Analysis and Annotation of Emotional Traits on Audio Conversations in Real-time

Analysis and Annotation of Emotional Traits on Audio Conversations in Real-time

韓　迪
Di Han
Graduate School of Computer and Information Sciences
E-mail: di.han.6u@stu.hosei.ac.jp

Abstract

It is a challenging task for computers to recognize humans' emotions through their conversations. Therefore, this research is aimed at analyzing conversation audio data, then labeling humans' emotions, finally annotating and visualizing the identified emotional traits of audio conversions in real-time. In order to make computer to process speech emotion features, the raw audio is converted from time domain to frequency domain and extract speech emotion by Mel-Frequency Cepstral Coefficients. In terms of speech emotion recognition, deep neural network and extreme learning machine are used to predict emotion traits. Each emotional trait is captured by speech recognition precision. There are four emotional traits which include sadness, happiness, neutral, anger in the dataset. The total precision value of four emotional traits is normalized into 1. In this study, the normalized precision is used as emotional trait relative intensity in which each emotional trait is labeled and displayed along with conversion. For better visualization, a Graphical User Interface is made to display the waveform graph, spectrogram graph and speech emotion prediction graph of a given speech audio. Meanwhile, the effect of voice activity detection algorithm is analyzed in this study. The timestamps for emotion annotation can be obtained by the result of voice activity detection.

Keywords—Emotional traits analysis; Normalized precision; Emotional traits annotation

1. Introduction

With the advancement of technology, our lifestyles are changing gradually. Human beings' usage of computers and mobile phones has increased significantly. For better user experience, smart products such as computers and mobile phones should have more ability to interact with humans. Nowadays, the main communication methods with computers and mobile phones are focusing on clicking the mouse, typing on the keyboard and touching the screen. However, speech interaction can also be used as an important communication tool. Generally speaking, some researchers pay more attention to understand the meaning of words in speech, so that they can predict the emotion of the corresponding speech through keywords. While, extracting the acoustic features of speech to directly predict the emotion of speech is a popular method right now.

Until now, other researchers have contributed to implementing the process of speech emotion recognition, the model of speech emotion recognition is continuously optimized so that the better recognition accuracy is obtained. However, the task of emotion annotation is not given enough attention. In fact, it is very necessary for people to find out the emotion parts in audio conversation. The task to annotate emotion in conversation can be helpful in several ways. For example, finding emotions during the conversation with a customer can improve the quality of customer service. Additionally identifying emotions might also be useful in editing or summarizing videos as the emotional part is usually corresponding to the important part of the video. Our work focuses on solving this problem by implementing a system to annotate and analyze emotional traits, in particular, our system could identify 4 emotions: happy, sad, anger and neutral. We hope our system can benefit users who are interested in analyzing emotions in their audios or videos.

Regarding the practical application case of speech emotion recognition, customer service telephone call emotion recognition system developed by NTT Inc. in Japan is a representative product [1]. To recognize the intent and gain a better understand of users' feelings, NTT's speech emotion recognition system not only recognizes more intense angry emotions by identifying the volume and pitch of the user's speech but also recognizes calmer angry emotions through the rhythm of the speech and the choice of words. Use these two recognition methods to identify whether the user has angry emotions, especially the unique emotional expression of Japanese users. If dissatisfaction with the automatic telephone voice service is identified, those data can be collected and the system is improved to provide better services. This is one of the purpose of speech emotion recognition. From this case, it shows that this research is valuable.

Supervisor: Prof. Runhe Huang

This research is focused on understanding humans' emotional traits and labeling them during two people's conversations. There are four types of emotional traits taken into consideration: Angry, Happiness, Sadness, and Neutral. The inputs are two people's conversion dataset and the outputs are the conversions with labeled emotional traits along the time axis. For better visualization, we make a conversation replay user interface in which both the original conversation and its labeled emotional traits are displayed. By using this system, the emotion changes in people's conversations can be visualized, and the parts with emotion in the conversation can be found quickly without listening to audio files, which helps people determine the emotions in the audio files, even find the main part of a conversation data directly. In addition, the VAD (Voice activity detection) algorithm is also used in this system, which can recognize the part that people are speaking in an audio file. Thus, the emotions of audio files without timestamps also can be annotated.

2. Overview of Emotional Traits Annotation

The overview of this system is given in Fig.1, it includes MFCC (Mel Frequency Cepstrum Coefficients) for converting speech from the time domain into the frequency domain with extracted low-level features, DNN (Deep Neural Network) takes the features from MFCC as inputs and then gets trained, and the trained DNN together with ELM (Extreme Learning Machine) forms the classifier for identifying four kinds of emotional traits occurred in the conversion data [2]. After that, an array including 4 kinds of emotion values is got.

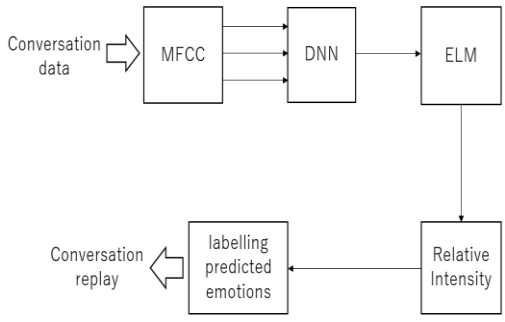

Fig.1. The system overview

There are four emotional traits are recognized from each segment along the time axis with recognition precision values. At a time t, the sum of precision values of 4 kinds of emotion trains is 1. With these precision values, each emotion traits the relative intensity of each emotional traits is calculated and the emotional traits are labeled accordingly. In order to make the system more widely used, the VAD (Voice activity detection) algorithm in the section 5 is used to solve the problem of time regions.

3. Methods for speech emotion recognition

In this section, the details of methods and models that are used to get the emotion state of speech are analyzed. Firstly, the method to extract low-level features of speech is introduced. MFCC is chosen in this paper. In the following paragraphs, the reason making a choice of MFCC is analyzed. What is more, the main extraction process of MFCC are introduced. Then, the low-level features extracted by MFCC are put into a DNN for deep learning. Finally, the features predicted by the DNN are collected, so that they are used in ELM as input for classification. Therefore, we can obtain a vector of each emotional trait's precision which is further used as each emotional trait's probability.

As for speech signal, waveform is a representation close to our daily life. At the first phase, a conversation audio data in the time domain is transferred into a waveform in the frequency domain. Any .wav speech file in the dataset and visualized the waveform. As shown in Fig.2, the x-axis represents the total length of speech file. The y-axis represents the amplitude at the moment. The place where many lines concentrate in the picture means the high energy at this time, in other words, it is more likely that two people are talking at the moment in this dataset.

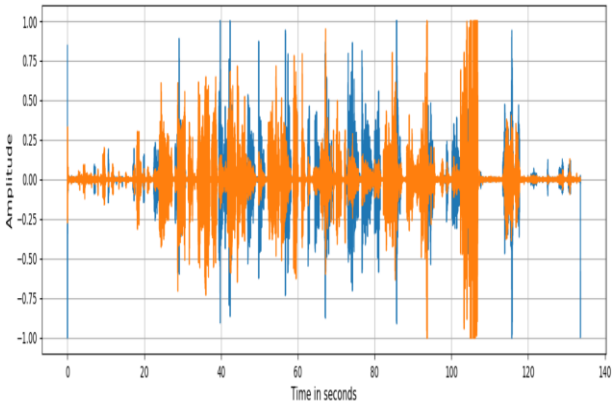

Fig.2. The waveform of a speech file

At the second phase, two people's conversation waveform has to be separated into two waveform graphs as given in Fig.3 and Fig.4. Since these speech dataset was collected by two microphones used by the two people who conduct conversation with each other. In Fig.3 and Fig.4, different colors are used to represent the waveform. The blue sound wave in Fig.3 is the blue part in Fig.2, we show them with the same color. Similarly, the orange sound wave in Fig.4 is the orange part in Fig.2.

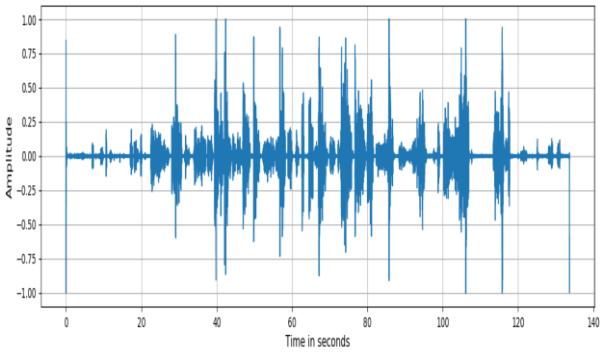

Fig.3. The waveform from one microphone

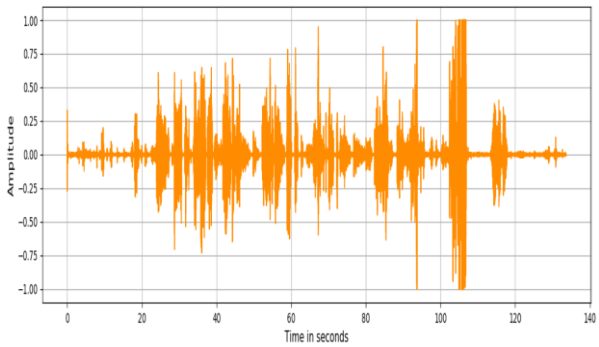

Fig.4. The waveform from another microphone

3.1. MFCC features extraction

As the first step of speech emotion recognition, the extraction of speech feature parameters is very important. In addition to MFCC used in this paper, there is also LPCC (Linear Prediction Cepstral Coefficient) [3]. Through comprehensive evaluations including accuracy, the most suitable MFCC is selected in this article.

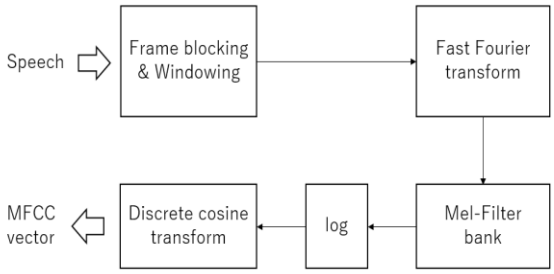

Fig.5. The diagram of MFCC

In the above, the visualized representation of the speech is shown in the time domain. However, the computer cannot directly extract valid feature information from the waveform graph. So the raw audio file from the time domain to the frequency domain need to be converted.

The FFT (Fast Fourier transform) is the most important step in MFCC, transforming the time domain signal into the frequency domain signal, so that computer can understand the information of the signal. Compared with the DFT (Discrete Fourier transform), FFT is faster because it calculates fewer multiplications.

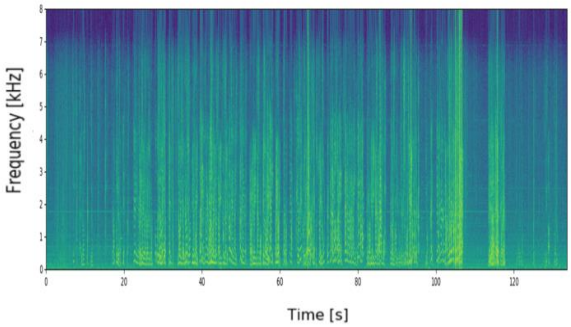

Fig.6. The spectrogram of MFCC

Fig.6 gives the spectrogram of MFCC. It can be seen that some places are with brighter color and some places are not. It seems that the color near at the 105 second in Fig.6 is much brighter than other places which indicates that the energy is higher at this time, which means the conversation is more intensively conducted.

3.2. DNN train speech emotion features

The internal structure of DNN (Deep neural network) is characterized by the hidden layer being multilayer, and the rest also includes the input layer and the output layer [4]. After getting the vector from MFCC, we train a DNN model to analyze the frame-wise emotion distribution. In addition to the input layer and the output layer, the number of hidden layers are 3 and every hidden layer's dimension is 256. The input is MFCC vector and the output is emotion trait of frames. Then the output of DNN is fed into ELM.

3.3. Emotion classification

To classify emotions, ELM (extreme learning machine) is used in this experiment. In short, ELM is a single hidden layer neural network. Compared with other learning algorithms, for example SVM (Support Vector Machine), ELM is characterized by faster speed to classification, high accuracy, and the ease of using ELM is guaranteed by its less manual intervention [5]. Thus, ELM is very suitable for this system.

4. Emotion relative intensity

In this section, in order to label the emotion, the softmax function is chosen to calculate the relative intensity of emotional traits. The softmax function can map the original output value from the speech emotion recognition model into probabilities and normalize these probabilities. The output value of the Softmax function is related to each other, and the sum of these

probabilities is always 1, the formula of the softmax function as follows.

$$\sigma(z)_j = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}}$$

for j = 1, ..., k

Therefore, in the Softmax function, if the probability of one emotion is increased, the probabilities of other emotions are correspondingly reduced. The outputs of the softmax function are interrelated and interact on each other, but the sum of these numbers representing emotions remains 1. Fig.7 shows that the array which is not easy to understand before using softmax function. Fig.8 shows that the emotional traits are normalized into 1 by softmax function.

```
[[-14.934883   26.262486   10.157334  -22.066013 ]
 [-15.9633     26.452768   11.420645  -20.959305 ]
 [-15.348414   26.395176   11.021883  -21.600857 ]
 ...
 [  4.5614715  -9.267523    2.7686992   6.470931 ]
 [  4.6459417  -9.5089855   3.0485294   6.853306 ]
 [  5.158002  -10.01465     2.7772632   6.7882895]]
```

Fig.7. The array before softmax function

```
[[1.2829510e-18 9.9999988e-01 1.0130275e-07 1.0261234e-21]
 [3.7925838e-19 9.9999976e-01 2.9623175e-07 2.5656592e-21]
 [7.4299953e-19 9.9999976e-01 2.1060254e-07 1.4308291e-21]
 ...
 [1.2632740e-01 1.2463548e-07 2.1033252e-02 8.5263926e-01]
 [9.7142644e-02 6.9183585e-08 1.9663578e-02 8.8319367e-01]
 [1.6134691e-01 4.1529937e-08 1.4921721e-02 8.2373130e-01]]
```

Fig.8. Normalized precision by softmax function

5. Emotion annotation

In the process of labeling, there are two aspects need to be pay attention, the first one is that time regions need to be known, and the second one is the emotion prediction corresponding to that time. There are many useless time regions where no emotion is contained. For example, there might be some silent regions between speaker's utterances. If the speech is chosen to annotate it in IEMOCAP (Interactive emotional dyadic motion capture) dataset, the timestamps are already prepared. Hence, only emotion prediction are needed [6]. While, using the speech is not in IEMOCAP dataset, the way to get timestamps is needed to find. One of the solution is applying VAD (Voice activity detection) algorithm to find timestamps in speech [7]. The VAD algorithm can be implemented with zero-crossing rates and energy. For example, if the current frame contains relatively higher energy than the normal frame, especially when it has higher energy distribution over the frequencies of human speech, then we know it has the voice activity. By applying VAD, we can divide speech into several parts and remove silences.

In Table 1, the effect of using VAD algorithm is shown. Choosing a speech in IEMOCAP dataset to prove the VAD algorithm is useful. As a result, VAD algorithm can divide the speech into 26 utterances. At the same time, 17 correct utterances are divided which meaning the precision rate is 60.7%.

Table 1. The voice activity detection algorithm

	VAD	Correct answer	Precision rate
All divided utterances	26	28	92.9%
Correct divided utterances	17	28	60.7%
Wrong divided utterances	9	28	32.2%

6. Implementation and results

In this paper, the IEMOCAP (Interactive emotional dyadic motion capture) is used as experimental dataset to train model. The IEMOCAP database is annotated by multiple annotators into category tags such as anger, happiness, sadness, neutrality. The whole speech files are divided into 5 sections. Every section approximately contains 30 conversation data. Every conversation data is about 2 minutes.

The dataset includes a lot of dialogues composed of a male vocal and a female vocal. The size of all speech files in dataset is 3G, so that these reliable dialogue datasets are useful to do machine learning. Before comparing the experiment results, all the speech emotion marked by professionals were pre-processed, as shown in Table 2.

Table 2. The emotion tags of sections in dataset

	Sad	Happy	Neutral	Anger
Section1	90	63	161	167
Section2	97	63	145	115
Section3	115	75	122	150
Section4	62	34	84	243
Section5	112	37	97	139

After comparing with the data in Table 1, the accuracies of speech emotion recognition are represented by WAR (weight average recall) and UAR (unweight average recall). The details of them shows in Table 2. Numbers in Table 3 mean percentages. Comparing

training 50 and 100 times, and the accuracy rate of epoch equaling 100 is better than that of epoch equaling 50.

Table 3. The accuracy with different training epoch

	WAR	UAR
Epoch=50	46.64	45.6
Epoch=100	47.14	46.66

In short, although the accuracy rates have not increased so much, it can be found that more training times will have better results.

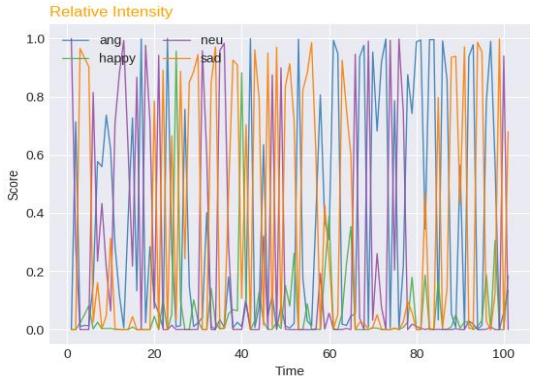

Fig.9. The emotion intensity of a speech file

After obtaining the low-level features through a series of operations, the deep neural network learning method is used to perform preliminary sentiment judgment on the utterance-sized fragments. Firstly, an array of emotions is obtained, and then the normalization method are performed to get relative intensity above all emotions. The relative emotion intensity of every utterance-level features is shown on Fig.9. The orange line means probability of sad, the green line means probability of happy, the purple line is probability of neutral, and the blue one is probability of anger. X-axis gives the times that we get the relative intensity. In Fig.9, 100 times are chosen. Y-axis shows the relative intensity, from 0 to 1, to identify the speech emotion should be labelled.

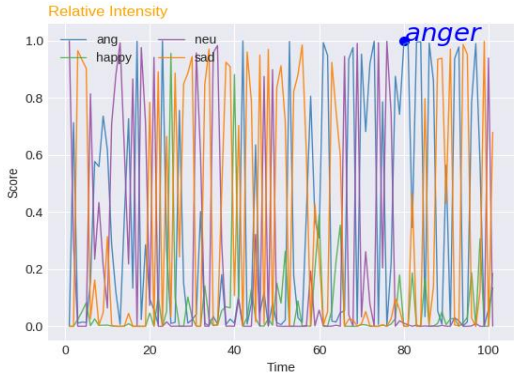

Fig.10. The correct prediction emotion

As Fig.10 shows that the blue dot is used to annotate the correct prediction emotion compared with correct answer, and the name of emotion is written over the blue dot.

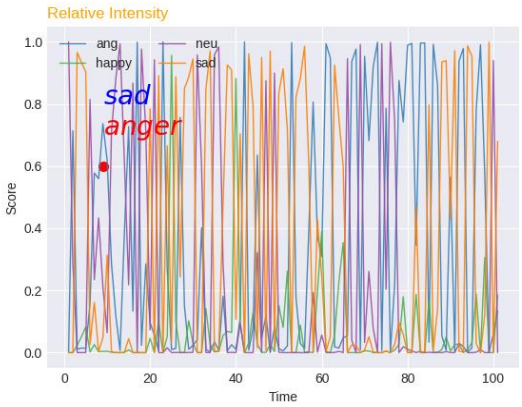

Fig.11. The wrong prediction emotion

As Fig.11 shows that the red dot is used to annotate the wrong prediction emotion compared with correct answer, and the name prediction emotion is written by red color. The correct answer of emotion using blue color over the wrong emotion.

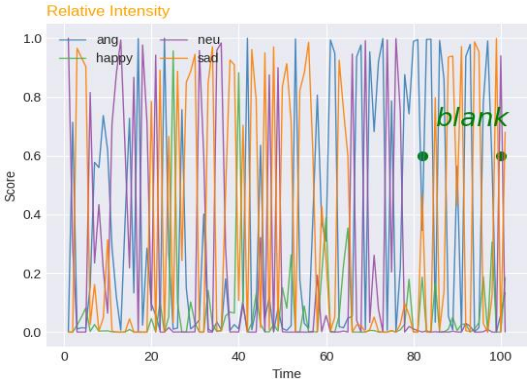

Fig.12. The blank space

As Fig.12 shows that the green dot is used to annotate the start and final points of blank space, and the blank space is written by green color.

Through the method is described earlier, the real-time emotion information of the speech are obtained on the GUI (Graphical user interface). Fig.13 shows the basic GUI. The GUI has three graphs including waveform graph, spectrogram graph, and emotional traits labeling graph. The path of audio file is shown on the playlist. The audio file can be added by Add button. Also, the useless audio file can be delete by Del button. When paly button is clicked, the three graphs update in real-time. When the audio file is played, the volume can be controlled.

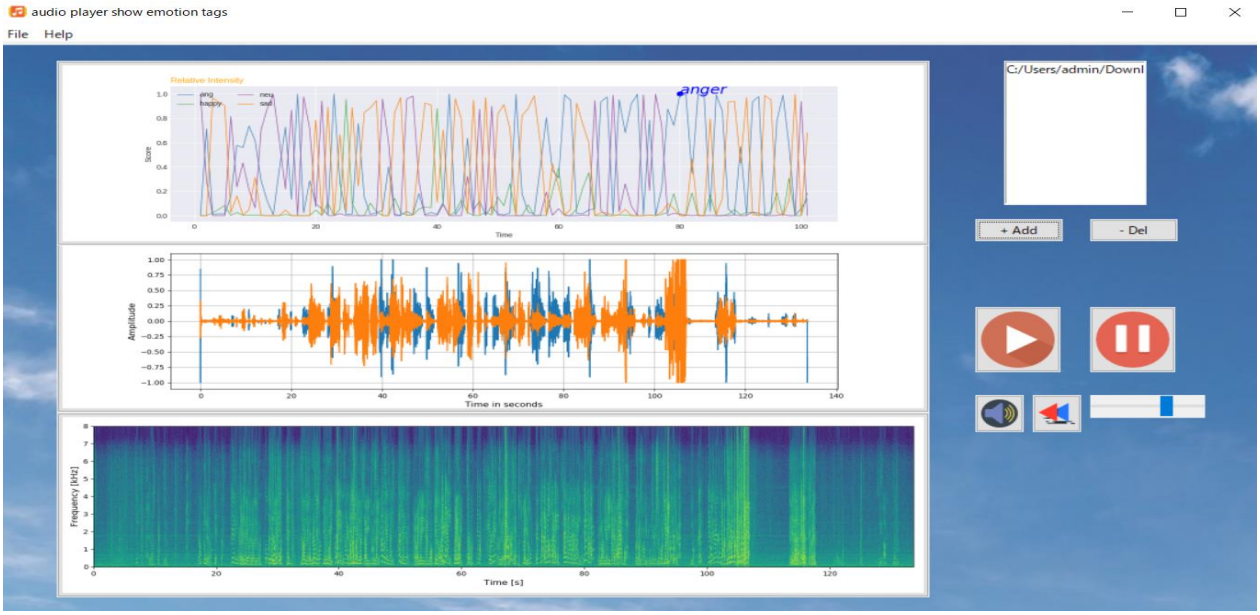

Fig.13. The GUI of emotional traits labeling

7. Comparison

Comparing with the previous speech emotion recognition systems, this paper completes the calculation of the relative intensity of emotional traits, and then annotates the emotion labels. Besides, in comparison with other real-time speech emotion recognition systems, the labels of emotions can be more clearly represented the relative intensity of emotions. Since the method named VAD is used, speech datasets without timestamps can also be used in the system to annotate emotions.

8. Conclusion and future work

In this paper, analysis and annotation of emotional traits in human conversation are introduced. First of all, MFCC is used to extract the features of the pre-processed speech dialogue. After the extraction, the deep neural network is trained to make the emotions to obtain a representation in a longer time range. Since the length of each sentence in our conversation may last 1-2 seconds, the emotion features of the ms unit frame can not express human true feelings very well. After that, the extreme learning machine proposed by Huang.et.al [5] is used to classify the obtained data into emotions, using its simple and fast features to make our system better. Besides, the array of four emotional traits is normalized, so that the emotional traits can be better represented and it is helpful for the subsequent labeling process. Finally, the emotions are annotated at the necessary positions in the figure. There are three kinds of dots are annotated in the figures which mean correct prediction emotion, wrong prediction emotion and blank space can be seen in the figure. At the same time, a GUI system has been made to visualize this intermediate process, so that people can understand this process more easily. In the future, the GUI need to be more aesthetically pleasing.

文　献

[1] 荒井和博. "コンタクトセンタ向け音声マイニングシステム ForeSight Voice Minig." 研究報告情報基礎とアクセス技術 (IFAT) 2019.6 (2019): 1-6.

[2] Han, Kun, Dong Yu, and Ivan Tashev. "Speech emotion recognition using deep neural network and extreme learning machine." Fifteenth annual conference of the international speech communication association. 2014.

[3] Muda, Lindasalwa, Mumtaj Begam, and Irraivan Elamvazuthi. "Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques." arXiv preprint arXiv:1003.4083 (2010).

[4] Deng, Li, Geoffrey Hinton, and Brian Kingsbury. "New types of deep neural network learning for speech recognition and related applications: An overview." 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2013.

[5] Huang, Guang-Bin, Qin-Yu Zhu, and Chee-Kheong Siew. "Extreme learning machine: theory and applications." Neurocomputing 70.1-3 (2006): 489-501.

[6] 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," Journal of Language Resources and Evaluation, vol. 42, no. 4, pp. 335-359, December 2008

[7] Sehgal, Abhishek, and Nasser Kehtarnavaz. "A convolutional neural network smartphone app for real-time voice activity detection." IEEE Access 6 (2018): 9017-9026.