PROCEEDINGS of the
23rd International Congress on Acoustics

9 to 13 September 2019 in Aachen, Germany

Investigation of acoustic models for emotion recognition using a spontaneous speech corpus

Tetsuo KOSAKA[1]; Yuka HANEDA; Daisuke MAKABE; Masaharu KATO

Yamagata University, Japan

ABSTRACT

In this study, we aimed to improve emotion recognition performance for spontaneous emotional speech. While many studies regarding emotion recognition for acting speech have been conducted, few have focused on spontaneous emotional speech because its speech corpora are not well developed. Recently, an online gaming voice chat corpus comprising spontaneous emotional utterances was developed. We conducted emotion recognition experiments to classify five emotions with this corpus. In our recognition system, we used acoustic features standardized in the *Interspeech 2009 Emotion Challenge*, and a deep neural network that exhibited a feed-forward type of architecture as an acoustic model. In the recognition experiments, we investigated the amount and quality of training data. Generally, it is necessary to increase the amount of training data to improve the recognition performance. However, a mere increase in the amount of training data resulted in poor performance in our experiments. The intensity of emotional expression is often low in spontaneous emotional speech, while it is high in acting emotional speech. We found that it was effective to select utterances with low emotional intensity from the acting speech corpus and add them to the training data.

Keywords: Emotion recognition, Emotional speech, Deep learning

1. INTRODUCTION

It is important to convey emotions in both human–human and human–machine interactions. Conveying emotions is performed by information obtained visually such as facial expression or obtained from speech. Emotions as nonverbal information included in speech are mainly represented by prosody like strength, pitch, and duration. Automatic emotion recognition by speech has been realized using classifiers such as support vector machine (SVM) (1). Currently, the advancement of deep learning technology has enabled significant research progress (2,3,4,5).

Conventionally, acted speeches by professional actors/actresses have been used primarily for emotion recognition. In this case, emotional speech data are collected by reading out prepared sentences with specified emotions. Using this method, emotional types are not required to be labeled by listening to the speech later. Therefore, speech data collection can be performed relatively easily. However, the emotion is often expressed more excessively than usual in this method. Furthermore, it is thought that acoustic features are different from ordinary emotional speech. Studies based on acted emotional speech are considered difficult to apply in the real world.

Some research groups have conducted studies regarding spontaneous emotional speech (6,7,8). However, it is generally difficult to collect a large amount of spontaneous emotional speech. Meanwhile, a large amount of data is required to create acoustic models for the recent statistical processing or deep learning. The interactive emotional dyadic motion capture database is an example of an effort to collect a large amount of spontaneous emotional speech (9). For developing this database, scenarios were prepared in advance to convey emotions that are more natural. For example, a scenario that induces anger was established to create a situation where the speaker created angry speeches. However, the scenario was set in advance, and professional actors/actresses were used to convey clear emotional expressions in this case. Therefore, it is thought that this speech corpus is different from natural emotion expression. Another example is the online gaming voice chat corpus with emotional label (OGVC) (10). OGVC uses game players' speech where they play a massive multiplayer online role-playing game (MMOPRG). In an MMOPRG, players play online games

[1] tkosaka@yz.yamagata-u.ac.jp

while talking with each other. They utter speech containing various emotions by concentrating on the game. Speech is uttered by people such as nonprofessional ordinary students. The corpus was completed by adding an emotional label to each utterance after collecting the emotional data. Because this corpus comprises spontaneous emotional speech uttered by nonprofessional speakers, highly accurate emotion recognition is considered difficult to achieve. Studies regarding spontaneous emotion recognition are scarce. In this study, we conducted basic studies of spontaneous emotion recognition based on a deep neural network using an OGVC. We compared the recognition methods, type of training data, and effect of emotion intensity.

2. EMOTION RECOGNITION METHOD

Figure 1 shows a block diagram of emotion recognition. First, low-level-descriptor (LLD) features as acoustic features are extracted from an input speech. A feature vector for the utterance is obtained by calculating various statistics from a time sequence of the LLD features. Emotion estimation is performed with a feedforward deep neural network (DNN) or an SVM using these statistics as input features. Generally, emotional expressions are classified into two types: using categories such as anger or sadness or using coordinates in emotional space with a pleasantness axis and an activation axis. In this work, we use the former five-category representation with categories such as “anger,” “joy,” “sadness,” “surprise,” and 'neutral'.

Emotional characteristics appear primarily in prosodic information such as fundamental frequency (F0). They are also related to the spectral features. Therefore, features called LLD shown in Table 1 are used. In the table, ZCR represents zero-crossing rate, and HNR represents harmonics-to-noise ratio. The LLD features are calculated for each frame and are used as a time sequence. Because emotions are considered as being expressed in the entire utterance, various statistics are calculated from the LLD sequence and are used as acoustic features of the utterance. In this study, we used a 384-dimensional feature vector that was standardized in *INTERSPEECH 2009 Emotion Challenge*, comprising a 32-dimensional LLD and 12 statistics ($32 \times 12 = 384$).

We used a feedforward DNN as a classifier based on deep learning. Parameter training was performed in two steps: unsupervised pretraining and supervised fine-turning. Pretraining was performed to avoid obtaining a local optimal solution and to obtain appropriate initial values. A restricted Boltzmann machine was used for the pretraining where unsupervised training was performed layer by layer. Fine-turning was performed by a backpropagation algorithm based on the stochastic gradient descent method by assigning a correct label for each frame. Cross entropy was used as a loss function. In the recognition step, an acoustic likelihood was calculated by scaling based on Bayes' rule, shown as follows:

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

where $p(\boldsymbol{x})$ is the occurrence probability of the input features, $p(s_i|\boldsymbol{x})$ is the output of the DNN for state s_i, and $p(s_i)$ is the state occurrence probability. The likelihood $p(\boldsymbol{x}|s_i)$ can be obtained at every frame, and its total sum for all frames in an utterance is calculated. Finally, the emotion that shows the highest value is determined as the recognition result.

For comparison, we investigated an SVM as a classifier that has been widely used for emotion recognition. An SVM can address linear nonseparable data using a kernel function. In this study, sequential minimal optimization was used as a training algorithm, and a polynomial kernel was used as the kernel function.

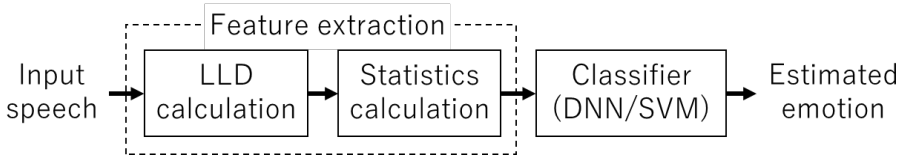

Figure 1 – Block diagram of emotion recognition system

Table 1 – Low-level descriptors (LLDs) and statistics

LLD (16× 2)	Statistics (12)
(△) ZCR	mean, standard deviation
(△) RMS Energy	kurtosis, skewness
(△) F0	min. and max: value, relative position
(△) HNR	difference between min. and max.
(△) MFCC (1-12)	linear regression: offset, slope, MSE

3. EMOTION SPEECH CORPUS

In this study, we used two emotion speech corpora: OGVC and Japanese Twitter-based emotional speech (JTES) for model training.

3.1 OGVC

OGVC comprises two types of speeches: spontaneous speech and acted speech. The former involves speech recording a conversation while playing an MMOPRG. In the latter, professional actors and actresses read aloud the transcripts of 17 dialogues extracted from the gameplay conversations. When they read them, the emotion type and intensity are specified. For a spontaneous emotion speech, the emotion type of each utterance is determined by the majority of three evaluators. The emotional intensity of each utterance is labeled in five levels by 18 evaluators. In our study, the average value by all evaluators is determined as the emotional intensity of the utterance.

3.2 JTES

In JTES, sentences containing emotional expression words were selected from tweets on Twitter by considering phonological and prosodic balance (11). Those sentences were read aloud to convey the intended emotion to a robot. The sentences were classified into four emotions (anger, joy, sadness, and neutral). The total number of sentences is 200, and the number of speakers is 100. Subsequently, the total number of utterances is 20,000.

4. RESURTS OF EMOTION RECOGNITION

4.1 Investigation on type of training data

In general, the emotional intensity of acted speech uttered by professional actors is strong. Meanwhile, the intensity of spontaneous emotional speech is low. The difference in emotional intensity affects speech signal variously. Hence, it is considered suitable to use acted emotional speech as training data for emotion recognition of acted emotional speech. The same applies for spontaneous emotional speech. In this section, we investigate the effect of the training data. Table 2 shows the conditions of the DNN used in these experiments, and Table 3 shows the test and training data. We conducted the following five spontaneous emotional speech recognition experiments where the type of training data is different:

ACT Acted speech data are used for model training. Because the types of training and test data are different, the recognition performance is expected to be low.

SPON The model is trained only by spontaneous emotional speech. Evaluation is performed by an 11-fold cross validation. The number of speakers for spontaneous speech is 11. In this experiment, utterances by 10 speakers are used for model training and those by another speaker are used for testing.

SPON + ACT To increase the amount of training data, acted speech data are added to the SPON above.

SPON + ACT1 The purpose is the same as the above, but the acted speech data to be added are limited to emotional intensity level 1 that is the weakest emotional expression.

SPON + ACT1-2 The acted speech data to be added are limited to intensity levels 1 and 2.

Table 2 – Experimental conditions for DNN

Structure of DNN		Pre-training		Fine-turning	
Input layer	384 nodes	# epoch	5	Initial learning rate	0.008
Hidden layer	3 layers	L2 regularization		# epoch	determined by early
	256 nodes	rate	0.002		stopping
Output layer	5 nodes	Batch size	100	Batch size	32

Table 3 – Test data and training data

Test data	
Spontaneous speech (2,438 utterances by 11 speakers)	
Training data	
ACT	Acted speech (2,656 utter. by 4 speakers)
SPON	Spontaneous speech uttered by 10 out of 11 speakers
SPON + ACT	SPON + acted speech (1,376 utter)
SPON + ACT1	SPON + acted speech (intensity 1. 344 utter.)
SPON + ACT1-2	SPON + acted speech (intensity 1 and 2. 688 utter.)

We conducted recognition experiments to classify five emotions (anger, joy, sadness, surprise, and neutral) using the SVM or DNN as a classifier. Figure 2 shows the recognition results. As expected, *ACT* demonstrates the worst result because the training data and test data are mismatched. The recognition performance improved in the matched condition (*SPON*), but simple data growth did not obtain good results (*SPON+ACT*). The best performance was obtained in the *SPON+ACT1-2* condition. The ACT1-2 dataset comprises utterances of low emotional intensity. This implies that the training data and test data must match in terms of emotional intensity. Regarding the classifier, the DNN performs better than the SVM; however, the difference between the two appears to be slight. Table 4 shows the results for each emotion. From the results of the SVM, *neutral* shows the best performance, while the recognition rate is only 13.1% for *anger*. Hence, the SVM is not suitable as a classifier in our experiments. Therefore, we used only the DNN in our subsequent experiments.

4.2 Investigation on amount of training data

In speech recognition, hundreds of hours or more of speech data are used for model training. Meanwhile, only approximately 20 min of speech data were used in the condition *SPON*. In this section, we describe the expansion of training data amount using JTES. JTES comprises 20,000 utterances uttered by 100 speakers. Using this corpus, the amount of training data can be expanded approximately eight times. JTES does not include utterances belonging to category "surprise." In the following experiments, four-class classification (anger, joy, sadness, and neutral) is conducted. As the amount of training data increases, the number of hyperparameters in the models is modified to the optimal number. The test data are the same as those in Sect. 4.1.

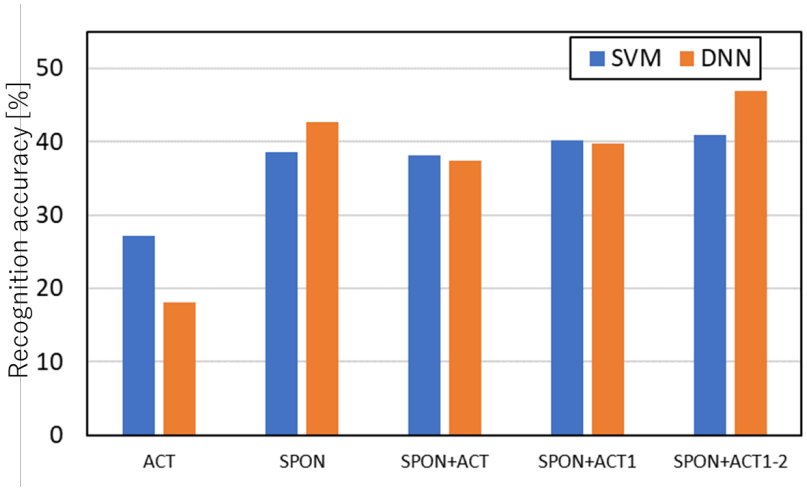

Figure 2 – Emotion recognition results with various conditions

Table 4 – Recognition rate for each emotion [%]

	anger	joy	sadness	surprise	neutral	average
SVM	13.1	48.7	30.0	52.9	59.9	40.9
DNN	44.3	38.3	43.2	50.4	58.0	46.9

SPON The model is trained only by spontaneous emotional speech in the OGVC. The number of hidden layers is four and the number of nodes for each layer is 256.
SPON + JTES To increase the amount of training data, speech data in JTES are added to the SPON above. The number of hidden layers is three and the number of nodes for each layer is 1024.

The recognition rate for SPON is 42.2%, while that for SPON+JTES is 41.3%. The recognition performance slightly decreased despite a significant increase in the data amount. In JTES, an utterance is recorded by reading the transcript aloud. Therefore, the emotional intensity may be stronger than that in SPON. The mismatch in intensity is considered to be caused by performance degradation. The emotional labels of the evaluation data are determined by the majority of three persons. In the SPON data, the percentage of data for which two or three evaluators agree on an emotion label is 77.7% and 22.3%, respectively. Table 5 shows the recognition rate for various labels. Comparing the two cases, SPON+JTES shows better results in the case of labels agreed by three. This suggests that the use of JTES data is effective when the utterance emotion is clear.

5. CONCLUSIONS

In this study, we examined the performance improvement in emotion recognition for spontaneous emotional speech. We used the OGVC as the emotional speech corpus for model training and evaluation. The OGVC uses a game player's speech in which they play an MMOPRG with voice chat. For emotion recognition, we used the speech features standardized in the *Interspeech 2009 Emotion Challenge*, and the number of dimensions was 384. A DNN was used as an acoustic model that exhibited a feed-forward type of architecture. In constructing the acoustic model, we investigated the type and amount of training data. In addition, we compared the DNN and SVM as classifiers. The results indicated that the DNN performed better in our task. Regarding the training data, the recognition performance improved in the matched condition where spontaneous speech data were used for model training. However, the amount of spontaneous data was only approximately 20 min. In general, hundreds of hours or more of speech data were used for model training in speech recognition. Performance improvement was limited with the amount of training data. To increase the amount of training data, we attempted to add different types of emotional data to the spontaneous emotional data. We found that merely increasing the amount of training data did not contribute to performance improvement. Generally, the intensity of spontaneous emotional speech is low. We found that the selective use of emotional data with low intensity was effective.

In this study, we used emotional intensity labels for data selection. However, it was difficult to prepare a large amount of intensity-labeled speech data. We plan to develop an automatic intensity determination for emotional speech as a feature task.

ACKNOWLEDGEMENTS

This study was supported in part by a Grant-in-Aid for Scientific Research (KAKENHI 19K12014) form the Japan Society for the Promotion of Science. We thank Dr. Nose, Tohoku University for providing the emotional speech corpus JTES.

Table 5 – Recognition rate for various labels [%]

Label	Majority decision	Agreed by two	Agreed by three
SPON	42.2	42.0	44.2
SPON+JTES	41.3	40.3	49.7

REFERENCES

1. Hassan A, Damper R. Multi-class and hierarchical SVMs for emotion recognition, Proc. INTERSPEECH 2010; 26-30 September 2010; Makuhari, Japan 2010. pp. 2354-2357.
2. Stuhlsatz A, Meyer C, Eyben F, Zielke T, Meier G, Schuller B. Deep neural networks for automatic emotion recognition: Raising the benchmarks, Proc. ICASSP 2011; 22-27 May 2011; Prague, Czech Republic 2011. pp. 5688-5691.
3. Amer M.R, Siddiquie B, Richey C, Divakaran A. Emotion detection in speech using deep networks, Proc. ICASSP 2014, 4-9 May 2014; Florence, Italy 2014. pp. 3752-3756.
4. Han K, Yu D, Tashev I. Speech emotion recognition using deep neural network and extreme learning machine, Porc. INTERSPEECH 2014; 14-18 September 2014; Singapore 2014. pp. 223-227.
5. Lee J, Tashev I. High-level feature representation using recurrent neural network for speech emotion recognition, 6-10 September 2015; Dresden, Germany 2015. pp. 1537-1540.
6. Devillers L, Vidrascu L. Real-life emotion detection with lexical and paralinguistic cues on human-human call center dialogs, Proc. INTERSPEECH 2006; 17-21 September 2006; Pittsburgh, USA 2006. pp. 801-804.
7. Arimoto Y, Kawatsu H, Ohno S, Iida H. Emotion recognition in spontaneous emotional speech for anonymity-protected voice chat systems, Proc. INTERSPEECH 2008; 22-26 September 2008; Brisbane, Australia 2008. pp. 322-325.
8. Mori H, Satake T, Nakamura M, Kasuya H. Constructing a spoken dialogue corpus for studying paralinguistic information in expressive conversation and analyzing its statistical / acoustic characteristics, Speech Commun. 2011; 53: 36-50.
9. Busso C, et al. IEMOCAP: interactive emotional dyadic motion capture database, Language Resources and Evaluation. 2008; 42(4): 335-359.
10. Arimoto Y, Kawatsu H, Ohno S, Iida H. Naturalistic emotional speech collection paradigm with online game and its psychological and acoustical assessment, Acoust. Sci. and Tech. 2012; 33(6): 359-369.
11. Takeishi E, Nose T, Chiba Y, Ito A. Construction and analysis of phonetically and prosodically balanced emotional speech database, Proc. O-COCOSDA 2016, 26-28 October 2016; Bali, Indonesia 2016. pp. 16-21.
12. Schuller B, Steidl S, Batliner A. The INTERSPEECH 2009 emotion challenge, Proc. INTERSPEECH 2009; 6-10 September 2009; Brighton, UK 2009. pp. 312-315.