might be caused by a mismatch between the training and test data.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Y. Y. conducted all of the experiments; Y. C. supervised all of the experiments; T. N. initiated the research and designed the total research. A. I. supervised all of the research and wrote the paper; all authors had approved the final version.

References

[1] A. B. Nassif, I. Shahin, I. Attili, M. Azzeh, and L. Shaalan, "Speech recognition using deep neural networks: A systematic review," *IEEE Access*, vol. 7, pp. 19143–19165, February 2019.

[2] Y. Ning, S. He, C. Xing, and L.-J. Zhang, "The development trend of intelligent speech interaction," in *Proc. Int. Conf. on Cognitive Computing 2019 (ICCC 2019)*, pp. 169–179.

[3] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, 2011.

[4] K. Wang, N. An, B. N. Li, Y. Zhang, and L. Li, "Speech emotion recognition using Fourier parameters," *IEEE Transactions on Affective Computing*, vol. 6, no. 1, 69–75, January 2015.

[5] S. Mirsamadi, E. Barsoum, and C. Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," In *Proc. 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2017)*, pp. 2227–2231.

[6] M. Yamanaka, Y. Chiba, T. Nose, and A. Ito, "A study on a spoken dialogue system with cooperative emotional speech synthesis using acoustic and linguistic information," in *Proc. 13th Int. Conf. on Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP 2018)*, pp. 101–108.

[7] K. Mukaihara, S. Sakti, and S. Nakamura, "Recognizing emotionally coloured dialogue speech using speaker-adapted DNN-CNN bottleneck features, in *Proc. SPECOM: International Conference on Speech and Computer*, 2017, pp. 632–641.

[8] T. Kosaka, Y. Aizawa, M. Kato, and T. Nose, "Acoustic model adaptation for emotional speech recognition using Twitter-based emotional speech corpus," in *Proc. 2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC 2018)*, pp. 1747–1751.

[9] E. Takeishi, T. Nose, Y. Chiba, and A. Ito, "Construction and analysis of phonetically and prosodically balanced emotional speech database," in *Proc. 2016 Conference of the Oriental Chapter of International Committee for Co-ordination and Standardization of Speech Databases and Assessment Techniques (O-COCOSDA)*, 2016, pp. 16–21.

[10] K. Maekawa, "Corpus of Spontaneous Japanese: Its design and evaluation." in *Proc. 2003 ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition (SSPR 2003)*, pp. 7–12.

[11] Y. Arimoto, H. Kawatsu, S. Ohno, and H. Iida, "Naturalistic emotional speech collection paradigm with online game and its psychological and acoustical assessment," *Acoustical Science & Technology*, vol. 33, no. 6, pp. 359–369, November 2012.

[12] K. Itou, M. Yamamoto, K. Takeda, T. Takezawa, T. Matsuoka, T. Kobayashi, K. Shikano, and S. Itahashi, "JNAS: Japanese speech corpus for large vocabulary continuous speech recognition research," *Journal of the Acoustical Society of Japan (E)*, vol. 20, no. 3, pp. 199–206, May 1999.

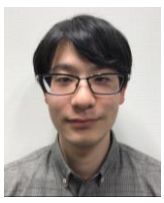
[13] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi speech recognition toolkit," in *Proc. IEEE 2011 Workshop on Automatic Speech Recognition & Understanding (ASRU)*, 2011.

[14] Y. Wu, R. Zhang, and A. Rudnicky, "Data selection for speech recognition," in *Proc. 2007 IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU)*, 2007, pp. 562–565.

[15] T. Asami, R. Masumura, H. Masataki, M. Okamoto, and S. Sakauchi, "Training data selection for acoustic modeling via submodular optimization of joint Kullback-Leibler divergence," in *Proc. INTERSPEECH 2015*, pp. 3645–3649.

[16] T. Nose, Y. Arao, T. Kobayashi, K. Sugiura, and Y. Shiga, "Sentence selection based on extended entropy using phonetic and prosodic contexts for statistical parametric speech synthesis," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 5, pp. 1107–1116, March 2017.

[17] L. F. Barrett, "Discrete emotions or dimensions? The role of valence focus and arousal focus," *Cognition & Emotion*, vol. 12, no. 4, pp. 579–599, July 1998.



Yusuke Yamada received his B.E. and M.E. degrees from School of Engineering, Tohoku University, Japan, in 2016 and 2017, respectively.

He had been engaged in development of speech recognition systems and speech synthesis technology.

Yuya Chiba received B.E., M.E. and Ph.D. degrees in engineering from School of Engineering, Tohoku University, Sendai, Japan in 2010, 2012, and 2015, respectively.

He is currently an assistant professor at the Graduate School of Engineering, Tohoku University, Sendai, Japan. His research interests include spoken dialogue systems, multimodal dialogue systems, and human-centric interfaces.

Prof. Chiba received the IEICE ISS Young Researcher's Award in Speech Field in 2014. He is a member of ISCA, ACL, IEICE, and ASJ.

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 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 of the Graduate School of Engineering, Tohoku University. His research interests include speech synthesis, speech recognition, spoken dialogue system, and music information processing.

Prof. Nose is a member of IEEE, ISCA, IEICE, IPSJ, and ASJ.

Akinori Ito was born in Yamagata, Japan in 1963. He received the B.E., M.E. and Ph.D. degrees from Tohoku University, Sendai, Japan, in 1984, 1986 and 1992 respectively.

He joined the Research Center for Applied Information Sciences, Tohoku University as an assistant professor in 1992, and then moved to Education Center for Information Processing, Tohoku University in 1993. In 1995, he joined the Faculty of Engineering, Yamagata University, Japan as a lecturer. He is now a professor of Graduate School of Engineering, Tohoku University, Sendai, Japan. He has engaged in spoken language processing, speech signal processing and music information processing.

Prof. Ito is currently the president of the Acoustical Society of Japan (ASJ), fellow of the Institute of Electronics, Information and Communication Engineers (IEICE), and a member of the Information Processing Society Japan, Human Interface Society and the IEEE.