Analysis of Emotional Expression by Visualization of the Human and Synthesized Speech Signal Sets

- A consideration of audio-visual advantage-

Kazuki Yamamoto
Hokusei Gakuen University
Faculty of Economics, Department of Management Information
Sapporo Japan
i15126@hokusei.ac.jp

Keiji Takahashi
Hokusei Gakuen University
Faculty of Economics, Department of Management Information
Sapporo Japan

Kanta Kishiro
Hokusei Gakuen University
Faculty of Economics, Department of Management Information
Sapporo Japan

Shunsuke Sasaki
Hokusei Gakuen University
Faculty of Economics, Department of Management Information
Sapporo Japan

Hidehiko Hayashi
Hokusei Gakuen University
Faculty of Economics, Department of Management
Sapporo Japan
hhayashi@hokusei.ac.jp

Abstract—**In recent years, speech synthesis technology has been developed, but the method of expressing emotions with synthesized speech was one of the issue. Therefore, in the latest speech synthesis technology, a method for expressing emotion by synthesized speech continues to be developed. For example, as one of them, it is a method of changing the physical values such as the height of the sound and the speaking speed, reproducing emotions such as joy and anger, and reading out the text. Emotional expression of synthesized speech is to reproduce human emotions by physically changing the height of sound, speech speed, intonation, and the like. However, we do not hear much comparison between synthesized speech expressing emotion and human one. Therefore, in this study, we conducted a control experiment on whether there is a difference between synthesized speech expressing emotion and human speech. In this control experiment method, two sounds, one synthesized by using one of the speech synthesis software and one recorded with human speech, are visualized as speech signals, and the differences that can be analyzed from these characteristics are listed. In this research, the speech signal is displayed as a waveform, frequency analysis and so on. As a result, various speech features were able to be grasped. For example, in the voice of anger by humans, as the anger becomes stronger, the strength of the vowel becomes stronger and the overall volume increases, it turned out that the synthesized speech also reproduced by imitating the feature.**

Keywords—Speech synthesis; Emotional expression; Emotional voice; Audio visual advantage

I. INTRODUCTION

In recent years, the development of speech synthesis software is accelerating. Speech synthesis software is software that reads input text as volume, inflection, speed etc. as set by the user as it is set, and is used in various situations such as voice guidance by telephone correspondence and creation of video contents. However, to improve the quality of speech synthesis software development, there are several problems in the part of reading text input by humans naturally like a real human being or reading it to express it with emotion. As one of the approaches to solving those problems, we are paying attention to emotional expressions of speech. Since there is no concept corresponding to "emotion" which expresses the psychological state of human beings in speech synthesis software, the expression of emotion expression of synthetic speech is defined as reproducing the voice when a person is angry or pleased. In order to reproduce various psychological states such as anger, pleasure, sadness and the like by human beings as "emotions" and reproducing them with the synthesized speech, it is essential to first know the characteristics of human voice.

In the latest advanced speech synthesis software which has been developed, techniques for reproducing human's angry voice, delightful voice, sad voice are developed just by changing parameters. Therefore, we focused on emotional expression by this latest speech synthesis software and visualized and analyzed speech signals. Also, we compare the synthesized speech with the human speech. For distinguishing emotional expressions, categorization from 2 to 18 categories has been studied mainly on cognitive psychology. In this study, as a signal processing of synthesized speech, focusing on representative emotional expressions being investigated also in previous research, we examined the distinction between the three emotions "anger" "sadness" "joy" based on "calm". In addition, we support the hypothesis that the degree of emotion is a continuous value, and adjusted the parameters in 5 stages. These speech signals were visualized and analyzed for their characteristics.

978-1-5386-2615-3/18/$31.00 ©2018 IEEE

II. PREVIOUS RESERCH

In recent years the development of robots and the like has evolved, mainly used for corporate activities and labor work / service work. Although there are still many issues and areas that are evolving, there is a need for a robot capable of acting as a human partner. Emotional expressions are one of the issues for making behavior that is closer to human beings. Since it is still difficult to artificially impart a continuous psychological state held by a human being to a robot, various researches have been done from the viewpoint of how to reproduce the human emotions. Therefore, there are several papers to implement emotional expression methods in speech synthesis technology [1][2][3].

For example, The paper [1] mainly focuses on communication including emotions that are often seen during human conversation. When communicating between robot and human being, it is assumed that communication with smooth sense of discomfort can be obtained by giving the characteristics of emotion to the voice output from robot. This paper focuses on four emotions of joy, anger, sadness and surprise, and analyze the three features of fundamental frequency, utterance speed, and sound magnitude using speech corpus voice data. In addition, in the paper [2], it is assumed that the emotion transmitted by voice has "speaker's emotion" and "listener's emotion", and in order to inform the listener of the emotions included in the synthesized voice, "emotions of the listener". It states that the physical characteristics of voice should be formulated. One of the physical features of the voice is to extract pitch trajectories from voices including various emotions and visualize the pitch of each emotional voice.

Furthermore, there are papers considering features of recognition from the point of view of listening side, such as paper [5] how human beings recognize emotions contained in speech. By reproducing the recognition features obtained by this research on a machine, it is possible to communicate with each other by speech dialogue between human and machine.

These papers have focused on human speech features and have used methods to analyze and visualize characteristics of human utterances as speech signals. However, little mention is made of feature comparison of human utterances and the latest speech synthesis software such as advanced text-to-speech software implementing additional function of emotional expression. Therefore, it is not very clear how the speech synthesis technology at the present stage differs from human speech. In this research, we focused on that point and analyzed the differences of these features.

III. PURPOSE OF THIS STUDY AND RESEARCH METHOD

The purpose of this research is to improve the quality of speech synthesis software, and as one of the problems to be solved for that, we focused on emotional expression of speech synthesis software. Therefore, paying attention to the difference between the latest voice synthesis software and the human voice, the differences are listed from each voice characteristic.

The latest synthesized speech software we used is "VOICEROID 2 Yuduki Yukari (VOICEROID 2)". Features of this software are equipped with a function to read the text with voice expressing each emotion only by changing parameters, 3 feelings of "joy", "anger", "sadness". This software was used because each emotional parameter can change the intensity with a numerical value of 0 to 1.0, and the five-step parameter adjustment is also easy.

Human speech uses speech data obtained by our previous recording "Evaluation of emotional expression by human voice". In this research, the speaker reads the text and records it, but when reading it, the speaker imaged and read the emotions of "calm", "anger", "sorrow" and "joy". This was set as emotional speech (hereinafter referred to as emotional speech), and the speaker practiced for one week's practice period, exercises to distinguish emotions to read out, and recorded. For speakers who read the text, they were looking for women around age 20 to compare with voice data of VOICEROID 2 and got cooperation. Also, since it was judged that theater experienced person was more appropriate than the general person (non-expert) to reproduce the image feeling, he made the condition of experiencing theater. As a result of selecting those under the condition of women and experienced players around 20 years old, they were carefully selected by three collaborators. Speaker A belonged to a theater company in 2015. Speaker B belonged to the theater department at junior high and high school, and belonged to the theater circle at university. Speaker C belonged to the theater department at high school and belonged to the theater circle at university.

We prepared two kinds of texts, "Oh, that's right" and "Another, than you" for the recorded sound. Since these texts are commonly used on a daily basis, we decided that it is suitable for use in this research. Furthermore, although the influence of meaning as language information on these texts is small, the influence of non-verbal information such as emotional expression by voice can be largely reflected. For this reason, we chose these two kinds of texts and recorded sounds in which three speakers read the text.

In the analysis and visualization, we used the free waveform editing software "Audacity" to list the differences from the features of the speech signal sets, such as waveform display of the speech signals of human speech and synthesized speech and checking the difference in time series data. In order to analysis we have set the labeling data for syllable data by using visualization and auditory signals.

IV. RESULTS AND CONSIDERATION

Some differences between synthesized speech and human speech are described. Compare the waveforms of "anger", "joy", and "sadness" by human and synthesized speech, with the strengths of each stage of emotion and five levels.

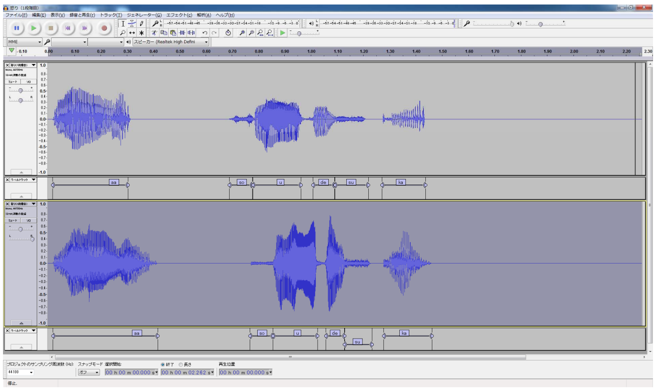

Fig. 1. Comparison of "angry" speech signals (synthetic speech)

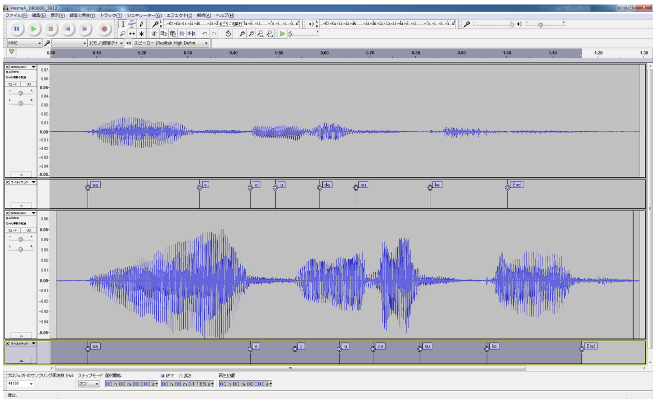

Figure. 2. Comparison of "angry" Speech Signals (Human)

The waveform shown in Fig. 1 is set to one step (0.2) and five steps (1.0) of "anger" by emotional expression of synthesized speech and reads out the text "Oh, that's right". Fig. 2 shows one step (the weakest degree) and five levels (the strongest degree) by the emotional sound "anger" of the speaker C, and similarly to Fig. 1, the waveform which reads out the text "Oh, that's right". Fig. 1 and Fig. 2 show waveforms of one level on the top and five levels on the bottom.

First of all, it is volume, but when you look at the human voice, it is obvious that the amplitude of the lower waveform is larger as a whole compared to the upper waveform. On the other hand, although it is the waveform of synthesized speech, the first two syllables "aa" both have amplitudes between -0.5 and 0.5, which shows that they are almost unchanged. However, in the latter half "sou" "u" "de" "su" four syllables, there is a difference in amplitude magnitude. From this, it turned out that by changing the parameter of emotion "anger", the difference in the size of the waveform can be obtained without changing the volume when synthesizing and reading out the voice.

Also, looking at the human voice signal in Fig. 2, it can be seen that the magnitudes of the waveforms differ between the first and fifth stages of anger. This can be presumed to be because there is a speech feature "when the human beings get angry, the volume becomes larger than usual". As mentioned earlier, since similar features were observed in synthesized speech waveforms, synthetic speech seems to reproduce this feature. The same applies to the syllable "ka" at the ending. Also, the fact that the first syllable "aa" is common to both human and synthesized speech means that it becomes longer as the degree of anger is strengthened. Especially the characteristic of human being was remarkable.

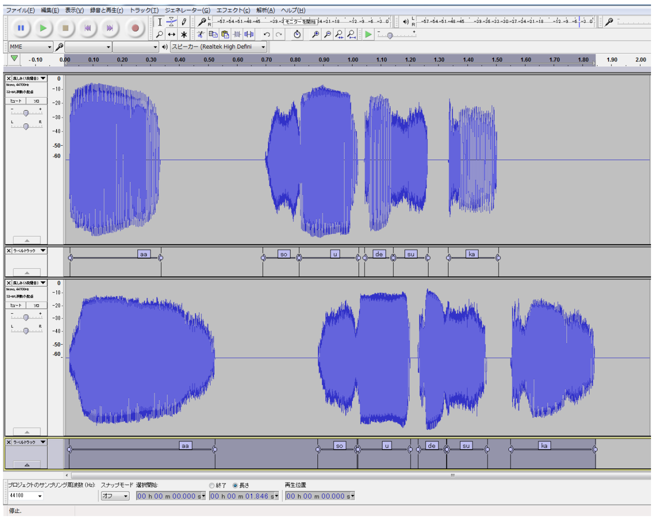

Fig. 3. Comparison of "sorrow" speech signals (synthetic speech)

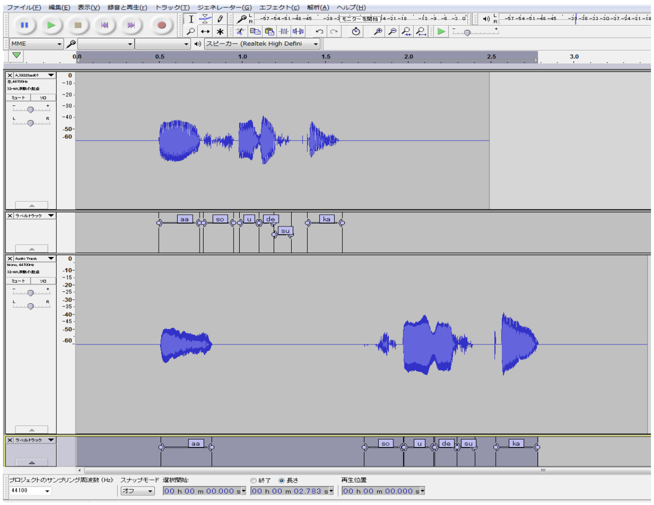

Figure. 4. Comparison of "sorrow" Speech Signals (Human)

The waveform of Fig. 3 is set to one step (0.2) and five steps (1.0) of "sorrow" by emotional expression of synthesized speech similarly to "anger", and reads out the text "Oh, that's right". Fig. 4 shows the waveforms in which the text of "Oh, that's right" are expressed as one step (the weakest degree) and five steps (the strongest degree) by the emotional sound "sorrow" of the speaker C. Fig. 3 and Fig. 4 show waveforms of one level on the top and five levels on the bottom. When comparing the interval between "aa" and "so" of synthesized speech and human speech, a remarkable difference was observed. First, it is understood that the interval of synthesized speech does not vary much between 1st step and 5th step, but the human voice is greatly different between 1st step and 5th step. The five-step interval became longer waveform.

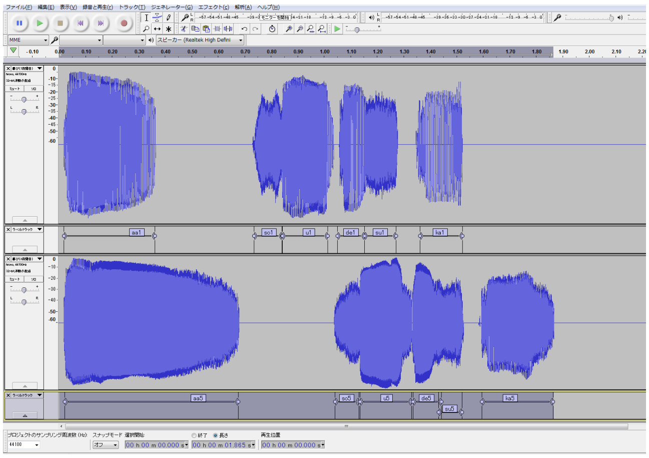

Fig. 5. Comparison of "joy" speech signals (synthetic speech)

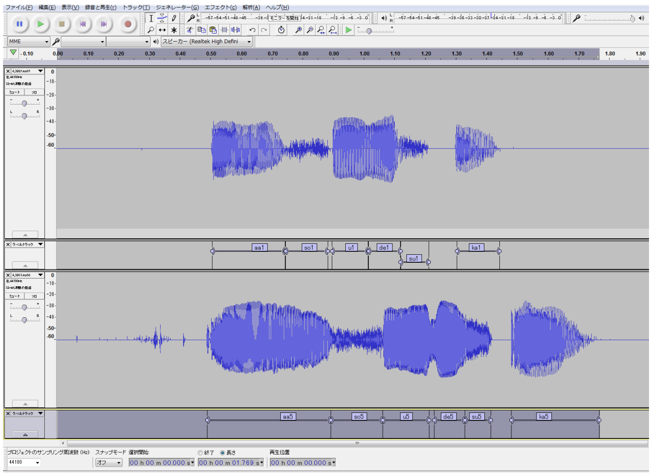

Figure.6. Comparison of "joy" Speech Signals (Human)

The waveform of Fig. 5 is set to one step (0.2) and five steps (1.0) of "pleasure" in the emotional expression of synthesized speech and reads out the text "Oh, that's right". Fig. 6 shows a waveform that expresses one stage (the weakest degree) and five levels (the strongest degree) by the emotional sound "pleasure" of the speaker C and reads the text of "Oh, that's right" as in the figure is there. Fig 5 and Fig. 6 show the waveforms of the upper one and the lower five, respectively. It is common in the waveforms of synthesized speech and human speech that the waveform width is larger in 5 stages and the length of "aa" is longer. Especially, it is understood that the synthesized speech has a length of "aa" in five stages as much as one step.

From the analysis of the above three emotions, it can be inferred that characteristics such as volume, sound height, sound and sound interval change, because significant differences are seen by changing emotional parameters. More detailed analysis results will be stated at the time of entry.

V. Conclusion

In this study, we investigate the differences in speech features in emotional expression of human speech and synthesized speech. Actual differences in waveform data in four emotions "joy", "anger", "sadness", and "calm" were investigated using each sound data. The purpose of this research was to evaluate the latest speech synthesis software and analyzed it focusing on the difference from human emotional expression. Because it was thought that emotions are peculiar to humans, it is innovative technology to reproduce the human "emotions" with synthetic speech, and there is still room for further development. As a result of this study, it was found that synthesized speech simulates human emotional speech and changes speech features (volume, sound height, intervals, etc.). Based on the difference revealed in this research, we can reproduce human emotional speech enrichedly by improving speech synthesis software in the future, and it can be expected that the utilization of synthesized speech will increase.

References

[1] S. Hirai, M. Imono, S. Tsuchiya, H. Watabe "Speech Synthesis Method with Emotion Considering Acoustic Features of Voice" 14th Information Science and Technology Forum Vol. 2, pp.289-290, 2015.

[2] T. Moriyama, S. Mori, S. Ozawa "Emotional Speech Synthesis Using Partial Spaces of Prosody" Transactions of Information Processing Society of Japan, Vol.50, No. 3, pp.1181-1191, 2009.

[3] M. Ochi, H. Kuroda, "Analysis of speech change including emotion by considering accent type and mora" 28th Annual Conference of the Society for Artificial Intelligence, 2014.

[4] M. Shuzo, Y. Yamamoto, M. Shimura, F. Kadoma, S. Mitsuyoshi, I. Yamada "Construction of Natural Voice Database for Analysis of Emotion and Feeling," Transactions of Information Processing Society, Vol. 3, pp.1185-1194, 2011.

[5] A. Masato "Recognition of emotional information contained in speech: how to express emotional space", Journal of Japan Acoustical Society, Vol.66, No.8, pp.393-398, 2010.

[6] K. Sakuraba, S. Imaizumi, K. Kakei "Sensibility information put on "Pikachu", Speech Research, Vol. 8, No. 1, pp.77 - 84, 2004.