Emotional Voice Detection using MFCC for Bunraku Chant

Zhou WENTING
Tsukuba University, Cave Lab

Abstract

The popularity of artificial intelligence has brought about the vigorous development of speech recognition. Speech emotion recognition is also a technology that has attracted much attention in recent years. This research considers the application of speech emotion recognition to Bunraku, a traditional Japanese culture. Since training data has a great influence on speech emotion recognition, and Bunraku of traditional culture does not have such an emotion data set, we consider using the general speech emotion corpus JTES [1] to detect Bunraku's emotions. Discuss whether this method is feasible.We build four classifiers with MFCC as the main feature, two SVMs using the SMO algorithm, and two DNN models. Then use Bunraku data and JTES corpus to train the classifiers, and the precision of SVM is 49.107% and 50%, the precision of DNN is 50.893% and 52.778%. Finally, we compared the features of the two datasets and verified the feasibility of detecting Bunraku emotions with the JTES corpus.

Keywords: bunraku chant, emotional voice detection, SVM, DNN, MFCC

Introduction

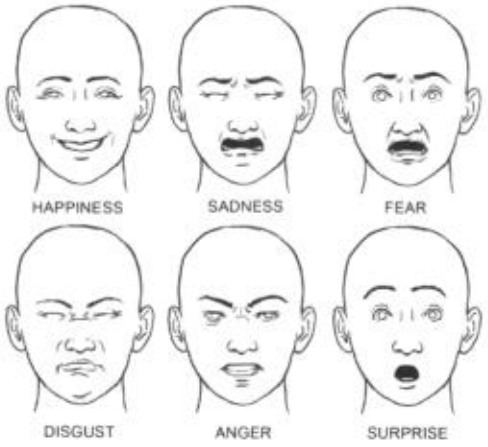
Bunraku is one of Japan's traditional theater arts. It is performed by a narrator, a shamisen player and puppeteers, and expresses colorful stories through puppet. In addition to expressing the emotions of puppets through body movements, the voices of narrator and shamisen are also important expression media.

With the rapid development of artificial intelligence, the interaction between computers and humans has become more and more. Speech recognition technology is an important part of human computer interaction. In addition to semantic recognition, emotion recognition is also particularly important [2]. Therefore, speech emotion recognition has received extensive attention in recent years. If computers can accurately recognize human emotions, they can better understand the user's intentions and effectively improve the quality of human computer interaction. In addition to the emotion recognition of speech, the emotion recognition technology of music is also gradually developed. M. A. Casey et al. [3] proposed a method of digital music emotion recognition, which can use commonly used acoustic characteristic parameters for emotion recognition. Nowadays, a large number of songs are released through the Internet and stored in large digital music databases, and the most commonly used words for retrieving and describing music are emotional words [4].

This research considers the application of speech emotion recognition in Bunraku chant. Emotion recognition of traditional music is more difficult than that of normal music, because the training data set has a great influence on speech emotion recognition. Many researchers have produced emotional

corpora of modern music, while traditional music rarely contains data with emotional labels. In speech emotion recognition, language also has a great influence on the recognition effect [5], people in different countries and languages have different features of expressing emotions. However, the emotional corpus of Japanese songs is not much. so in this research we consider the Japanese speech corpus. Whether the emotion of Bunraku chant can be accurately identified through the normal speech corpus is the purpose of this research.

Figure 1. The six basic human emotions (Image source: "The Minds Journal"[2])



Bunraku

Bunraku is performed by a narrator, a shamisen player, and three puppeteers, called "Sangyo", which means three professions. This thesis will mainly study the emotions expressed by Narrator's voice, the narrator expresses Bunraku's emotions through voice changes and narrative skills, portraying the emotions of each character with voice.

The fragment of the experiment in this thesis is the Sugisakaya. Sugisakaya is the first scene of the 4th act of "Imoseyama Onna Teikin" [1]. The Bunraku data used in this research comes from the Osaka University of the Arts in March 2017. This study uses the video of Sugisakaya taken at that time as the original data, and extracts audio from the video. The audio is divided into three segments with durations of 2 minutes 32 seconds, 5 minutes 26 seconds, and 3 minutes 13 seconds.

JTES

Since this research is based on Bunraku and Bunraku chant is Japanese, the Japanese voice emotion database is selected. This study uses the emotional sound corpus Japanese Twitter-based emotional speech (JTES) [1]. The data set includes 50 sentences for each of the four emotions of joy, anger, sad, and natural, and the colloquial sentences expressing emotions in Twitter, taking prosodic balanced into consideration. There are 20,000 speech fragments of 100 people with 50 people each for male and

female.

This data set uses a discrete sentiment model. The corpus is a simulated emotion corpus, that is, the speaker is allowed to imitate different emotions to read the specified content. This method has strong operability and easier data acquisition. And the obtained corpus meets the emotional requirements and is highly distinguishable.

Figure 2. A fragment of Sugisakaya

お三輪　　イエイエイエ、私がまだ用がある。いなすことはなりませぬ
姫　　イヽヤ、こゝには置きはせぬ。邪魔せずと、そこ通しゃ
と、手を引っ立てゝ立ち出づれば、
お三輪　　イヤ放さじ
とお三輪もまた、あなたへ引けばこなたへ引く、訳も渚にたはれる雁、つばさ振袖振り分け姿、恋を争ふその折から、

Mel-Frequency Cepstral Coefficients (MFCC)

MFCC was proposed by Davis and Mermelstein in 1980 [6], and is a feature widely used in automatic speech and speaker recognition, and has also been used in music classification in recent years.

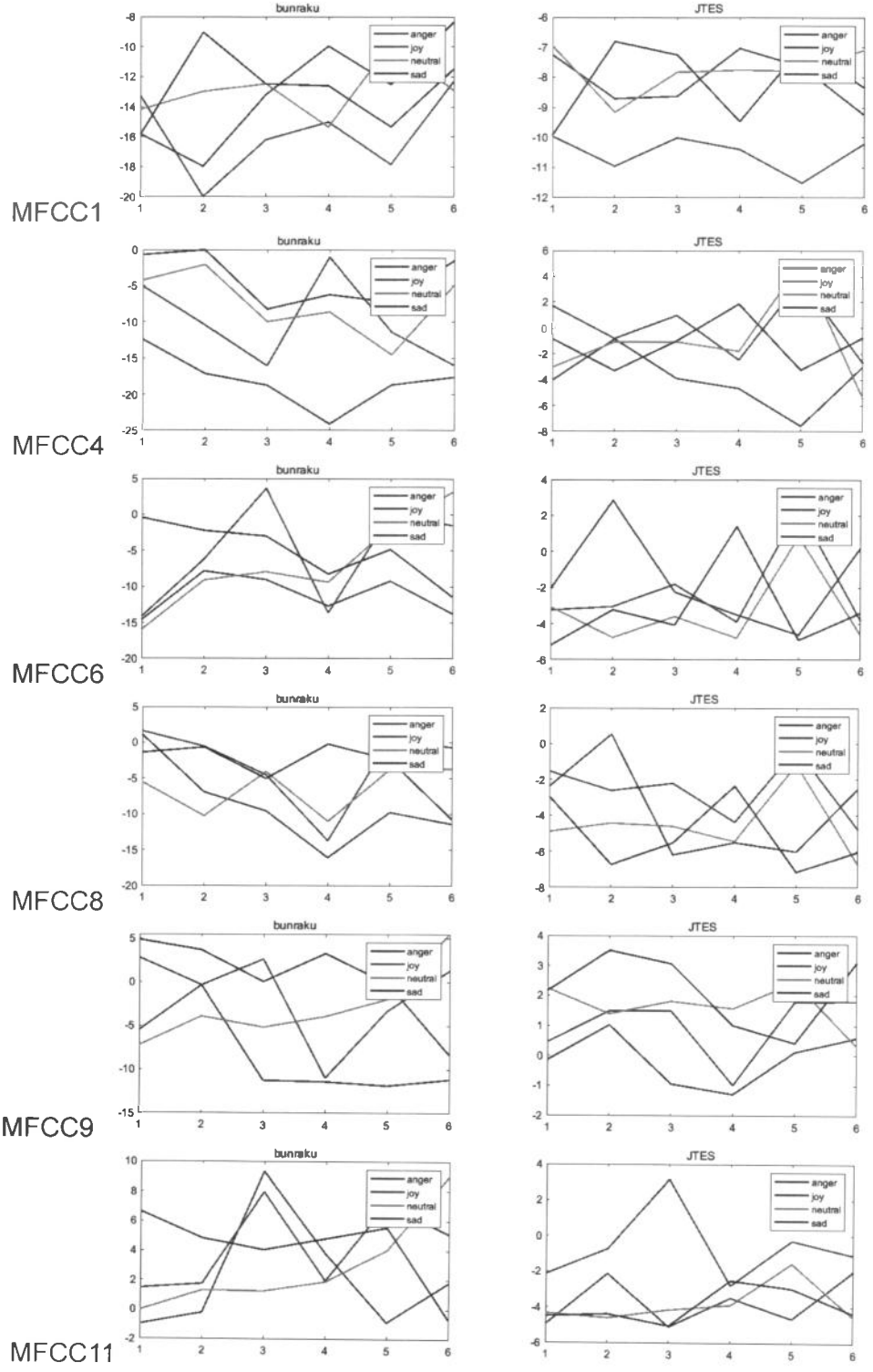
We compare the MFCC of different emotions in Bunraku and JTES. Bunraku data select 4 votes data. Equal amounts of data were randomly selected for comparison in JTES. Bunraku's MFCC (1-12) are generally smaller than JTES. The specific comparison is as follows. The horizontal axis is the different data, and the vertical axis is the value of the feature.

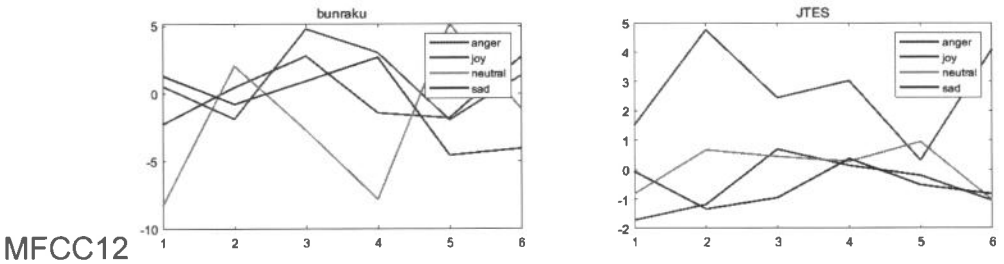
MFCC1, the value of joy in both datasets is the smallest. MFCC2, the distribution of each emotion is different. MFCC3, the distribution of each emotion is different. MFCC4, the value of anger in both datasets is the smallest, the value of sad is similar. MFCC5, the distribution of each emotion is different. MFCC6, the value of sad is similar. MFCC7, the distribution of each emotion is different. MFCC8, the value of joy, neutral, sad are similar. MFCC9, the value of anger in both datasets is the smallest, the value of joy and sad are similar. MFCC10, the distribution of each emotion is different, but the value ranges are similar. MFCC11, the value of joy is similar. MFCC12, the range of values for each emotion is similar, and the distribution of anger is the highest among the four emotions. Some MFCC are shown in Figure 3.

Speech emotion recognition

Voice is a very important way for humans to express emotions. Human voices contain a lot of information. In addition to semantic information expressed through language, the speaker's identity information, etc., there is also emotional information. When the same person speaks the same sentence with different emotions, different semantics can be conveyed. Humans can capture the emotional changes of the speaker through voice, such as changes in intonation, volume, or special modal particles. Speech emotion recognition is the simulation of the above-mentioned human emotion perception and understanding of speech through a computer. The computer extracts the sound features in the speech signal, learns the correspondence between these features and emotions, and makes predictions.

Figure 3. MFCC of Bunraku and JTES





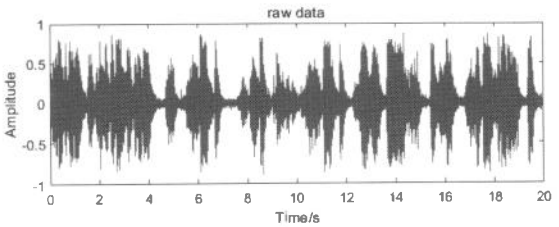
Speech emotion recognition is used in many aspects, such as driving safety. Nowadays, voice interaction devices in vehicles are gradually becoming popular, and voice emotion recognition is also beginning to be used to monitor emotions and improve driving safety. When the driver is in an emotionally unstable state, it will increase the risk of driving. If the system can provide reminders or care based on the driver's emotions, it can improve the driving experience [7]. Voice emotion recognition is also used in polygraph detection. The voice emotion recognition system can assist the polygraph to detect the tension or anxiety of the speaker according to the rhythm and pitch of the voice. More applications include patient emotion detection [8,9] and automatic translation [10].

There are two main types of emotion models commonly used today. One is the discrete emotional model, which is the model used in this research. Another type of emotion classification is the dimensional emotion model, which can be constructed in a two-dimensional or higher-dimensional space to describe continuous emotions. This emotion model is more detailed in the classification of emotions, and each of the above emotions can be represented by dimensions.

Data preprocessing

The bunraku data used in this research is extracted from the video. The audio contains noise. The data must be denoised first. Then the Bunraku data in this experiment was extracted from the video, the voice of the narrator and the shamisen appear together, without a separate sound source. The training data set JTES used in this study is pure human voice, so we need to separate the human voice from the shamisen sound.

Figure 4. Waveform of Sugisakaya first segment 40s to 60s



Then divide the processed bunraku data into phrases according to the lyrics. Since the audio is not a continuous complete segment, using "Imoseyama Onna Teikin: Commentary. Sugisakaya" [1] a reference, I sorted out the lyrics that appear in the audio. The Sugisakaya audio has a total of 11 minutes 11 seconds, divided into 219 phrases. I invited 4 native Japanese speakers to judge the emotion

of Bunraku data. The reference emotion is determined according to the number of votes, and the emotion with the largest number of votes is selected as the final emotion. The number of votes is shown in Table.

Figure 5. Waveform of denoised data and narrator voice and Shanmisen sound

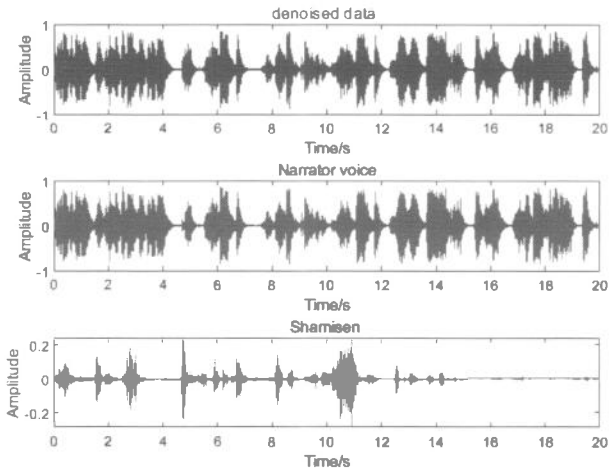


Table 1. Voting result

Number of votes	Count
2	20+15
3	72
4	112
Total	219

Among the data with 2 votes, 15 of them are 2 to 2, and the emotions cannot be determined, each of these 15 data has two emotions, we will not discuss these 15 data later.

Use OpenSMILE [11], where feature set INTERSPEECH (2009) [12] is used to extract the features of JTES and Bunraku data. In addition to the 384 features of INTERSPEECH (2009), there are also the name feature and the class feature. A total of 386 features.

Table 2. Extracted features

Name	string
384 features of INTERSPEECH(2009)	numeric
Class	{anger, joy, neutral, sad}

Classifier

We use JTES and Bunraku (4 votes) as training data to test the classification of emotions. SVM and DNN are chosen as classification models. Sequential minimal optimization (SMO) is an efficient optimization method of SVM dual problem, mainly used to solve the optimization problem of SVM objective function. SMO is proposed by John C. Platt [15]. SMO decomposes a large optimization

problem into multiple small optimization problems to solve, and the sequential solution results of these small optimization problems are consistent with the overall solution result, but the calculation time is greatly shortened. Build SVM and DNN models for the two datasets respectively.

Assuming that y_i is the predicted value and x_i is the true value, the evaluation standards are,

$$\text{Precision} = \text{Correctly Classified Instances/Total Number of Instances}$$

$$\text{Relative absolute error(RAE)} = \frac{\sum_{i=1}^{N}|x_i - y_i|}{\sum_{i=1}^{N}|x_i - \bar{x}|}, \qquad \bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

$$\text{Root relative squared error(RRSE)} = \frac{\sum_{i=1}^{N}(x_i - y_i)^2}{\sum_{i=1}^{N}(x_i - \bar{x})^2}, \qquad \bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Table 3. SVM (10-fold Cross Validation)

training set	Precision	RAE	RRSE
JTES	75.75%	73.954%	81.355%
Bunraku	50.893%	101.844%	102.321%

For JTES as training data, our DNN consists of an input layer, 4 hidden layers and an output layer. Using the Adam [14] optimization, it is a first-order gradient-based stochastic objective function optimization algorithm. To avoid overfitting, we added dropout, ignoring 20% of the nodes in each layer. There are 4 neurons in the output layer, corresponding to the four categories, using the softmax function. Where $x_i(i = 1, \ldots, n)$ is neurons, ω is the weight, b is the bias, C is the number of output nodes, that is, the number of categories.

$$z = \sum_{i=0}^{n} \omega_i x_i + b$$

$$softmax(z)_i = \frac{\exp(z_i)}{\sum_{c=1}^{C}\exp(z_c)}, \qquad i = 1, \ldots, C,$$

Because the amount of Bunraku data is small, for Bunraku data as training data, we design a simpler DNN.The structure of DNNs are shown in Table 4.

Table 4. The structure of DNN(left-JTES, right- Bunraku)

Model: "sequential"

Layer (type)	Output Shape	Param #
dense (Dense)	(None, 512)	197120
dropout (Dropout)	(None, 512)	0
dense_1 (Dense)	(None, 1024)	525312
dropout_1 (Dropout)	(None, 1024)	0
dense_2 (Dense)	(None, 1024)	1049600
dropout_2 (Dropout)	(None, 1024)	0
dense_3 (Dense)	(None, 1024)	1049600
dropout_3 (Dropout)	(None, 1024)	0
dense_4 (Dense)	(None, 4)	4100

Total params: 2,825,732
Trainable params: 2,825,732
Non-trainable params: 0

Model: "sequential"

Layer (type)	Output Shape	Param #
dense (Dense)	(None, 512)	197120
dropout (Dropout)	(None, 512)	0
dense_1 (Dense)	(None, 1024)	525312
dropout_1 (Dropout)	(None, 1024)	0
dense_2 (Dense)	(None, 1024)	1049600
dropout_2 (Dropout)	(None, 1024)	0
dense_3 (Dense)	(None, 4)	4100

Total params: 1,776,132
Trainable params: 1,776,132
Non-trainable params: 0

Result

We use all the JTES data and 4 votes Bunraku data as training sets to train the four models separately. For the model using JTES as the training set, all Bunraku data are tested and counted separately. For models trained with 4 votes Bunraku data, test Bunraku data with 2 votes and 3 votes. The test results are shown in Table 5.

Table 5. Classification result

Classifier	Number of votes	Count	Correct count	Precision
(JTES-SVM)	2	20	7	35%
	3	72	32	44.444%
	4	112	55	49.107%
	Total	204	94	46.078%
(JTES-DNN)	2	20	10	50%
	3	72	33	45.833%
	4	112	57	50.893%
	Total	204	100	49.200%
(Bunraku-SVM)	2	20	6	30%
	3	72	36	50%
	Total	92	42	45%
(Bunraku-DNN)	2	20	10	50%
	3	72	38	52.278%
	Total	92	48	52.174%

Since the training data used by the second classifier is four votes' data, we use three votes' data to compare the test data.

Table 6. Classification effect comparison

classifier	training data	test data	Precision
1-SVM	JTES	three votes	44.444%
2-DNN	JTES	three votes	45.833%
3-SVM	four votes	three votes	50%
4-DNN	four votes	three votes	52.278%

JTES has more data, 20,000, while four votes' training data is only 112. However, Bunraku still performs better than JTES as a training set. If there is enough Bunraku emotion data, the classification effect is better, but for Bunraku, which does not have a large emotional corpus, the Bunraku emotion detected by JTES as a training set still has reference value. In general, the model of JTES as a training set is not much different from Bunraku. It can be said that it is feasible to use JTES to detect Bunraku's emotions.

Discussion

In this research, we discussed Bunraku's emotion detection method. We propose the feasibility of detecting bunraku emotions with daily emotion corpus. Because the traditional culture of Bunraku does not have existing emotional labels, we try to use the JTES corpus to predict Bunraku's emotions. SVM and DNN were trained using speech emotion corpus, and its effect was discussed. And apply it to Bunraku data to predict Bunraku emotions. Then we trained SVM and DNN using Bunraku data to compare the classification effects of the two SVMs and two DNNs.

We preprocessed the data, including noise reduction and separation of the narrator's voice. The bunraku data was divided into 219 phrases according to the lyrics, and four native Japanese speakers were invited to judge the bunraku data. Finally got Bunraku's emotional label by voting. In the experiment, we use a SVM trained on the JTES corpus to predict the emotion of Bunraku data. The precision of three votes is 44.444%, the precision of four votes is 49.107%. Use DNN, the precision of three votes is 45.833%, the precision of four votes is 50.893%. Then we use the Bunraku data of the four votes as the training data. The precision of SVM for three votes is 50%. The precision of DNN for three votes is 52.278%.

We analyze the consistency of MFCC, the main feature used to train the model, on Bunraku and JTES. And the effect of JTES and Bunraku as training set models are similar, we verify the feasibility of JTES to detect Bunraku's emotion.

Appendix

1. https://dl.ndl.go.jp/info:ndljp/pid/1024775/4?tocOpened=1
2. https://themindsjournal.com/basic-emotions-and-how-they-affect-us/

References

[1] 武石笑歌 他“感情音声データベース構築に向けた音韻・韻律バランス感情音声の収録と分析”日本音響通信学会講演論文集,1-R-47 (2016-3).

[2] Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., & Taylor, J. G. (2001). Emotion recognition in human-computer interaction. IEEE Signal processing magazine, 18(1), 32-80.

[3] Casey, M. A., Veltkamp, R., Goto, M., Leman, M., Rhodes, C., & Slaney, M. (2008). Content-based music information retrieval: Current directions and future challenges. Proceedings of the IEEE, 96 (4), 668-696.

[4] Woo, W., Park, J. I., & Iwadate, Y. (2000, December). Emotion analysis from dance performance using time-delay neural networks. In Proceedings of the Fifth Joint Conference on Information Sciences, JCIS 2000 (pp. 374-377).

[5] Latif, S., Qayyum, A., Usman, M., & Qadir, J. (2018, December). Cross lingual speech emotion recognition: Urdu vs. western languages. In 2018 International Conference on Frontiers of Information Technology (FIT) (pp. 88-93). IEEE.

[6] Davis, S., & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE transactions on acoustics, speech, and signal processing, 28(4), 357-366.

[7] Pravena, D., & Govind, D. (2017). Significance of incorporating excitation source parameters for improved emotion recognition from speech and electroglottographic signals. International Journal of Speech Technology, 20(4), 787-797.

[8] Paris, M., Mahajan, Y., Kim, J., & Meade, T. (2018). Emotional speech processing deficits in bipolar disorder: The role of mismatch negativity and P3a. Journal of affective disorders, 234, 261-269.

[9] Schelinski, S., & von Kriegstein, K. (2019). The relation between vocal pitch and vocal emotion recognition abilities in people with autism spectrum disorder and typical development. Journal of autism and developmental disorders, 49(1), 68-82.

[10] Swain, M., Routray, A., & Kabisatpathy, P. (2018). Databases, features and classifiers for speech emotion recognition: a review. International Journal of Speech Technology, 21(1), 93-120.

[11] Eyben, F., Wöllmer, M., & Schuller, B. (2010, October). Opensmile: the munich versatile and fast open-source audio feature extractor. In Proceedings of the 18th ACM international conference on Multimedia (pp. 1459-1462).

[12] Schuller, B., Steidl, S., & Batliner, A. (2009). The interspeech 2009 emotion challenge.

[13] Platt, J. (1998). Sequential minimal optimization: A fast algorithm for training support vector machines.

[14] Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

MFCC を用いた文楽語りの感情音声検出

周 文婷
（筑波大学大学院）

【要旨】

人工知能（AI）の普及にともなって、AI を用いた音声認識も急速に発展している。その中でも音声感情認識が近年注目されている。本研究では、日本の伝統文化である人形浄瑠璃文楽の語りに音声感情認識の適用することを検討する。音声感情認識には学習データが大きく影響するが、文楽にはそのような感情データがないため、一般音声感情コーパス JTES を用いて文楽の語りから感情表現の検出を試みた。MFCC を主な特徴とする 4 つの分類器、SMO アルゴリズムを用いた 2 つの SVM、および 2 つの DNN モデルを構築する。そして文楽の語りと JTES コーパスを用いて分類の機械学習を行ったところ、SVM の精度は 49.107%と 50%、DNN の精度は 50.893%と 52.778%であった。最後に、2 つのデータの特徴を比較し、JTES コーパスを用いた文楽語りの感情表現検出が可能であることを検証した。

キーワード：文楽語り、感情音声検出、SVM、DNN、MFCC