2019 RISP International Workshop on Nonlinear Circuits, Communications and Signal Processing (NCSP2019)
Honolulu, Hawaii, USA, March 4-7, 2019

Study on Relationship between Degree of Emphasis and Acoustic Feature for Synthesizing Emphasized Speech

Yasuhiro Ohtani and Masato Akagi

Graduate School of Advanced Science and Technology
Japan Advanced Institute of Science and Technology
1-1 Asahidai, Nomi, Ishikawa 923–1292 Japan
E-mail: [yasu6triplejump, akagi]@jaist.ac.jp

Abstract

Humans can perceive not only presence/absence of emphasis but also degrees of emphasis from actual emphasized speech. However, humans cannot fully do from synthesized speech. This paper focused on two properties of Fundamental frequency (F0) contours: amount of decay from the accent nucleus and variation between each accent nucleus, and hypothesized that the two properties of F0 contours are important for synthesizing emphasized speech. To discuss this hypothesis, this paper clarified relationships between degrees of emphasis and F0 contours. To clarify relationships, it was necessary to compare relationships for each stimulus. To compare relationships, it was necessary to know the degree of emphasis of each stimulus and analyze variations of F0 contours. A listening test was carried out to obtain the degrees of emphasis of stimulus. A value which is frequency at the barycentric point of the vowel was extracted from F0 contours to analyze the variation of F0 contour. From these results, we had two findings; degree of emphasis is increasing when amount of decay from accent nucleus to next mora is increasing, and the variation of accent nuclei is different with/without emphasis. The experiment was carried out to evaluate hypothesis. Synthesized stimuli from non-emphasized voice by varying amount of decay and variation of accent nuclei are used for the experiment. The results showed that the participants of the experiment can perceive emphasis with degrees from the synthesized stimuli. This result clarified the relationships between presence/absence of emphasis and two findings. In addition the hypothesis is important for synthesizing emphasized speech which convey presence/absence of emphasis.

1. Introduction

Speech communication of humans is rich in expressions. It includes not only linguistic information but also para and non-linguistic information. However, synthesized speech cannot fully convey para-linguistic information yet. Emphasis is one of the important elements of para-linguistic information to convey intentions of speech contents. According to phonetics, speech emphasis is a part that makes differences with other parts. The emphasized part is made outstanding from other parts. Speech emphasis is realized by making voice size, length, and prominence[1]. Humans can perceive not only presence/absence of emphasis but also degrees of emphasis from actual emphasized speech. However, humans cannot fully do from synthesized speech. Human can perceive strength of the speaker's intention from degrees of emphasis. Perceiving strength of intention from synthesized speech make speech communication using synthesized speech rich. Thus, it is necessary to synthesize emphasized speech that can convey degrees of emphasis to listeners.

Pitch is the most important prosodic attribute for perceiving para-linguistic information [2]. Fundamental frequency (F0) contours are one of the acoustic features related to pitch. Many previous studies have synthesized emphasized speech focusing on F0 contours [3]. In Japanese, pitch decreases rapidly from accent nucleus to next mora. Mora is the relative length of the sound which becomes the unit of strength and intonation. Accent nucleus is a mora just before the pitch decreasing. In addition, the peak of pitch decreases after the decreasing of the accent nucleus (catathesis or downstep)[4]. In the case of emphasized speech, this phenomenon is hindered. Thus, we focus on the two features of F0 contours which related the variation of pitch in Japanese: decreasing of F0 from the accent nucleus and difference between accent nucleus of emphasized part and other accent nuclei in sentences. Also, we hypothesize that these features are important for synthesizing emphasized speech.

This study aims to clarify relationships between these features and degree of emphasis in order to evaluate the hypothesis. To clarify relationships, degrees of emphasis from the recorded voices are evaluated. In addition, decreasing of F0 contours from the accent nucleus of emphasized word to next mora are analyzed. F0 contours are expressed by using F0 at the barycentric point of the vowel (point pitch) [5]. Features in F0 contours are represented by calculating the difference of point pitch.

2. Speech stimuli

In order to discuss relationships, it is necessary to know the degree of emphasis of each stimulus. In addition, it is necessary to know segment information of speech stimulus in order to extract point pitch from the F0 contours. Thus, a listening test is carried out to evaluate the degrees of emphasis of stimuli.

2.1 Experiment 1: listening test for degree of emphasis

A listening test was carried out to evaluate the degrees of emphasis of the stimuli. This test is named as the experiment 1. Tokyo dialect utterances were used as the stimuli. Participant in the recording is one speaker. Stimulus were three-word sentences and the words were noun consisted of 4-mora. The pitch accent of each word was unaccented, initial accented, medial accented or final accented word[4]. Each stimulus was recorded with instruction of emphasizing one of the three noun words or non-emphasizing all words. Therefore, one stimulus of the same utterance content was uttered in four manners (without emphasis, emphasizing first word, second word, or third word). Ten native Japanese students with normal hearing were participated in the experiment 1. The listening test was performed in a soundproof-room. The stimuli were randomly presented to the listener via a headphone. They were asked to evaluate not only presence/absence of emphasis but also degrees of emphasis in four steps. Degrees of emphasis were averaged in each stimulus.

2.2 Segmentation

It was necessary to segment speech stimulus to obtain the point pitch. Speech stimuli were segmented manually by using result of analysis obtained by using Praat. Segmentation was based on the knowledge of spectrogram. Speech stimuli were segmented into vowel, voiced consonant, and unvoiced consonant portions.

3. Comparisons of relationships between degree of emphasis and features

In order to clarify the relationships, it is necessary to analyze the two features and discuss the relationships.

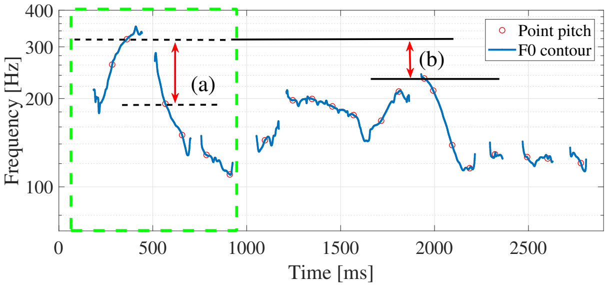

Figure 1: Amount of decay: (a)from accent nucleus to next mora and (b) between accent nucleus in sentences.

Point pitch, which was the value of F0 at the time of energy barycentric point, was extracted from F0 contour to analyze the two features. The F0 contour of each voice is obtained by using STRAIGHT(V40_005b)[6] with frame length 40 ms, frame shift 1 ms and boundary of F0: 80 Hz - 600 Hz. Point pitch is extracted from the F0 contours. This study focuses on two amounts of decays: variation from accent nucleus to next mora and difference between accent nuclei in sentences. Figure 1 shows the F0 contour of emphasized voice and two features. The green dashed box indicates F0 contour of the emphasized word. The red circles indicate point pitches. The arrow (a) in Figure 1 indicates the variation from accent nucleus to next mora. The arrow (b) in Figure 1 indicates difference between accent nuclei in sentences. Then, relationship between degree of emphasis and features is compared to clarify the relationships.

3.1 Decreasing from accent nucleus to next mora

Figure 2(a) shows the relationship between degrees of emphasis and amount of decay in medial accented words. Amount of decay is decrease of point pitch from accent nucleus to next mora. Figure 2(b) shows the relationship between degrees of emphasis and amount of growth in medial accented words. The degree of emphasis is a value derived from the result of the listening test. Amount of growth is increase of point pitch from mora before accent nucleus to the accent nucleus. From each point pitch, the variations were calculated (Eq. 1 and 2).

$$Amount\ \ of\ \ decay = \log x_2 - \log x_1 \tag{1}$$
$$Amount\ \ of\ \ growth = \log x_2 - \log x_3 \tag{2}$$

x_1 is the value of point pitch at next mora of accent nucleus. x_2 is the value of point pitch at accent nucleus. x_3 is the value of point pitch at mora of one before accent nucleus. Figure 2(a) indicates that degrees of emphasis increase with amount of decay increase. On the other

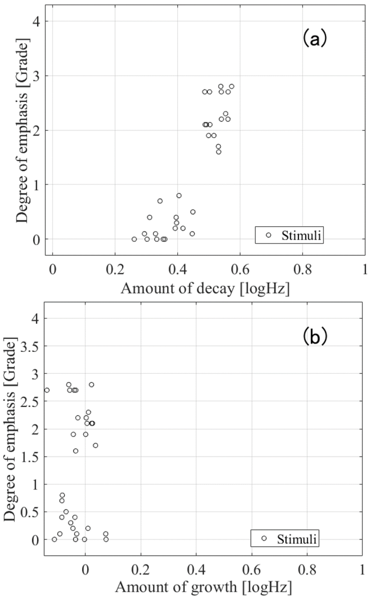

Figure 2: Relationship between degree of emphasis and variation before and after accent nucleus: (a) Relationship between degree of emphasis and amount of decay, (b) Relationship between degree of emphasis and amount of growth.

hand, Figure 2(b) indicates that the amount of growth does not change even if the degree of emphasis varies. In addition, point pitches do not change regardless of the presence/absence of emphasis or the change in degree of emphasis. From the two results, it is considered that point pitch at accent nucleus of emphasized word increases as the degree of emphasis increases. In addition, point pitch at mora of one before accent nucleus increase according to increasing of point pitch at accent nucleus. Therefore, the presence or absence of emphasis changes when the amount of decay change.

3.2 Difference of accent nuclei in sentences

Figures 3(a), (b) and (c) shows the relationship between the degrees of emphasis and amount of decay. Amount of decay is the difference between first and second word, difference between second and third word or difference between first and third word respectively. The emphasized word is medial accented word. Amount of decay is calculated from point pitches of accent nuclei (Eq. 3, 4 and 5).

$$Amount\ \ of\ \ decay(a) = \log a_1 - \log a_2 \tag{3}$$

$$Amount\ \ of\ \ decay(b) = \log a_2 - \log a_3 \tag{4}$$

$$Amount\ \ of\ \ decay(c) = \log a_1 - \log a_3 \tag{5}$$

a_1 is the value of point pitch at the accent nucleus of the first word. a_2 is that of the second word. a_3 is that of the third word. When the word is unaccented word, Point pitch with the highest value is selected as point pitch of accent nucleus. Black asterisk indicates in the case of without emphasis. The red circle, the blue square, the green triangle indicate the case that the first word, the second word or the third word which are emphasized respectively. Figure 3(a) shows that the amount of decay (a) is less than 0 when the second word was emphasized. Thus, the F0 peak of second word was higher than the F0 peak of the first word when the second word was emphasized. Figure 3(b) illustrates that the amount of decay (b) is more increasing when the second word was emphasized. On the other hand, amount of decay (b) is less than 0 when the second word was emphasized. Figure 3(c) shows that the amount of decay (c) is almost 0 when the third word was emphasized. Thus, the F0 peak of the third word is same as the F0 peak of the first word. From three results, difference of point pitches of accent nuclei between emphasized word and next word is more increasing. In addition, difference between emphasized word and before word is more decreasing. However, amount of decrease does not change even if degree of the emphasis changed. Therefore, the presence or absence of emphasis changes when the amount of decay change.

3.3 Experiment 2: evaluation of hypothesis

A listening test is conducted to evaluate whether people can perceive emphasis from stimuli or not, when modifying the F0 contour according to the two findings: amount of decay on emphasized word and difference of accent nucleus in sentences. This test is named as the experiment 2. Three kinds of stimuli are used in the experiment 2. The first stimuli are synthesized by using F0 contours analyzed with STRAIGHT. The second stimuli are synthesized by using F0 contour obtained from the point pitch. These are used to clarify whether the quality of synthesized speech using F0 contour obtained from point pitch is suitable[7]. The third stimuli are synthesized by using F0 contours obtained from the point pitch, which are manipulated according to the two findings. The experiment 2 was carried out under the same procedure as the experiment 1. The evaluation results are averaged in each stimulus. As a result of the experiment 2, emphasis was perceived from a word manipulated according to findings with 88.9 percent chance in the third stimuli. On the other hand, emphasis was perceived from the word manipulated in the third stimulus with 22.2 and 16.6 percent chance respective in the first and the second stimuli. From the experiment 2, we clarify that participants of the experiment 2 can perceive emphasis from synthesized speeches which were synthesized by us-

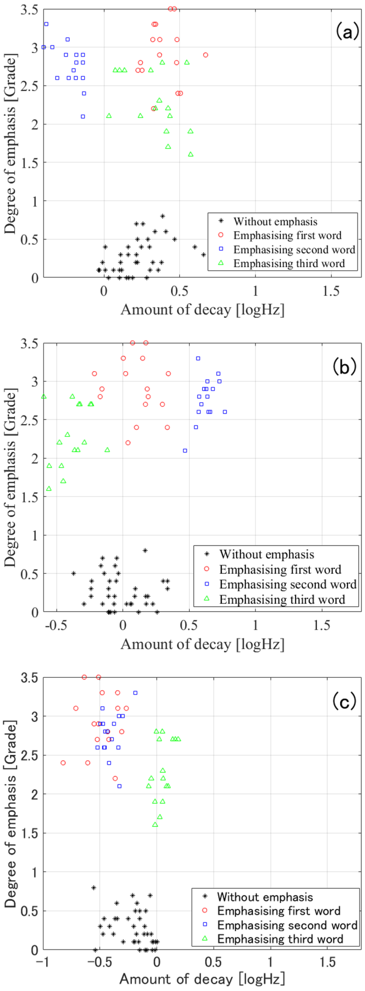

Figure 3: Relationship between degree of emphasis and amount of decay: (a) The variation between first and second word, (b) The variation between second and third word and (c) The variation between first and third word.

ing manipulated F0 contour. Therefore, the hypothesis is important for synthesizing emphasized speech.

4. Conclusions

In this paper, we hypothesized that the two features of F0 contours (amount of decay from the accent nucleus, variation between each accent nucleus) are important for synthesizing emphasized speech. This study aimed to clarify relationships between degrees of emphasis and two features of F0 contours in emphasized words to discuss this hypothesis. To clarify relationships, we compare relationships for each stimulus and evaluate the relationships. To compare relationships for each stimulus, we carry out experiment 1 and analyze variation of two features for each stimulus by using point pitch. We carried out experiment 2 to evaluate hypothesis. From the experiment 2 result, it can be concluded that the relationships between presence/absence of emphasis and two findings are clarified and the hypothesis is important for synthesizing emphasized speech. In addition, Figure 2 indicated that there are variations of degree of emphasis when amount of decay from accent nucleus varies. Therefore, it is considered that the variations may affect degrees of emphasis in human perception and relationship between degrees of emphasis and F0 contours may clarify by modeling variations.

Acknowledgment

This work was supported by SECOM Science and Technology foundation and JST-Mirai Program (JPMJMI18D1).

References

[1] Tanaka, H., et al.: Gendaigengogakujiten, Seibido, 1988, (in Japanese).

[2] Ladd, R., K. Silverman, F. Tolkmitt, G. Bergmann and Scherer, R.: Evidence for the independent function of intonation contour type, voice quality, and F0 range in signaling speaker affect, JASA, vol. 78, No. 2, pp. 435–444, 1985.

[3] Shirai, K., Iwata, K.: Rrosodic Rules for Speech Synthesis Representing Word Emphasis, IEICE vol. 70, No. 5, pp. 861-–821, 1987, (in Japanese).

[4] Takubo, Y., maekawa, K., kubozono. H., Honda, K., Shirai, K., Nakagawa, S.: gengonokagaku 2 onsei, Iwanamishoten, pp. 44–46, 1998, (in Japanese).

[5] Hashimoto, S.: Several Feature of Japanese Word Accent, IEICE D, vol. 56, No. 11, pp. 654–661, 1973, (in Japanese).

[6] Kawahara, H., Masuda-Katsuse, I., De Cheveigne, A.: Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds, Journal of Speech Communication, vol. 27, No. 3, pp. 187-207, 1999.

[7] M. Abe, M., Sato, H.: Two-stage F_0 control model using syllable based F_0 units, Acoustical Science and Technology, vol. 49, No. 10, pp. 682–690, 1993, (in Japanese).