Generation and perception of F_0 markedness in conversational speech with adverbs expressing degrees

Takumi Yamashita, Yoshinori Sagisaka

Graduate School of Global Information and Telecommunication Studies,
Waseda University, Nishi-Waseda 1-3-10, Shinjuku-ku, Tokyo 169-0051 Japan
takumi@toki.waseda.jp, sagisaka@giti.waseda.ac.jp

Abstract

Aiming at natural F_0 control for conversational speech synthesis, F_0 characteristics are analyzed from both generation and perception viewpoints. By systematically designing conversational situations and utterances with adverb phrases expressing different degree of markedness, their F_0 characteristics are compared. The comparison shows the consistent F_0 control dependencies not only on adverbs themselves but also on the attribute of neighboring adjective phrases. Strong positive/negative correlation is observed between the markedness of adverbs and F_0 height when an adjective phrase with a positive/negative image is followed to the current adverb phrase. These consistencies have been perceptually confirmed by naturalness evaluation tests using the same two-phrase samples with different F_0 heights. These results indicate the possibility of F_0 control for natural conversational speech using lexical markdness information and adjacent word attributes.

1. Introduction

Recently corpus-based approach has been successfully applied to the prosody control [1]-[4]. In reading style Japanese speech, as F_0 contours can be predicted quite well by phrase dependency structure, the phrase length, position in a phrase and phrase accent type, reasonable prosodic quality is obtained for Japanese text-to-speech. Research efforts have been mainly devoted to the extraction of these parameters from text. As synthetic speech becomes popular, it started to be used in other applications where reading style speech is no more adequate. We can only synthesize speech with a reading style that is quite unnatural in many cases when conversational speech is expected.

Real human conversation speech is controlled by a lot of factors, and there is little knowledge on what factors affect how in conversational speech generation. From the engineering viewpoint, it is quite difficult to synthesize conversational speech, as not only many of control factors are unknown but also it looks quite difficult to automatically extract them from conversation contexts even if we know them. In order to synthesize more natural conversation speech, we have to analyze and model control mechanisms with sufficient generality. Under these circumstances, we thought that word intrinsic prosodic information can be one of the easiest and useful information for natural conversational speech generation.

In this paper, as a first step towards conversational speech synthesis, we have analyzed F_0 control characteristics based on word intrinsic markedness. If there exist any consistent relation between F_0 control characteristics and word intrinsic markedness, they can be added to the dictionary as lexical attributes and easily referred in conversational speech generation process. As a pilot study, we selected adverb phrases expressing different degree of markedness and measured their F_0 differences. In Section 2, the selection of utterances with adverb phrases is explained. The collection of speech utterances under controlled conversational contexts and F_0 measurement results are described in Section3. In Section 4, perceptual naturalness evaluation results are presented in relation to the F_0 control characteristics observed in the previous section. Finally, the coincidence of generation and perception characteristics is summarized and the usefulness of word intrinsic markedness for conversational speech synthesis is confirmed.

2. Design of conversation utterances with adverbs expressing degrees

To quantitatively analyze the relationship between F_0 and word intrinsic markedness, we selected adverbs expressing degrees since they can be quantified by their subjective magnitude that can be interpreted as their intrinsic markedness. We chose very popular seven Japanese adverbs of four-mora length with the same accent type (flat accent) and made

shortest phrase utterances by appending adjective phrases after them. The adverbs used in the experiment are listed in Table 1 in the order of subjective magnitude with corresponding English expressions. Please notice that Japanese phrase order (adverb phrase + adjective phrase) is reverse of English one. To eliminate the F_0 control differences resulting from phrase dependency structure, phrase length, position in a phrase and phrase accent types, we chose these two-phrase utterances that are frequently used in real conversations. Except the differences resulting from negative expressions required by the last two adverbs, all utterances of these adverb phrases were compared in the same condition followed by the same adjective phrases.

For adjectives in the following phrases, two types of five categories were selected to express either positive or negative image. These ten adjectives are listed in Table 2 with corresponding English expressions. Most of these adjectives have penultimate accent type except two (ki'rei and busa'iku). In total, we used forty-five different utterances for generation experiments. They consist of the combination of six adverb phrases followed by five positive image adjectives and the three adverb phrases ("very", "normally" and "not so much") followed by five negative image adjectives. For perception experiments of naturalness evaluation presented in Section 4, we used sixty phrases. They consist of seven adverb phrases followed by five positive image adjectives and five adverb phrases ("extremely", "very", "normally", "not so much" and "not at all") followed by five negative image adjectives.

Table 1 Adverbs expressing degrees selected for the utterances in the experiments

Japanese adverbs	Corresponding English expressions
hijooni	extremely
sootoo	very
wariai	quite
sokosoko	relatively
futsuuni	normally
annmari	not so much
zenzen*	not at all*

*only used in perception experiment

Table 2 Adjectives used for the following phrases of the utterances in the experiments

Positive-image adjectives		Negative-image adjectives	
Japanese	Corresponding English expression	Japanese	Corresponding English Expression
kirei	beautiful, clean	kitanai	dirty
umai	delicious	mazui	unsavory
kawaii	charming	busaiku	ugly
yasasii	mild	kibisii	strict
omosiroi	interesting	tsumaranai	boring

3. Collection and analysis of conversational speech with adverbs

3.1. Conversation context specification for speech collection

To collect as much as natural conversational speech, we asked the subject speakers to utter them naturally as a response to casual questions that are familiar to them in everyday conversations. For example, when recording an utterance including the adverb /hijooni/ (extremely) followed by the adjective /umai/ (delicious), they were asked to reply to the question /ajiwa doo ?/ (How does it taste ?). Before asking this question, subjects were asked to imagine the situation where they have taken something tasty, and show their satisfaction with the phrase /hijooni umai/ (It's extremely delicious) without any particular intentional emphasis.

In total, forty-five different two-phrase combinations were pronounced by four subjects and recorded in quiet condition. In addition, as control data, these samples were uttered in reading style after the recording of conversational speech. To confirm the subjective lexical markedness of adverbs used in the experiments, we asked all subjects to put score ranging from 1 (unmarked) to 10 (marked) to these adverbs.

3.2. Analysis results and discussions

In reading style Japanese speech, it is well known that all (adverb phrase + adjective phrase) combinations mentioned in Section 2 have the same F_0 contour except local discrepancies resulting from micro-prosody. We extracted the F_0 pattern at vowel centers of reading style speech. Figure 1 shows the average F_0 values of reading style speech at the adverb phrase positions when adjectives with a positive image follow, which shows the equal F_0 height control for all adverbs. In the Figure 1, the vertical axis shows the average F_0 of 2^{nd}, 3^{rd} and 4^{th} vowel center positions and the horizontal axis shows the markedness of adverbs in ascending order from left to right.

The result for conversational speech is shown in Figure 2. In Figure 2, the vertical axis shows the F_0 average differences (in log scale) of same phrases between the reading style speech and the conversational one at each adverb position when positive-image adjectives follow. As shown in the figure, F_0 contour becomes consistently higher in proportion to the increase of markedness of adverbs. The correlation between the score expressing the subjective markedness of adverbs and F_0 height ranged from 0.76 to 0.91 among subject speakers. The average value over subject speakers was 0.85.

Figure 3 shows the contrast of the effects of following adjectives attribute, positive image versus negative image, on

F_0 height in adverbs ("very", "normally" and "not so much"). As shown in the figure, high correlations were observed between the markedness of adverbs and F_0 differences (between the reading style and the conversational speech) in both adjective groups but their sign was opposite. Correlations between the score expressing the subjective markedness of adverbs and F_0 height for positive/negative adjective groups ranged from 0.94 to 1.00 and from -1.00 to -0.66 respectively. These high correlations with different signs suggest the possibility of F_0 control of conversational speech with word intrinsic markedness and neighboring attributes.

4. Perception on the naturalness of adverb phrases with different F_0 heights

4.1. Perceptual evaluation test

To confirm the naturalness of adverb phrases with different F_0 height, we have designed a perceptual evaluation test. As fully synthesized speech sometimes spoils judgment accuracy, we used natural speech with different F_0 heights. In total 720 speech samples were recorded for sixty combinations (adverb phrase + adjective phrase) with twelve different F_0 heights. These twelve speech samples were uttered by a male subject adjusting the maximum F_0 of the utterances (F_0 at the second mora position in the adverbs) to a given pure tone signal. Pure tone signals range from G sound (98.00Hz) to F# sound (185.00Hz) with a semitone interval. Through the analysis of these samples, we have confirmed that all these twelve different F_0 samples of each phrase combination have the same F_0 contours as an accent component of adverb phrases and that the other prosodic differences are little.

Ten Japanese subjects with normal auditory ability were asked to evaluate twelve speech stimuli using five categories from 1 (very bad) to 5 (very good) corresponding to naturalness rating as scores of each sample. Icons of the twelve speech stimuli of the same phrase were displayed on the console screen for the subjects to listen freely and they were asked to give a score.

4.2. Results and discussions

For each speech stimulus, the average score over subjects was calculated. For positive or negative adjective groups, the results are shown in Table 4 (a) and (b) respectively. As seen in these figures, we listed maximum F_0 values at adverb position in descending order from top to bottom columns and the markedness of adverbs in ascending order from left to right rows. The darkness of each cell in the Table corresponds to the naturalness score. (The darker, the preferred.)

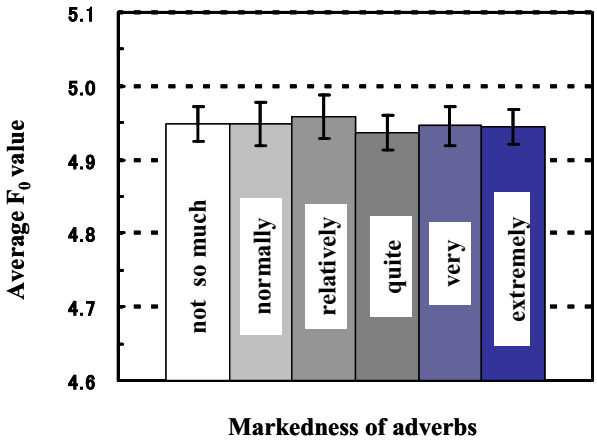

Figure 1 Average F_0 at adverbs with different markedness in reading style speech

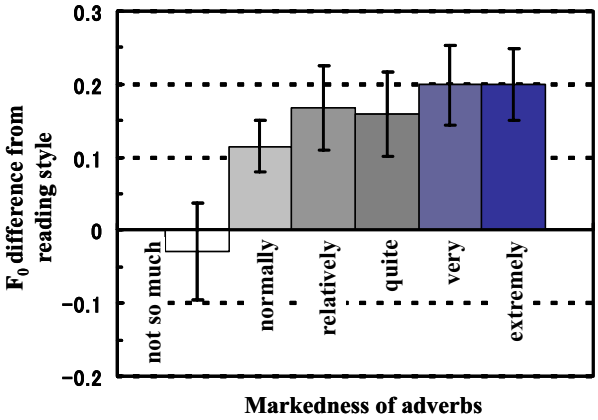

Figure 2 The increase of F_0 average difference between reading and conversation in proportion to the increase of markedness of adverbs when positive adjectives follow

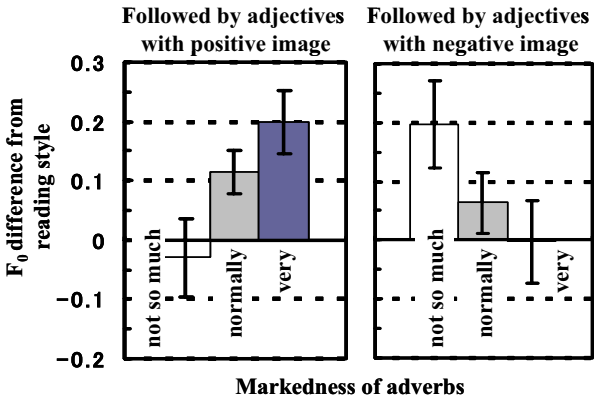

Figure 3 The effect of following adjective attributes on the F_0 differences between reading and conversation

Table 4 Average naturalness scores for the uttrences with adverb phrases of different F_0 heights
(a) Followed by adjectives with positive image

F_0 maximum at adverbs [Hz]	Markedness of adverbs						
	zenzen (not at all)	annmari (not so much)	futsuuni (normally)	sokosoko (relatively)	wariai (quite)	sootoo (very)	hijooni (extremely)
185.00 (F#)	**1.56**	**1.42**	**1.70**	**1.96**	**2.26**	**3.48**	**3.78**
174.61 (F)	**1.76**	**1.62**	**2.14**	**2.48**	**2.60**	**3.74**	**4.10**
164.81 (E)	**2.10**	**2.00**	**2.62**	**2.94**	**3.18**	**4.00**	**4.16**
155.56 (D#)	**2.36**	**2.56**	**3.20**	**3.48**	**3.82**	**3.98**	**4.06**
146.83 (D)	**2.84**	**2.88**	**3.52**	**3.72**	**4.04**	**3.84**	**3.98**
138.59 (C#)	**3.20**	**3.16**	**3.96**	**4.14**	**4.14**	**3.56**	**3.50**
130.81 (C)	**3.48**	**3.50**	**4.12**	**4.18**	**4.00**	**3.28**	**3.12**
123.74 (B)	**3.80**	**3.90**	**4.10**	**3.98**	**3.64**	**2.94**	**2.70**
116.54 (A#)	**3.98**	**4.08**	**3.66**	**3.60**	**3.30**	**2.50**	**2.38**
110.00 (A)	**4.34**	**3.92**	**3.12**	**3.00**	**2.66**	**2.20**	**1.94**
103.83 (G#)	**4.34**	**3.72**	**2.56**	**2.56**	**2.36**	**1.84**	**1.64**
98.00 (G)	**4.18**	**3.54**	**2.30**	**2.32**	**2.12**	**1.70**	**1.54**

(b) Followed by adjectives with negative image

F_0 maximum at adverbs [Hz]	Markedness of adverbs				
	zenzen (not at all)	annmari (not so much)	futsuuni (normally)	sootoo (very)	hijooni (extremely)
185.00 (F#)	**2.80**	**2.32**	**1.52**	**1.96**	**2.20**
174.61 (F)	**3.12**	**2.72**	**1.78**	**2.26**	**2.32**
164.81 (E)	**3.46**	**3.22**	**2.24**	**2.46**	**2.46**
155.56 (D#)	**3.50**	**3.56**	**2.82**	**2.62**	**2.52**
146.83 (D)	**3.64**	**3.72**	**3.22**	**2.88**	**2.84**
138.59 (C#)	**3.66**	**3.86**	**3.66**	**3.22**	**3.26**
130.81 (C)	**3.52**	**3.80**	**3.92**	**3.64**	**3.50**
123.74 (B)	**3.10**	**3.60**	**4.14**	**3.80**	**3.84**
116.54 (A#)	**2.74**	**3.10**	**4.06**	**4.04**	**3.92**
110.00 (A)	**2.38**	**2.54**	**3.76**	**4.12**	**4.14**
103.83 (G#)	**2.18**	**2.34**	**3.54**	**4.22**	**4.12**
98.00 (G)	**1.94**	**2.08**	**3.44**	**4.04**	**3.94**

As shown in Table 4 (a) and (b), it is found that the highest naturalness score is consistently changing over these samples. Table 4 (a) shows that higher naturalness score is assigned to speech with higher F_0 as the markedness of adverbs increases when the adjectives are positive. The totally opposite perceptual evaluation is given, as seen in Table 4 (b). These consistent but neighboring adjective dependent perceptual characteristics nicely match to the results of generation shown in Figure 3. These results confirmed the effectiveness of the F_0 control in conversational speech by word intrinsic markedness of adverbs and the attributes of neighboring adjectives.

5. Summaries

The effect of word intrinsic markedness on F_0 control characteristics were analyzed from both generation and perception viewpoints. Systematic control of adverb selection and conversational contexts show the consistent effects of markedness of adverbs on F_0 height. It is found that this effect is contextually dependent on the parity of the following adjective; positive versus negative. Furthermore, exactly the same characteristics are found in the perceptual experiments on naturalness evaluation for speech with different F_0 heights. These consistent results indicate the possibility of F_0 control for natural conversational speech using lexical markedness information and adjacent word. The generalization of this phenomenon to other utterance samples and conversational contexts is under planning aiming at effective use of word information and interaction with contexts for the synthesis of natural conversation speech.

6. References

[1] Sagisaka Y.: “On the prediction of global F_0 shape for Japanese text-to-speech” , Proc. ICASSP, pp.325-328, 1990

[2] Riley M.D.: “Tree-based modeling of segmental durations" p.265-274 in “Talking Machines” edited by G.Bailly et al North-Holland, 1992

[3] C. Traber: “SVOX: The implementation of a Text-to-Speech System for German” , TIK-Schriftenreihe Nr 7, 1995

[4] Tokuda, K., Masuko, T., Miyazaki, N., and Kobayashi, T.: “Hidden Markov models based on multispace probability distribution for pitch pattern modeling”, Proc. ICASSP, pp.229-232, 1999