

# Speech Synthesis with Attitude

Yoshinori Sagisaka, Takumi Yamashita and Yoko Kokenawa

Global Information and Telecommunication Institute, Waseda University

sagisaka@giti.waseda.ac.jp, takumi@toki.waseda.jp, Yoko.Kokenawa@toki.waseda.jp

## Abstract

$F_0$  characteristics were analyzed and modeled for the output of speech with natural prosody in communication systems. Lexicons were selected to express speaker's attitude during the human speech generation process. We modeled the prosody using information of constituent lexicons expressing attitude and markedness. Motivated by preliminary observations of prosodic variations in conversational speech,  $F_0$  characteristics were quantitatively analyzed using simple phrases consisting of adjectives expressing positive or negative attitude and adverbs expressing different degrees of markedness. Strong positive/negative correlations were observed between the markedness of adverbs and  $F_0$  height when an adjective phrase with a positive/negative attitude follows the current adverb. These consistencies have been perceptually confirmed by naturalness evaluation tests. Finally,  $F_0$  control is modeled using lexical information expressing positive or negative attitude and markedness.

## 1. Introduction

For reading-style speech synthesis, from text input, a considerable amount of research has been successfully carried out to optimize the controls using statistical techniques, and corpus-based approaches have been successfully applied to prosody control [1]-[4]. As the prosody of reading-style speech is largely characterized by linguistic information extracted from the input text, e.g., phrase dependency structure, phrase length, phrase position and phrase accent type, reasonable quality of prosody can be obtained for its output. The corpus-based approach is quite effective for problems where the control mechanism is known and factors can be listed up. However, as synthetic speech becomes popular, it is starting to be used in other applications where reading-style speech is no longer adequate. In particular, reading-style prosody is far from satisfactory in most outputs from Q&A systems and human-like agents such as humanoid robots.

Real human communicative speech is governed by a range of factors. There is little knowledge of what factors affect the way speech is modulated by prosody. Even in a theoretical framework, traditional linguistics and phonetics cannot provide us any useful basis for the analysis of prosodic variations in the bi-directional speech observed in daily human communications. To obtain speech output with natural prosody, we need observations of prosody variations and a fundamental understanding of underlying principles.

From a conventional modeling viewpoint, it looks quite difficult to synthesize conversational speech, as not only many of the control factors are unknown but also it is hard to

automatically specify their values using conversation contexts even if we know them. In order to synthesize more natural conversational speech, we have to analyze and model control mechanisms with sufficient generality. Through the observation of prosodic variations in conversational speech, we have found that a word intrinsic attribute by itself could be good information for natural conversational speech generation.

In this paper, first, as a preliminary study on prosodic variations, the  $F_0$  characteristics of a single utterance /n/ are qualitatively analyzed to know how  $F_0$  variations are to be characterized and represented. Next, we quantitatively analyzed  $F_0$  characteristics of simple phrases consisting of adjectives expressing positive or negative attitude and adverbs expressing different degrees of markedness. The usefulness of these observed  $F_0$  control characteristics was perceptually confirmed by subjective naturalness evaluation tests. Finally, a computational model of  $F_0$  control is proposed for conversational speech using attitude attributes and markedness of constituent lexicons.

## 2. A preliminary observation of $F_0$ heights and styles in relation to speaker's attitude

To classify prosody variations from a communicative viewpoint, we have recorded commonly used short utterances of /n/ and standardized consistent descriptions of their variations. For recording, we asked four speakers to have a casual conversation on any topic of their choice. The total recording data lasted twenty-five minutes and contains forty two /n/ tokens. We measured the  $F_0$  variations and correlated them with speakers' attitudes.

Table 1 shows our classification of  $F_0$  variations by height and dynamics. The functional meanings of the /n/ could be roughly categorized into seven groupings, which are "surprise", "request to repeat", "negative response (No)", "positive response (Yes)", "reluctance", "agreeable response" and "showing agreeable attitude". The speakers were trying to express those messages just by /n/ instead of making an actual statement. Moreover, as seen in Table 1, the speakers seem to make a use of  $F_0$  height and dynamics to show their attitudes.

When a rather negative statement follows /n/, or the speakers utter /n/ with a negative attitude, the  $F_0$  tends to be lower in general. This seems to have nothing to do with the meaning of /n/ itself; however, and simply reflects the speaker's attitudes. We found that these prosodic characteristics were highly related to the corresponding lexical forms expressing speaker's attitudes. This indicates that lexicons by themselves can provide us prosody control information as default values, e.g. positive or negative attitude is expressed in prosody that can be estimated respectively by positive or negative lexicons with their intrinsic prosodic properties.

### 3. Analysis of conversational short utterances consisting of an adjective and an adverb

As a next step towards computational modeling of  $F_0$  control, we analyzed the  $F_0$  characteristics of short utterances consisting of an adjective expressing the speaker's positive or negative attitude and an adverb expressing its degree. To eliminate the  $F_0$  control differences resulting from phrase dependency structure, phrase length, position in a phrase and phrase accent type, we chose two-phrase utterances that are frequently used in real conversations. In total, we used forty-five different utterances for  $F_0$  measurements. These utterances consisted of the combination of six adverb phrases followed by five adjectives expressing positive attitude and the three adverb phrases ("very", "normally" and "not so much") followed by five adjectives expressing negative attitude as listed in Table 2. Most of these adjectives have

Table 2 (a) Adjectives expressing speaker's attitude

Positive attitude		Negative attitude	
Japanese	Corresponding English expression	Japanese	Corresponding English Expression
kirei	beautiful, clean	kitanai	dirty
umai	delicious	mazui	unsavory
kawaii	charming	busaiku	ugly
yasasii	gentle	kibisii	strict
omoshiroi	interesting	tsumaranai	boring

(b) Adverbs expressing degrees

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

\*used in perception experiment only







penultimate accent types except two (ki'rei and busa'iku). All adverbs have four-mora length with the same accent type. Except for the differences resulting from negative expressions required by the last two adverbs, the adverb phrases were compared under the same conditions followed by the same adjective phrases.

To collect as much natural conversational speech as possible, we asked all speakers to utter the phrases naturally, as a response to casual questions that were familiar to them in everyday conversations. In addition, after the recording of conversational speech, these samples were uttered again in reading style for comparison. We also asked all subjects to score lexical markedness of these adverbs ranging from 1 (unmarked) to 10 (marked).

As expected, all reading-style speech samples showed very similar  $F_0$  contours except for local discrepancies resulting from micro-prosody and phrase length differences. On the contrary,  $F_0$  contours of conversational speech samples differed at the adverb phrase position. Figure 1 shows the  $F_0$  average differences (in log scale) of same phrases between the reading style speech and the conversational speech at each adverb position when adjectives expressing positive-attitude follow. As shown in the figure, the  $F_0$  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 correlation over speakers was 0.85.

Figure 2 shows the contrast of the effects of attitude attribute adjectives, positive versus negative, 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 even though the adjectives reflected attitude contrast. 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 lexical markedness and neighboring attributes.

Table 1 Classification of  $F_0$  variations by their heights and dynamics

dynamics \ height	rise 	flat 	fall 	rise & fall 
higher 	surprise - more positive request to repeat - more interest negative response - more polite	agreeable attitude - more emphatic agreeable attitude - emphatic agreeable attitude - more acceptable	agreeable attitude - emphatic positive response - willingly	negative response - more polite
lower 	surprise - rather neutral request to repeat - rather neutral  request to repeat - less interest negative response - less polite	agreeable attitude - acceptable  agreeable response - insignificant  reluctance - hesitation agreeable response - serious reluctance - doubt reluctance - stronger doubt	agreeable response - insignificant  agreeable response - serious  positive response - unwillingly	negative response - less polite  negative response - much less polite

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

To confirm the naturalness of adverb phrases with different  $F_0$  heights, we carried out a perceptual evaluation test. As fully synthesized speech sometimes interferes with judgment accuracy, we used natural speech with different  $F_0$  heights.

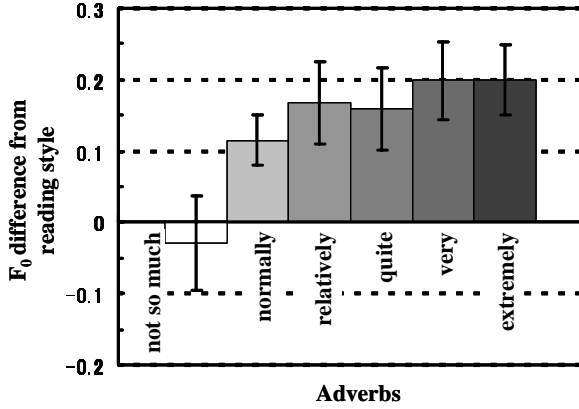


Figure 1 The increase of  $F_0$  average difference between reading and conversation in proportion to the increase of markedness of adverbs when positive adjectives follow

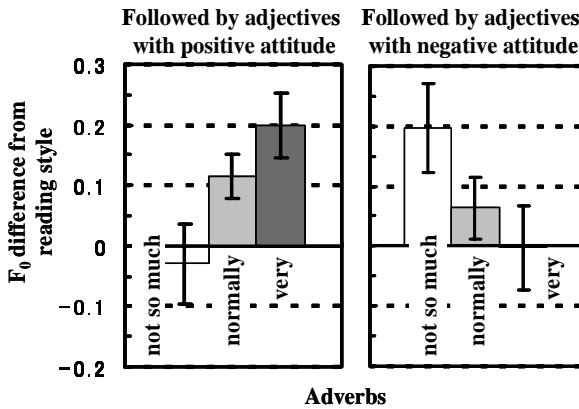


Figure 2 The effect of adjective attributes on the  $F_0$  differences between reading and conversation

For the perception experiment, we used sixty phrases consisting of seven adverb phrases followed by five adjectives expressing positive attitude and five adverb phrases (“extremely”, “very”, “normally”, “not so much” and “not at all”) followed by five adjectives expressing negative attitude. 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. The pure tone signals ranged from a G note (98.00Hz) to an F# note (185.00Hz) with a semitone interval. Through the analysis of these samples, we have confirmed that all twelve different  $F_0$  samples of each phrase combination have the same  $F_0$  contours for the accent component of the adverb phrases and that other prosodic differences are small.

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) as scores of each sample, corresponding to naturalness ratings. Icons of the twelve speech stimuli having the same phrase were displayed on the console screen for the subjects to listen freely before they were required to enter a score.

For each speech stimulus, the average score of each subject was calculated. As seen in Table 3, we listed maximum  $F_0$  values at the 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 darkest is the most preferred.) The table shows that the highest naturalness score is consistently changing over these samples. The results with adjectives expressing positive attitude show that a 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 the results with adjectives expressing negative attitude. These consistent but neighboring adjective dependent perceptual characteristics nicely match to the results of generation shown in Figure 2. These results confirmed the effectiveness of the  $F_0$  control in conversational speech by lexical markedness of adverbs and the attributes of neighboring adjectives.

Table 3 Average naturalness scores for the utterances with adverb phrases of different  $F_0$  heights

max $F_0$ at adverbs [Hz]	followed by an adjective with positive attitude							followed by an adjective with positive attitude				
	zenzen (not at all)	annmari (not so much)	futsuuni (normally)	sokosoko (relatively)	wariai (quite)	sootoo (very)	hijooni (extremely)	zenzen (not at all)	annmari (not so much)	futsuuni (normally)	sootoo (very)	hijooni (extremely)
185.00(F#)	1.56	1.42	1.70	1.96	2.26	3.48	3.78	2.80	2.32	1.52	1.96	2.20
174.61(F)	1.76	1.62	2.14	2.48	2.60	3.74	4.10	3.12	2.72	1.78	2.26	2.32
164.81(E)	2.10	2.00	2.62	2.94	3.18	4.00	4.16	3.46	3.22	2.24	2.46	2.46
155.56(D#)	2.36	2.56	3.20	3.48	3.82	3.98	4.06	3.50	3.56	2.82	2.62	2.52
146.83(D)	2.84	2.88	3.52	3.72	4.04	3.84	3.98	3.64	3.72	3.22	2.88	2.84
138.59(C#)	3.20	3.16	3.96	4.14	4.14	3.56	3.50	3.66	3.86	3.66	3.22	3.26
130.81(C)	3.48	3.50	4.12	4.18	4.00	3.28	3.12	3.52	3.80	3.92	3.64	3.50
123.74(B)	3.80	3.90	4.10	3.98	3.64	2.94	2.70	3.10	3.60	4.14	3.80	3.84
116.54(A#)	3.98	4.08	3.66	3.60	3.30	2.50	2.38	2.74	3.10	4.06	4.04	3.92
110.00(A)	4.34	3.92	3.12	3.00	2.66	2.20	1.94	2.38	2.54	3.76	4.12	4.14
103.83(G#)	4.34	3.72	2.56	2.56	2.36	1.84	1.64	2.18	2.34	3.54	4.22	4.12
98.00(G)	4.18	3.54	2.30	2.32	2.12	1.70	1.54	1.94	2.08	3.44	4.04	3.94

## 5. F<sub>0</sub> generation of conversational speech using lexical information

For the F<sub>0</sub> control of conversational speech, we extracted the F<sub>0</sub> generation parameters proposed by Fujisaki [5] for the speech samples used in section three. All F<sub>0</sub> contours were approximated by the same amplitude of phrase command (**Ap**) but amplitudes of the accent commands (**Aa**) were varied in proportion to the degree of the adverbs. To control F<sub>0</sub> based on a generation model, we tried to find a mapping function from the lexical markedness to **Aa**.

To measure the control accuracies for open lexicons and speakers, we repeatedly split the data into training sets and test sets. A mapping function was fitted for each training set using a sigmoid function by minimizing the RMS error between observed F<sub>0</sub> contours and generated ones.

For an F<sub>0</sub> generation experiment using unrestricted lexicons, one hundred-twenty **Aa** values of adverb phrases were extracted from six adverbs followed by five adjectives expressing a positive attitude, from speech uttered by four speakers. They were normalized for each speaker by using the average value and the standard deviation of the corresponding speaker. One hundred normalized **Aa** values of other adverbs were used for mapping function training and the remaining twenty samples of the corresponding adverb were tested. The lexical markedness of adverbs is given by the average of scores expressing the subjective markedness over four subjects.

Table 4 shows open-lexicon RMS errors of normalized **Aa** for an adverb phrase and (i) output of the mapping function given an input lexical markedness of the corresponding adverb, (ii) mean observed value of each adverb and (iii) the mean observed value of all adverbs. As shown by the comparison between these three kinds of average RMS errors, the input lexical markedness can account for 74% of the reduction of observed variances. As the number of adverbs is quite small, the mapping function may not be correctly estimated in some training data set. We found that the RMS errors varied from 0.0658 to 0.1189 when we changed the training data by the combination of three adverbs. The lowest error 0.0658 which

is comparable to 0.0665 of (ii) the mean observed value of each adverb was obtained when adverbs expressing the strongest, middle, and weakest degree of modification (i.e. "not so much", "quite" and "extremely") were used for the mapping function training.

To determine the speaker dependencies of F<sub>0</sub> control, three speakers' data (ninety tokens) were used for the training of the mapping function and one speaker's data (thirty tokens) were used as a test set. As for the open-lexicon experiment, speaker normalized **Aa** data for adverb phrases were used for the training. Table 5 shows open RMS errors of normalized **Aa** for an adverb phrase and (i) predicted values using lexical markedness, (ii) the mean of each speaker and (iii) the mean over all speakers.

As shown in the smallness of differences between (i) and (ii) of Table 5, prediction from other speaker's characteristics looks quite effective. The RMS error differences are smaller than those in the open-lexicon experiments.

## 6. Conclusions

F<sub>0</sub> characteristics were analyzed and modeled for prosody control in communication systems. A preliminary study on the prosodic variations of a single utterance /n/ in conversational speech revealed that F<sub>0</sub> height and dynamics reflected the speaker's attitude and that they were highly related to the corresponding lexical forms expressing attitudes. Systematic control of short utterance selection in conversational contexts showed the consistent effects of lexical information on F<sub>0</sub> height in expressing speaker's attitude and markedness of adverbs. The same control characteristics were supported by perceptual experiments on naturalness evaluation for speech with different F<sub>0</sub> heights. These consistent control characteristics were modeled by training a monotonously increasing mapping function between subjective scores of lexical markedness and amplitudes of accent command **Aa** for an F<sub>0</sub> generation model. Open experiments on the prediction of **Aa** showed the usefulness of the training of this mapping function.

## References

- [1] Sagisaka Y., 1990, On the prediction of global F<sub>0</sub> shape for Japanese text-to-speech. Proc. ICASSP., 325-328.
- [2] Riley M.D., 1992, Tree-based modeling of segmental durations in Talking Machines edited by G.Bailly et al North-Holland., 265-274.
- [3] C. Traber, 1995, SVOX: The implementation of a Text-to-Speech System for German, TIK-Schriftenreihe Nr 7.
- [4] Tokuda K.; Masuko T.; Miyazaki N.; and Kobayashi T., 1999, Hidden Markov models based on multispace probability distribution for pitch pattern modeling, Proc. ICASSP, 229-232.
- [5] Fujisaki H., Hirose K., 1984, Analysis of voice fundamental frequency contours for declarative sentences of Japanese, J. Acoust. Soc. Japan (E), Vol.5, No.4, 233-242.

Table 4 Lexicon open RMS errors of normalized accent command **Aa** (Errors between observed values and (i) predicted values using lexical markedness, (ii) the mean of each adverb and (iii) the mean over all adverbs)

adverb	not so much	normally	relatively	quite	very	extremely	average
(i)	0.1198	0.0694	0.0836	0.0613	0.0573	0.0513	0.0738
(ii)	0.0889	0.0580	0.0841	0.0583	0.0572	0.0527	0.0665
(iii)	0.1736	0.0568	0.0825	0.0579	0.0873	0.0847	0.0905

Table 5 Speaker open RMS errors of normalized accent command **Aa** (Errors between observed values and (i) predicted values using lexical markedness, (ii) the mean of each speaker and (iii) the mean over all speakers)

subject	speaker A	speaker B	speaker C	speaker D	average
(i)	0.0478	0.0595	0.0854	0.0556	0.0621
(ii)	0.0429	0.0502	0.0798	0.0568	0.0574
(iii)	0.0780	0.0955	0.1621	0.1132	0.1122

