DOI: 10.15439/2017F533
ISSN 2300-5963 ACSIS, Vol. 13

# Extracting Acoustic Features of Japanese Speech to Classify Emotions

Takashi Yamazaki
Human System Science
Tokyo Institute of Technolgy
Tokyo 152-8552, Japan
Email: yamazaki@nk.ict.e.titech.ac.jp

Minoru Nakayama
Information and Communications Engineering
Tokyo Institute of Technolgy
Tokyo 152-8552, Japan
Email: nakayama@ict.e.titech.ac.jp

*Abstract*—**An emotional detection technique which extracts acoustic features from audio recordings of speech was developed. Though the formant frequency of individual voices may contribute to emotional variations in speech, the differences between vowels have an influence on feature extraction. To reduce the influence, a simple procedure was developed to extract relative features of vowels for every mora. The estimation performance of this emotional detection technique was improved by 11% using relative formant frequencies instead of formant frequencies. The strengths of some emotional expressions were also reflected in some features. The effectiveness of using acoustic features to estimate the category of emotionally inflected speech was confirmed.**

## I. Introduction

EMOTIONAL presentation is a key factor for sharing information in human communications. Its importance is not only for human-human communications, but for human-robot communications and human computer communications as well. Possible resources for estimating the emotional state of people using natural verbal communication habits while speaking is verbal information (context of speech), acoustic feature of voices (speaking tone), and also visual features (facial expressions, behavioral actions).

Uemura [1] compared the degree of ability to detect emotional inflections using these resources, and suggested the order of intensity of presenting emotions by level of importance was voice audio > facial expressions > verbal contents. Therefore, any systems which include communication robots may be able to recognize the emotional states of speakers and may well play an important role as a communication partner once features of emotional expressions can be extracted and utilized. Although some robots have already been designed to recognize or present emotions during communications with humans using features of language and behavior, more specific features are required to improve their performance. As voice audio is the effective source of providing emotional inflections, acoustic feature of voices should be used for the communication aids. The appropriate technique to extract significant features from voice audio and their applications are required. When these techniques were developed, robots may be possible to detect emotional states of a human speaker as well as synthesizing an appropriate speech in response to human's emotional conditions. These fundamental techniques, in particular, emotional recognition and presentation in voice audio, should be developed.

The emotional information processing for speech as voice audio is still tough work in comparing with speech recognition and speech synthesis for even a specific speaker. As the problems are caused by phonological issue of language, additional lexical information are also used for predicting emotional state [2]. Previous phonological studies suggest that phonological components can represent factors of emotions [3]. The formant frequency is one of the features of phonological information of speech. When the formant frequency was used to predict emotion, the accuracy was 90% when the spoken words were identical [3]. Therefore, the formant frequency is confirmed as being one of the significant features which can be used to detect the emotional state of the speaker. However, performance for mixed sentences which contain several vowels decreases 15–20% [3]. Since formant frequencies for each vowel are different, phonological information depends on the combinations of vowels in sentences.

As mixed sentences are naturally used in our communications, formant frequency should be employed as a feature of speech when considering vowels in mixed sentences. Therefore, an appropriate procedure for processing phonological information from mixed sentences is required in order to improve the performance of predicting emotional states, and this technique may also be capable of providing features of phonological information for use with the synthesis of voices containing emotions [4].

This paper will address the development of a procedure for the extraction of formant frquencies as phonological information of speech from spoken words which contain multiple vowels, and the effectiveness of these features is evaluated in order to predict the emotional state of mixed sentences. The details are as follows:

1) To develop a procedure for extracting acoustic features using the formant frequencies of every vowel, and to also develop a procedure for specifying the emotional state (emotional category) of speech using the extracted features.
2) The contributions of each acoustic feature are compared in order to determine the key information necessary for

TABLE I
MEAN F1S [HZ] AND EXAMPLES OF RELATIVE RATIOS

| Emotions | Ka | Ze | Bu | Ki |
|---|---|---|---|---|
| Neutral | 450.5 | 488.8 | 384.3 | 299.5 |
| Surprised (Strength=1) | 521.0 | 509.3 | 398.1 | 409.7 |
| Relative ratio | 1.16 | 1.04 | 1.04 | 1.37 |

TABLE II
RESULTS OF F-TEST OF VOWELS (EXP. VALUES)

| Features | df | MSE | F | $p$ |
|---|---|---|---|---|
| Fundamental Freq. | (4,96) | 4198.97 | 2.10 | n.s. |
| Power | (4,96) | 28.34 | 4.25 | $< 0.01$ |
| F1 | (4,96) | 17272.32 | 17.48 | $< 0.01$ |
| F2 | (4,96) | 134190.75 | 20.69 | $< 0.01$ |
| F3 | (4,96) | 116989.82 | 1.24 | n.s. |

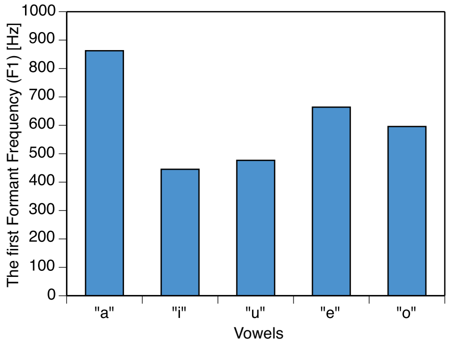

Fig. 1. Comparison of mean F1s between vowels

TABLE III
RESULTS OF F-TEST OF VOWELS (RELATIVE RATIOS)

| Features | df | MSE | F | $p$ |
|---|---|---|---|---|
| Fundamental Freq. | (4,96) | 0.06 | 1.92 | n.s. |
| Power | (4,96) | 0.01 | 0.05 | n.s. |
| F1 | (4,96) | 0.06 | 5.04 | $< 0.01$ |
| F2 | (4,96) | 0.15 | 3.46 | n.s. |
| F3 | (4,96) | 0.01 | 0.45 | n.s. |

prediction of emotional states and strength levels of an emotion.

The results of these analyses will provide basic information to be used to produce emotionally inflected speech using speech synthesis engineering technologies.

## II. METHOD

### A. Source of speech sound

An audio processing procedure was developed using a set of voice audio corpus for short speech phrases (a voice audio corpus which consists of short phrases spoken by voice actors/actresses). This corpus was originally developed to record emotionally inflected voice audios (voice audio spoken by voice actors/actresses portraying emotions) in online gaming using voice chats which are used to compose an audio library [6]. Features of acoustic characteristics used to estimate emotional states are comparable when extracted from phrases or whole sentences [5]. In addition, two male and female voice actors recorded emotionally inflected voice audios of some phrases from chats from online games. The categories of emotions are based on Ekman's 8 basic emotions [7]. All phrases are presented and recorded in a neutral tone and 3 strength levels of every emotion (for example, very happy, moderately happy, and slightly happy). Though the extracted phrases have no meanings as independent terms, the emotional impressions of the phrases can be easily recognized.

### B. Analysis of audio data

Three phrases were randomly selected from the corpus and used as a set of test data for four emotions: "happy", "angry", "sad" and "surprised". These were the same emotions that were used in a previous study [4], because the emotional features may be present in those categories of emotions to a significant degree. The overall number of test phrases is 48, which consists 3 phrases × 4 emotions × 4 levels of emotional strength: neutral, and strength levels 1∼3.

Audio data was analyzed using software known as "Praat" [8]. Seven acoustic features, such as fundamental frequency, power, duration, speech speed, and 3 formant frequencies (F1-F3: the first, second and third frequencies as the formant frequency) were extracted manually.

## III. ACOUSTIC FEATURES

### A. Features between vowels

To compare the lexical features of phrases, the "mora" unit was introduced. The mora is a unit of lexical sound and corresponds with a piece of speech that includes a vowel. Therefore, for every mora the acoustic features mentioned above were extracted, except for duration and speed of speech. Exceptions were evaluated phrase by phrase.

Though the mora can be a useful unit for acoustic features, means of features of mora are used for evaluation, as the duration of a mora is too short. Therefore, an appropriate procedure that compensates for this is required in order to consider combinations of vowels.

Here, an example of analysis of the phrase "KaZeBuKi", which consists of 4 moras, is shown in Table I. Table I shows F1 frequencies for neutral and surprised speech (strength level=1) for every mora. In comparing the frequencies of the two speech conditions, the differences show the contributions of emotionally inflected speech.

In order to examine the factors for 5 categories of vowels using acoustic features, one-way ANOVA as F-test was conducted. This analysis was conducted using features of every mora, and the number of samples for each category differed. The strength of emotions was ignored during this analysis.

The F values for the main effects of variations, ratios of two variances, and their degrees of freedom (df) are summarized in Table II, and the effects are significant for power, F1 and F2 ($p < 0.01$). The results confirm the effects of acoustic

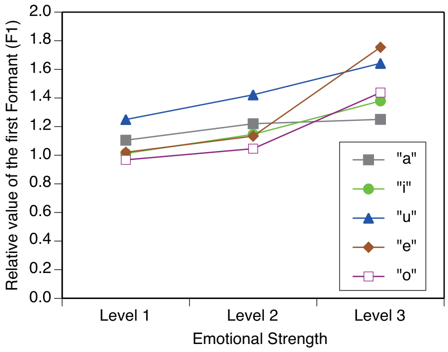

Fig. 2. Comparison of F1 relative ratios between vowels

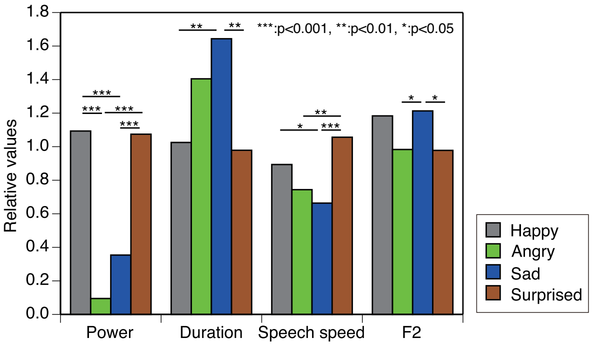

Fig. 3. Comparison of feature values between vowels

features between vowels. Figure 1 shows mean frequencies of F1 between vowels, and the results of sub-effect tests are presented. There are some significant differences between vowels, and the tests have confirmed that the fequency of the vowel “a” is the highest.

### B. Relative features

Even with neutral emotion, the acoustic features between vowel categories are different. The relative values of features of emotional speech features which use neutral expression were introduced. The relative values as “relative ratio” are displayed in the third line in Table I.

The same analysis was applied to relative values. The results are shown in Table III. The Table shows that the factor for vowels is significant for F1 ($p < 0.01$), however. Mean relative values of F1 between levels of emotional strength are summarized in Figure 2. Some changes are observed at the level 3 of emotional strength. For the further analysis, multiple comparison between vowels was conducted as the sub-effect tests. The results show that there are no significant differences between vowel categories. Therefore, the factor of vowels with acoustic features decreased when the relative values were introduced.

### C. Differences in features between emotional categories

The overall features of phrases can be generated using relative features. To extract the differences between the types of emotional presentations, one-way ANOVA was conducted on each acoustic feature. In the results, the effectiveness of the features of some types of emotional presentations, such as power, duration, speech speed and F2, is significant. The effectiveness is illustrated in Figure 3, which shows the results of sub-effect tests of the four emotions with significant levels.

These results show that emotional presentation changes some of the means of acoustic features.

TABLE IV
RESULTS OF DISCRIMINANT ANALYSIS WITHOUT FORMANT FREQUENCIES

| Source emotions | HAP | ANG | SAD | SUP | Error (%) |
|---|---|---|---|---|---|
| Happy | 5 | 0 | 0 | 4 | 44.44 |
| Angry | 0 | 7 | 2 | 0 | 22.22 |
| Sad | 1 | 3 | 4 | 1 | 55.56 |
| Surprised | 1 | 0 | 1 | 7 | 22.22 |
| Error rates | | | | | 36.11 |

TABLE V
RESULTS OF DISCRIMINANT ANALYSIS WITH FORMANT FREQUENCIES

| Source emotions | HAP | ANG | SAD | SUP | Error (%) |
|---|---|---|---|---|---|
| Happy | 7 | 0 | 0 | 2 | 22.22 |
| Angry | 0 | 7 | 2 | 0 | 22.22 |
| Sad | 1 | 1 | 7 | 0 | 22.22 |
| Surprised | 3 | 0 | 0 | 6 | 33.33 |
| Error rates | | | | | 25.00 |

## IV. DISCRIMINANT ANALYSIS OF EMOTIONS USING ACOUSTIC FEATURES

### A. Procedure and performance

As some acoustic features reflect emotionally inflected speech, the possibility of predicting the type of emotional presentation using discriminant analysis exists. The amount of data was limited, and the randomForest technique was used for classification. The strengths of emotions were ignored in this analysis.

As mentioned in the introduction, the effectiveness of formant frequency should be tested using discriminant analysis. An analysis was therefore conducted with and without using formant frequencies.

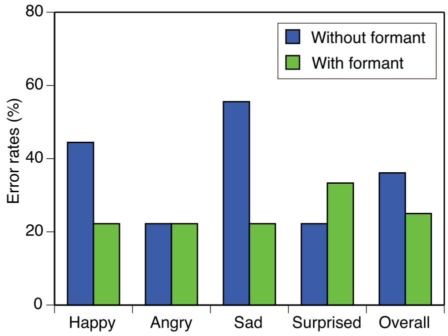

Fig. 4. Error rates of Emotions

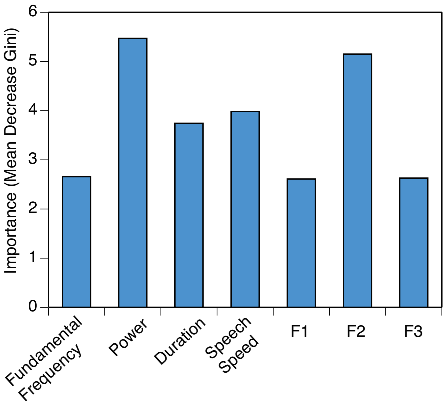

Fig. 5. Importance of acoustic features

TABLE VI
RESULTS OF F-TESTS OF THE STRENGTHS OF EMOTIONAL VOICES

| Features | df | MSE | F | $p$ |
|---|---|---|---|---|
| Fundamental Freq. | (2,102) | 0.05 | 13.35 | $< 0.01$ |
| F1 | (2,102) | 0.08 | 7.21 | $< 0.01$ |

Table IV shows the results of the discriminant performance without formant frequencies, and Table V shows the performance with formant frequencies. In both tables, the vertical cells represent emotional categories of speech, and the horizontal cells represent the discriminant results. The orthogonal components represent correct discriminations. The percentages of misclassification frequencies are calculated for each emotion, and they are summarized as error rates in the far right columns. In sum, the overall error rate decreased 11% when features of formant frequency were used. The error rates for the four emotions are summarized in Figure 4. The figure shows that formant frequency contributes to discriminant performance for "happy" and "sad" emotions, but the improvements are not present for all emotions.

### *B. Importance of acoustic features for classifying emotional states*

The degree of importance of features, which means the contributions of classification of emotional state discrimination, are summarized in Figure 5. The most prominent features are power and the 2nd formant frequency (F2), which is related to the position of the speaker's tongue. The factor of emotionally inflected speech may be a concern when the tongue is moving.

### *C. Relationship between acoustic features and emotional strengths*

The factor of emotional strengths for the three voice levels (since voices are presented using three levels of emotional strength) was tested for the acoustic features which were extracted using one-way ANOVA. In the results shown in Table VI, the main factor is significant for fundamental frequency and for 1st formant frequency (F1). Changes in the values for emotional strengths are summarized in Figure 6. Sub-effect tests show that there are some significant differences between the strengths of emotionally inflected speech, and the results are illustrated in Figure 6. These two features may respond to the level of the emotional presentation, and a detailed analysis of this will be the subject of our further study.

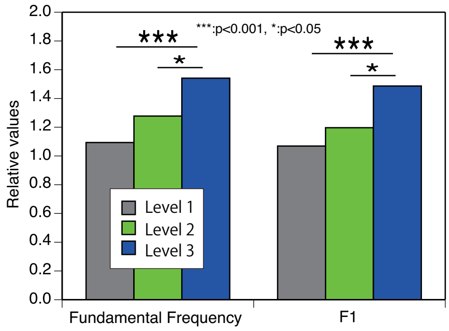

Fig. 6. Results of multiple comparisons of the strengths of emotionally inflected voices

## V. SUMMARY

The relationships between emotionally infected speech and its acoustic features were analyzed. In particular, a signal processing procedure for use with formant frequencies between vowels was developed to examine their contributions to detecting emotional factors. During the procedure acoustic features, including formant frequencies, were evaluated using relative values for emotional speech which were compared to relative values for natural speech for every mora unit. The features of phrases were evaluated as means of features between moras.

This technique improves discriminant performance between the four types of emotional speech by comparing formant frequencies such as F2. Performance was improved 11% for speech without formant frequencies. Power and the 1st formant frequency are the two acoustic features which influence the strength of emotional speech the most.

The development of practical application procedures including adaptations for phonological features of other languages system such as English will be the topics of further study.

## REFERENCES

[1] J.Uemura, K.Mera, Y.Kurosawa, T.Takezawa, "Analysis of Inconsistency among Emotions Estimated from Linguistics, Acoustic, and Facial Expression Features and A Proposal of the Inconsistency Detecting Method," Proc. of 78th annual meetings of IPSJ, 6Y-04, 4, 321–322, 2016.

[2] T. Matsui, M. Hagiwara, "A Dialogue System with Emotion Estimation and Knowledge Acquisition Functions," Trans. of Japan Society of Kansei Engineering, 16(1), 35–42, 2017. doi: 10.5057/jjske.TJSKE-D-16-00058

[3] M. Shigenaga, "Features of Emotionally Uttered Speech Revealed by Discriminant Analysis," IEICE Trans., Vol.J83-A, No.6, 726–735, 2000.

[4] M. Shigenaga, "Characteristic Features of Emotionally uttered Speech Revealed by Discriminant Analysis (III): Discrimination of both Mixed Sentences and Test Data," IEICE Technical Report, SP, 97(396), 65-72, 1997-11-21, 1997.

[5] M. Shigenaga, "Characteristic Features of Emotionally uttered Speech Revealed by Discriminant Analysis (VI)," Proc. of Acoustic Society of Japan, 3-3-12, 1999.

[6] NII Speech Resources Consortium, "Online gaming voice chat corpus with emotional label (OGVC)," URL http://research.nii.ac.jp/src/OGVC.html

[7] P. Ekman, W.V. Friesen, Unmasking the face, Prentice-Hall, Inc., NJ, USA, 1975.

[8] Paul Boersma, David Weenink, http://www.fon.hum.uva.nl/praat/ (accessed 27th Jan., 2017)

[9] H. Jouo, NIHONGO ONSEI KAGAKU, Badai Music Entertainment, Tokyo, Japan, 1998.