Development of Interactive Robot
-Emotion Estimation System Using Speech by 1dCNN-

Yugo Kawachi
Department of Mechanical Information Science and Technology, Kyushu Institute of Technology
680-4, Kawazu, Iizuka-City, Fukuoka, 820-8502, Japan

Eiji Hayashi
Department of Mechanical Information Science and Technology, Kyushu Institute of Technology
680-4, Kawazu, Iizuka-City, Fukuoka, 820-8502, Japan

E-mail: kawachi.yugo846@mail.kyuutech.jp, haya@mse.kyutech.ac.jp
http://www.kyutech.ac.jp/

Abstract

In order for robots to interact smoothly with people, they need to recognize human emotions and express its own emotions through both verbal and non-verbal communication. In non-verbal communication, we naturally do things such as estimating emotions from the tone of the other person's voice while talking. In this study, we developed a system to estimate emotions from features of speech rather than the speaker's words by comparing two different data sets, and compared what features each data set has and how they differ for each subject.

Keywords: personal robot, emotion estimation, nonverbal communication, 1dCNN

1. Introduction

With the expansion of the robotics industry market, the development of service robots is becoming more and more popular. These robots are intended to be used in the home, medical care, welfare, and other places where people communicate with each other, and it is necessary for them to behave and talk in a friendly manner. In this research, we are developing an interactive robot that pursues human-like movements by focusing on non-verbal interactions (cooperative behavior) such as facial expressions and body language.
In conversations, we can infer the other person's emotions from the inflection of their speech. We thought it was necessary to have a function to estimate emotions from speech. Therefore, in this study, we developed an emotion estimation system for speech using a machine learning method called 1 dimensional convolutional neural network(1dCNN).

2. Emotion Estimation System for Speech Using 1dCNN

In this system, we used a model called emotion-classification-Ravdess[1]. In addition, user status was defined 'Positive' and 'Negative' in line with previous studies.

2.1. *Structure of 1dCNN*

The structure of the 1dCNN for emotion estimation is shown in Fig. 1. In Fig. 1, A is Convolution layer and B is Pooling layer. The parameters of the Convolution and Pooling layers are shown in Table. 1, and the ReLU and softmax functions are used as activation functions.

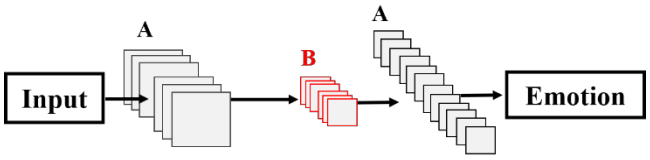

Fig. 1 Structure of network

	1dCNN	MAX_pooling	1dCNN
Size of filter	5	8	5
stride	1	1	1
Number of filters	128	128	128

Table. 1 parameters of each layers

2.2. *Feature Values*

The Mel Frequency Spectrum and Mel Frequency Cepstrum Coefficient (MFCC) were extracted as features from speech, and the features were used as input data for training. The MFCC is a discrete cosine transformed coefficient of the Mel Frequency Spectrum. The discrete cosine transform is said to improve the performance of the feature.

2.3. *Leaning data*

For training, we used the Ravdess[2] dataset and a newly created dataset.

Ravdess was recorded two types of sentences, "Kids are talking by the door" and "Dogs are sitting by the door," read by 24 professional actors (12 men and 12 women) in eight different emotions. The total number of data was 1440. The eight emotions were Classified “neutral, calm, happiness, sad, anger, fearful, disgust, surprised”. In this research, neutral and calm were removed, joy and surprise classified as positive, and sadness, anger, fear, and disgust classified as negative. The new dataset was created using the author's voice, and as in Ravdess. It was recorded by emotion and classified into positive and negative. The new dataset contains English greetings of two words or less, not sentences. The total number of data is 120, 60 for positive and 60 for negative.

2.4. *Accuracy Evaluation*

Table. 2 shows the classification results by each model of the results of using 1dCNN with MFCC and Mel frequency spectrum as input, and it used author's voice.

Dataset	Ravdess		Newly Created Dataset	
Feature Values	MFCC	Spectrum	MFCC	Spectrum
Positive	0%	0%	92%	89%
Negative	67%	67%	93%	90%

Table. 2 Classification Accuracies for each feature

First of all, focusing on the classification accuracy of each dataset, the classifier using Ravdess was classified as negative in both MFCC and Mel frequency spectrum. On the other hand, the evaluation using the newly created data set showed that the classifier was able to classify more unbiasedly than the classifier using Ravdess. The reason is conceivable that Ravdess dataset is a sentence. When words are given as input, it is necessary to focus on the inflection of the words for classification. However, in the case of sentences, the recognition rate drops due to the influence of syllables in between sentences.

Next, classification accuracy for each feature value in the newly created dataset, we can see that MFCC has a higher accuracy of about 3%. However, since there was only one difference in the number of data that failed to be classified, it was concluded that there was no difference in the classification accuracy between the two features.

Based on the above results, the classifier trained on the newly created data set using MFCC and Mel frequency spectrum features respectively was adopted as the emotion estimation system.

3. Verification

Test data containing the voices of four people were input to the emotion estimation system for classification. Fig. 2 shows the results of classification using MFCC and Mel Frequency Spectrum. A and C are the results of the system's classification of the positive emotions, and B and D are the results of the system's classification of the negative emotions. Table. 3 shows the classification accuracies of each class obtained from the classification results of this experiment.

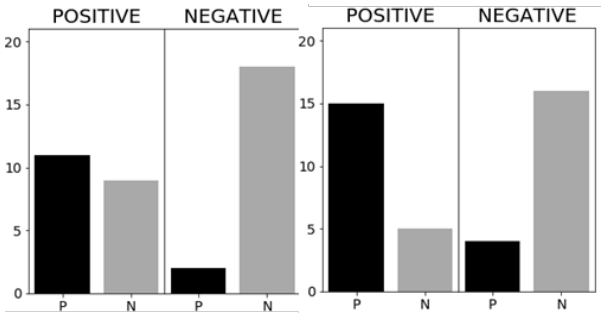

Fig. 2 Result of classification for each feature

Table. 3 Classification accuracy of emotion for each feature

	MFCC	spectram
Positive	**66.7%**	**76.9%**
Negative	**76.6%**	**78.0%**

From the results of this study, it was found that the classification accuracy of the model using Mel Frequency Spectrum was on average 5.9% higher than the model using MFCC for feature extraction. In addition, the model using MFCC showed a bias towards the negative, however it was not the case when the Mel Frequency Spectrum was used. The shape of each feature was output as a graph for comparison.

Fig. 3 shows the MFCC features of subject A's speech, and Fig. 4 shows the same speech as in Fig. 3 with the Mel frequency spectrum. From these figures, It can be seen that in MFCC, there is not much difference in shape between positive and negative and it is difficult to judge, whereas in Mel frequency spectrum, there was a clear difference in shape.

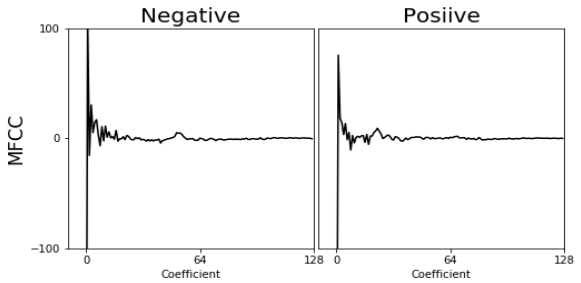

Fig. 3 Shape of MFCC

Table. 4 shows the emotion classification accuracy for each subject. First, focusing on subject C, the accuracy is lower than the others in the classification using the Mel frequency spectrum. This is thought to be because the distance between the subject and the microphone is close and the volume of the sound is larger. This suggests that in the classification using the Mel frequency spectrum, the factor of the loudness of the sound is related to the classification.

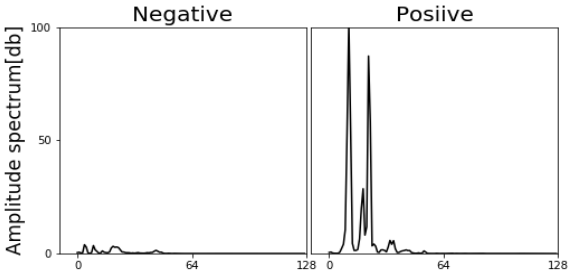

Fig. 4 Shape of Mel Frequency Spectrum

Feature Value	emotion	A	B	C	D
MFCC	Positive	75.0%	75.0%	88.9%	50.0%
	Negative	83.3%	83.3%	90.9%	66.7%
Mel Frequency Spectrum	Positive	57.1%	88.9%	71.4%	88.9%
	Negative	76.9%	90.9%	33.3%	90.9%

Table. 4 Accuracy of classification for each subject

4. Conclusion

In this study, we created and evaluated a classification system for emotions using 1dCNN. As a result, we were able to develop a system to classify users' emotions with an accuracy of 76.9%. However, we are able to identify emotions even when we use speakers, which means that a system in which the loudness and height of the voice affect the classification is not appropriate. This means that a system in which the loudness and height of the voice affect the classification is not appropriate. As a future prospect, it is necessary to search for an emotion estimation system that does not depend on the volume and height of the voice, and also to construct an emotion estimation system that combines facial expression and body expression recognition.

5. References

1. **Marco Giuseppe de Pinto,**https://github.com/marcogdepinto /Emotion-Classification-Ravdess(2019)

2. Livingstone, Steven R., Katlyn Peck, and Frank A. Russo. "Ravdess: The ryerson audio-visual database of emotional speech and song." *Annual meeting of the canadian society for brain, behaviour and cognitive science*. 2012.