Feature extraction of speech with contrastive emphasis
focused on accent nucleus

1710042 Otani Yasuhiro

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. 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. Fundamental frequency (F0) contours are one of the acoustic features elated to pitch. Many previous studies have synthesized emphasized speech focusing on F0 contours. 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 down step). 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). Features in F0 contours are represented by calculating the difference of point pitch. 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. A listening test was carried out to evaluate whether the stimuli are useful for analysis and the degrees of emphasis of the stimuli. This test is named as the experiment 1. Tokyo dialect utterances were used as the stimuli. Each stimulus was recorded with instruction of emphasizing one of the three noun words or non-emphasizing all words. 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 (1 to 4). Degrees of emphasis were averaged in each stimulus.

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.

In order to clarify the relationships, it is necessary to analyze the two features and discuss the relationships. 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) 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. Then, relationship between degree of emphasis and features is compared to clarify the relationships. In order to discuss the relationship between degrees of emphasis and amount of decay, Amount of decay and Amount of growth are calculated. Amount of decay is decrease of point pitch from accent nucleus to next mora. Amount of growth is increase of point pitch from mora before accent nucleus to the accent nucleus. The degree of emphasis is a value derived from the result of the listening test. From the result, degrees of emphasis increase with amount of decay increase. On the other hand, 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 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.

In order to discuss the relationship between the degrees of emphasis and amount of decay, Amount of decay is calculated from point pitches of accent nuclei. 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. When the word is unaccented word, point pitch with the highest value is selected as point pitch of accent nucleus. From the 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. 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. 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. From the experiment 2, we clarify that participants of the experiment 2 can perceive emphasis from synthesized speeches which were synthesized by using manipulated F0 contour. Therefore, the hypothesis is important for synthesizing emphasized speech. In addition, 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.