F0 control characterization by perceptual impressions on speaking attitudes using Multiple Dimensional Scaling analysis

Yoko Kokenawa[a] Minoru Tsuzaki[b] Hiroaki Kato[c] Yoshinori Sagisaka[d]

a)d) GITS, Waseda University 1-3-10 Nishi-waseda Shinuku-ku, Tokyo, 169-0051, Japan
b) Kyoto City University of Arts 13-6 Kutsukake-cho, Oe, Nishikyo-ku, Kyoto, 610-1197, Japan
c) ATR Human Information Science Labs, 2-2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0288, Japan

Email: a) yoko.kokenawa@toki.waseda.jp, b) minoru.tsuzaki@atr.jp, c) kato@atr.jp, d) sagisaka@giti.waseda.ac.jp

ABSTRACT

Aiming at prosody control for speech synthesis expressing speaking attitudes, F0 shapes were characterized by their perceptual impressions. To directly correlate F0 shapes with perceptual impressions, single word utterances “n” extracted from daily conversations were employed. The analysis showed that speaking attitudes were manifested in the global F0 control of "n" as the differences of their average height (high-low) and dynamic patterns (rise, flat, fall and rise&fall). Next, controlled utterances of “n” were perceptually examined through Multiple Dimensional Scaling analysis to confirm F0 control freedoms found in the analysis. The result showed the three-dimensional structure of a perceptual impression space and factor dependent F0 control characteristics. The *positive-negative* attitude can be controlled by average F0 height while those of *confident-doubtful* or *allowable – unacceptable* are manifested through dynamic F0 patterns. These findings provide new possibilities of systematic F0 control for conversational speech synthesis with speaking attitudes using corpus-based approach.

1. INTRODUCTION

A corpus-based approach has successfully improved output speech quality in text-to-speech (TTS) systems [1]-[4]. As the increase of application domains, the insufficiencies of output speech prosody become one of the serious problems in conversational situations. However, it is difficult to specify input control factors and model the prosody characteristics in conversational speech generation. For this reason, only limited works have been carried out to model prosody control for conversational speech generation.

To improve the naturalness of conversational speech, we have proposed to reflect speaker's attitude to F0 control by employing the information of output words by themselves [5]. This study quantitatively showed the possibility of prosody control expressing positive/negative attitude with magnitude through markedness of constituent adverbs and adjectives expressing positive/negative attitudes. Though this study showed the possibility of prosody control with speaking attitude in conversational situations, the control is limited where lexicon intrinsic information plays main role. We have to consider other control freedoms to express various speaking attitudes.

As shown in emotional speech studies [6] or in speech data collection under actual living environments [7,8], considerable amount of prosody variations exist in conversational speech expressing attitudes. To expand our corpus-based approach to express speaking attitudes, we need to know how many independent prosody controls exist and how possibly output speaking attitudes are independently described.

In this paper, we have tried to characterize the F0 shapes by their perceptual impressions. Through these analyses, we aimed at finding prosodic control freedoms and the dimension of word expressions used for perceptual impressions. In particular, we analyzed frequently seen single word utterances “n” with different prosody that speakers consciously or unconsciously generate based on their attitudes or intensions.

The word "n" is, acoustic-phonetically, a sustained utterance of either [m], [n], or eng. In Japanese, it can be an interjection, rejoinder or filler depending on the context or situation. Most importantly, it does not have any particular lexical meaning, default intonation or accent, so that its F0 pattern most likely conveys the speaker's intention. Therefore, we adopted “n” as an ideal sample to see the function of F0 characteristics. The analysis on F0 characteristics of “n” with perceptual impression not only quantitatively shows prosodic differences among themselves but also gives some hints on the relation between perceptual impression as input and F0 characteristics as control freedoms for speech synthesis with speaking attitudes.

In the following sections, first we analyzed F0 characteristics of “n” observed in daily conversation in Section2. In Section3, perceptual experiments were conducted using controlled utterances of “n” to quantitatively confirm observed F0 characteristics using and to obtain regular expressions for perceptual impressions. Section4 explains the Multi-dimensional Scaling analysis results on perceptual impression expressed in basic word vectors using. It is noted that F0 control can be nicely characterized by three-dimensional independent factors expressing speaking attitudes. Finally, we summarize all findings and discuss further works to implement current findings to speech synthesis with speaking attitudes.

2. SPEAKING ATTITUDE CLASSIFICATION BY F0 HEIGHT AND DYNAMIC PATTERNS

In order to specify the speaking attitudes conveyed by conversational prosody, we have observed the prosodic variations of single word utterance “n” frequently used in conversations.

Table1. Speaking attitudes of "n" in the subsequent utterance expressions classified by F0 average height and dynamic patterns

dynamic pattern / height	rise	flat	fall	rise & fall
high	Really? ♪ Then, what happen?? ♪	And? and? ♪ ♪ Then what? ♪ I agree!!	Really? I did not know that ♪ Of course!! ♪	Never mind ♪
↑	Liar!! ♪	okay	That is nice ♪	That is okay.
↓	Really? Is that true?	I do not know… Is that okay ?	That is fine.	No, I do not like that ↓
low	I didn't know that. Really ↓	I am not sure… I do not quite agree…	okay… ↓ I did not know that… ↓ yes… but… ↓ ↓	

Speech in this "n" category were used to express quite a wide variety of attitudes/intensions coupled with different prosodic patterns, although they could be labeled with a single phonemic category. Therefore, it enabled us to compare linguistic and/or perceptual effects of prosodic variation without being bothered by other linguistic factors.

For the analysis, we have collected forty-two "n" samples from thirty-minute conversations by four female adults who share a friendly relationship. The F0 characteristics of each sample and the subsequent utterances were extracted. After classifying the subsequent utterances by the average height and dynamics of F0, we found that the content of utterances following "n" could be utilized to infer the speakers' attitudes that occasionally might be difficult to specify only from the prosodic information.

Table1 shows the classification of speaking attitudes of "n" using the subsequent utterance expressions. As shown in Table1, they were classified by average height and dynamics of F0. To specify speaking attitudes in these samples more precisely and quantitatively, we have conducted a perceptual experiment to get quantitative expression of speaking attitudes in multidimensional space.

3. QUANTIATIVE EXPRESSION OF SPEAKING ATTITUDES BASED ON PERCEPTUAL IMPRESSION

3.1. Selection of basic word descriptions

In order to quantify what people perceived on speaking attitudes and to obtain more precise and general impression expressions of speaker's attitudes that can be conveyed by F0 patterns, we decided to regulate word usage to describe perceptual impression. Based on the previous observation, twelve single word utterances "n" that were controlled by F0 average height (three kinds) and F0 dynamics (four kinds) were prepared as speech stimuli. Those speech stimuli were uttered by the first author, who tried not to set up any conversational situation in order to avoid showing intentional attitude. In Table2, the maximum and minimum F0 of each sample are listed respectively. Five subjects (two male; three female) with normal perception ability listened to each sample and were asked what word or phrase they could imagine follows "n" in the sample. The subjects were then asked to subjectively presume speaking attitudes.

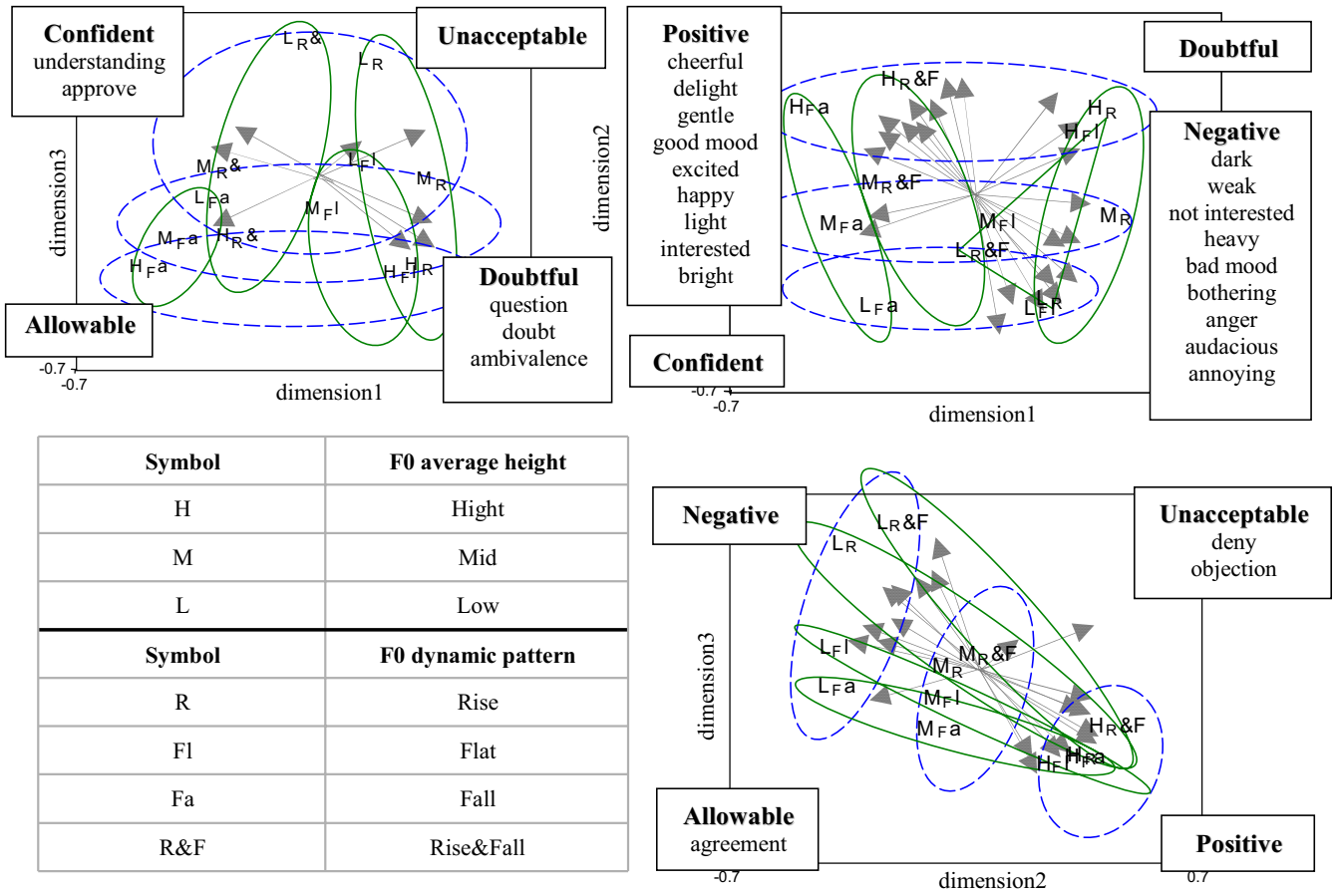
As the result, sixty-seven impression words were obtained to indicate speaking attitudes. More than two subjects chose twenty-six identical words. These basic impression words could be classified into the following three groups, *doubtful- confident* (doubt, ambivalence, understanding, approve), *unacceptable–allowable* (deny, objection, agreement) and *negative- positive* (dark, weakly, not interested, bad mood, heavy, bothering, audacious, anger, annoying, cheerful, delight, gentle, good mood, excited, happy, light, interested, bright)

3.2 Vector expression of each perceptual impression by basic words

To treat each perceptual impression in a quantitative way, we decided to "approximate" each perceptual impression by the above twenty-six basic impression words. Though these twenty-six basic impression words are still mutually dependent and redundant to express speaking attitudes, they look still sufficient enough to use for approximation. In order to see how each sample was possibly perceived based on basic impression words, the identical twelve single word utterances of "n" from section 3.1 were again used as speech stimuli.

Table2. Maximum and minimum F0 of speech samples in the listening test to get descriptions for perceptual impressions

height / dynamic pattern	High-range		Mid-range		Low-range	
	max	min	max	min	max	min
Rise	354.55Hz	182.20Hz	282.58Hz	142.17Hz	194.21Hz	98.24Hz
Flat	264.55Hz	232.39Hz	213.23Hz	178.90Hz	162.77Hz	124.97Hz
Fall	305.50Hz	119.81Hz	234.22Hz	90.87Hz	155.94Hz	65.98Hz
Rise&Fall	363.46Hz	222.15Hz	273.08Hz	153.48Hz	163.17Hz	111.44Hz



Symbol	F0 average height
H	Hight
M	Mid
L	Low
Symbol	**F0 dynamic pattern**
R	Rise
Fl	Flat
Fa	Fall
R&F	Rise&Fall

Figure1. Projection of impression words on plane surface on 3 dimensions by INDSCAL (circle the same F0 height with dash line and the same F0 shape with full line)

Explanatory Note: Symbol notation of F0 characteristics in figure1

A new group of subjects were asked to use an eight level scaling system, 0(not at all)- 7(very much), to measure how much each speech sample related to each impression word. Hence, each subject completed three hundred twelve operations in total. The subjects were five native Japanese adults (one male; four female). Also they were allowed to listen to the samples repeatedly. The total necessary rating time per subject was from thirty to forty minutes.

4. FREEDOMS IN SPEAKING ATTITUDES AND F0 CONTROL

4.1 MDS analysis on perceptual impression of speaking attitudes

In order to obtain the quantitative expression of speaking attitudes based on the basic impression words, Multi-Dimensional Scaling (MDS) analysis [8] was applied. MDS analysis can reveal the independent dimensions based on data of distance or similarity and the underlying structure and constraints where the samples follow. We decided to use Individual Differences Multidimensional Scaling (INDSCAL) algorithm [8] to treat all subjects' responses in the same framework. The INDSCAL algorithm treats different subject data in a common multiple dimensional space where the individual disparity is expressed by the differences of scaling for each axis.

Input data from five subjects were prepared from the rating differences of the twelve samples expressed in the twenty-six basic impression words. The decision of the number for appropriate dimension could be made using empirical knowledge. In our analysis, three-dimensions were taken by reference to Variance Accounted For (VAF) shown in Table3. As shown in Table4, every subject seemed to use appropriate weights toward the three-dimensional spaces, so all five subjects' data were employed as input. To interpret each axis meaning, the average scores corresponding to each basic impression words were projected to the three-dimensional spaces using multiple linear

Table3. Variance Accounted For toward each number of dimension (VAF)

	Dimension			
	1	2	3	4
VAF	0.7398	0.8036	0.816	0.5952

Table4. Individual weight to each dimension

Subjects	Dimension		
	1	2	3
SY	0.9459	0.7651	0.8470
CA	0.4933	0.7668	0.7051
YY	0.6428	0.7821	0.8761
FY	0.7063	0.4332	0.6148
KK	0.5833	0.7868	0.6851

regression analysis. In Figure1, the actual basic impression words that acquired the high multiple correlation coefficient and regression coefficient toward the three-dimensions are shown.

As shown in these figures, we can approximate each basic impression word in three dimensions expressing the speaking attitudes of *positive–negative*, *confident–doubtful* and *allowable–unacceptable*. The axes of *confident–doubtful* and *positive–negative* can be projected on the plane spanned by the first and second dimension. The axes of *allowable–unacceptable* and *confident–doubtful* can be interpreted in the plane spanned by the first and third dimension. The axes of *allowable–unacceptable* and *positive–negative* are interpreted in the plane by the second and third dimension.

These results nicely coincide with our intuitive grouping of twenty-six basic expressions given in section 3.1 and support the possibility of treatments of perceptual impressions by the restricted number of freedoms conveyed just by F0 average height and shapes.

4.1. F0 control using three-dimensional speaking attitudes

From Figure1, it was found that the F0 control characteristics were also well organized in the three- dimensional space expressing *positive-negative*, *confident–doubtful* and *unacceptable–allowable*. As shown in the figures, F0 average heights and F0 dynamic patterns were independently corresponding to perceptual impressions of speaking attitudes. Speaking attitudes corresponding to *positive-negative* are highly related with F0 average heights while the ones of *confident–doubtful* and *unacceptable – allowable* were related with F0 dynamic patterns.

The finer analysis of F0 control characteristics indicates that the higher/lower the F0 average height was, the more *positive/negative* attitude appeared to be indicated respectively as shown in the plane spanned by the first and second dimension and the second and third dimension. F0 dynamic patterns were located in order of Rise, Flat, Rise+Fall, and Fall, and this order was corresponded to speaking attitudes of *doubtful-confident* to be shown in the plane spanned by the first and second dimension and the first and third dimension. Moreover, as observed in the plane by the second and third dimension, the speech samples were located with the order of Rise+Fall , Rise, Flat, and Fall from the *unacceptable* speaking attitudes toward *allowable*.

To reconsider the result from F0 control viewpoint, those results indicate the possibility of F0 control based on the speaking attitudes. *Positive-negative* speaking attitudes control the degree of F0 average height while the speaking attitudes of *confident – doubtful* and *unacceptable – allowable* control the differences of F0 dynamic patterns. Though the analysis object in this study was single word utterance “n”, this propensity should be also able to expand to ordinary sentences. As observed from daily conversation in Section2, the subsequent utterances were strongly related semantically and in prosody.

5. SUMMARY AND FUTURE WORKS

This paper has described the attempt to find the dimensions of expression words obtained by perceptual impressions and the freedoms of prosodic control. After classifying single word utterance “n” from daily conversations, we found that the speaking attitudes were expressed through the differences of F0 average height and dynamic patterns. Based on this finding, we conducted perceptual experiments using F0 controlled utterances. First, we collected the basic impression words to be perceived through F0 contour. And those impression words were approximated through MDS analysis. The result showed that perceptual impression spaces dependent F0 average height and dynamic patterns were described in the three-dimensions, which were *positive-negative*, *confident-doubtful* and *allowable-unacceptable*. Moreover, speaking attitude of *positive-negative* was perceived by the differences of F0 height while those of *confident-doubtful* and *allowable-unacceptable* were highly related with F0 dynamic patterns. This result indicates the possibility of F0 control by speaking attitudes.

These F0 control findings are useful especially for speech synthesis with speaking attitudes using corpus-based approach. Many prosodic factors are involved in two-way dialogues to move the conversations along smoothly. In order to create a conversational F0 control model, we need to specify those prosodic factors. Even though this study is the first step toward the conversational F0 control model, the result surely showed the prosodic factor in conversational speech. At the same time, this indicates realizable possibility of the conversational speech synthesis. In the future, from the viewpoint of how this finding can be related with the default F0 contour of each content word/sentence, we will expand our research in order to develop to F0 control model.

Acknowledgment

Work supported in part by the Grant-in-Aid for Scientific Research (A) (2) No. 16200016, JSPS

6. REFERENCES

[1]Riley M.D., Tree-based modeling of segmental durations, Talking Machines edited by G.Bailly et al, North-Holland, pp.265-274, 1992

[2] Sagisaka Y., “On the prediction of global F0 shape for Japanese text-to-speech”, Proc. ICASSP, pp.325-328, 1990

[3] 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

[4] Traber C., SVOX: The implementation of a Text-to-Speech System for German, 1992, TIK-Schriftenreihe Nr 7

[5] Sagisaka,Y., Yamashita, T., and Kokenawa, Y., "Speech Synthesis with Attitude" Proc. Speech Prosody 2004, pp.401-404, 2004

[6] Campbell, N., “Speech & Expression; the value of a longitudinal corpus”, Processing of the Language Resources & Evaluation Conference Lisbon, 2004

[7] Campbell, N., and Erickson, D., “What do People Hear? A study of the Perception of non-verbal Affective Information in Conversational Speech” Journal of the Phonetic Society of Japan, Vol.8, no1, pp.9-28, April 2004

[8] Campbell, N., “Perception of Affect in Speech - towards an Automatic Processing of Paralinguistic information in Spoken Conversation” Proc. INTERSPEECH 2004 - ICSLP 8th Int. Conf. on Spoken Language Processing, Vol.II, pp.881-884, 2004

[9] Borg, I., Groene, P., “Modern Multidimensional Scaling: Theory and Application”, Springer, N.Y., 1997