A trial of communicative prosody generation based on control characteristic of one word utterance observed in real conversational speech

Yoko Greenberg[a] Nagisa Shibuya[a] Minoru Tsuzaki[b] Hiroaki Kato[c] Yoshinori Sagisaka[a]

a) GITS, Waseda University 1-3-10 Nishi-waseda Shinuku-ku, Tokyo, 169-0051, Japan
b) Kyoko 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

yoko.kokenawa@toki.waseda.jp, n.shibuya@fuji.waseda.jp, minoru.tsuzaki@atr.jp, kato@atr.jp, sagisaka@giti.waseda.jp

Abstract

Aiming at prosody control for conversational speech synthesis, communicative prosodies were generated based on the prosodic characteristics derived from one word utterance “n”. The grouping of F0 patterns using VQ revealed four F0 dynamic patterns (rise, gradual fall, fall, and rise&fall) for large amounts of one-word utterance “n” in daily conversations. Through the analysis using an F0 generation model, different control characteristics were found for these patterns. A communicative prosody control scheme is proposed for short utterances reflecting these control characteristics for three dimensional representative perceptual impressions, *confident-doubtful, allowable-unacceptable* and *positive-negative* previously obtained by MDS analysis. The naturalness evaluation tests for synthesized conversational speech showed superiority in naturalness of the proposed prosody control. These results indicate the possibility of communicative prosody generation for conversational speech synthesis through perceptional impression expressions 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.

In our previous studies on communicative prosody control modeling [5],[6], we have tried to describe in terms of perceptual impressions twelve one-word utterances of “n” with three F0 average height (low, mid and high) and four F0 dynamic patterns (rise, fall, flat and rise&fall). Through MDS analysis, we found that the dimension of perceptual impressions could be reduced in three dimensions expressing basic impressions (*confident-doubtful, allowable-unacceptable* and *positive- negative*), which could then be used to identify output prosodic characteristics [5]. Furthermore, it was indicated that this correspondence of three dimensional perceptual impressions to prosodic characteristics could be applied not only to one word utterances of “n”, but also to actual phrases [6]. Throughout these investigations, we also found that duration played an important role to give the different perceptual impressions to the listeners.

In this paper, as a further step toward a conversational speech generation, we first characterized the prosodic patterns of “n” recorded in daily conversations. In the following section, using a large amount of live data, we tried to group the prosody variations and analyzed their control characteristics using a fundamental frequency (F0) generation model proposed by Fujisaki [7]. In section 3, based on the analysis results of “n”, a communicative prosody generation scheme is proposed for other phrases using perceptual impressions. To confirm the validity of the proposed scheme, we replaced the prosody of read phrases representing three-dimensional perceptual impressions to communicative F0 parameters and duration using STRAIGHT synthesis [8]. In section 4, naturalness evaluation results are described to show the effectiveness of proposed communicative prosody generation scheme and we summarize all the findings.

2. Prosodic characterization of one-word utterances “n” in daily conversations

2.1. One-word utterance data for communicative prosody characterization

In our pilot studies to correlate the perceptual impressions to the prosodic characteristics of one-word utterance “n”, we employed twelve prototypical prosody variations which consisted of three kinds of F0 average height and four kinds of dynamic patterns [5]. To generalize these findings, we need to know the actual variations of prosodic patterns of “n” in a large amount of live data. Moreover, quantitative analysis is needed based on F0 generation model to see how the prosodic variations were generated. For these purposes, we analyzed large amounts of one-word utterance “n” in the natural conversational speech recorded by JCT/CREST ESP project [9]. The 6,271 samples were randomly selected to analyze among 23,648 one word utterances of “n” from the 150 hours speech data spoken by one Japanese female.

2.2. F0 generation parameters of one-word utterance “n”

In order to characterize the prosodic patterns of one-word utterance “n” recorded in an actual living environment, vector quantization was applied to the extracted data. Before the quantization, the data was classified into seven categories by utterance duration. The durations ranged from 50 milliseconds to 400 milliseconds and were broken into 50 milliseconds increments. For quantization, F0 patterns were resynthesized by modifying control command time based on F0 generation model fundamental frequency (F0) generation model proposed by Fujisaki. The resynthesis using generation model parameters can normalize F0 contours of various time lengths and fill out F0 drops caused by voiceless sounds and thus prevent unnatural normalization by simple liner time lengthening and shortening.

The utilized F0 patterns were resynthesized by lengthening and shortening of manually extracted accent command positions in conformity with the duration. K-means method was employed for VQ. The VQ division was determined to twenty clusters for each duration class by observing the decrease of total distortion shown in Figure 1. Accordingly, 140 clusters were obtained. The obtained F0 contours were decomposed into three generation control parameters; the minimum value of F0 (F_{min}), the amplitude of the phrase commands (A_p) and the amplitude of the accent commands (A_a).

2.3. Comparison of F0 parameters among F0 dynamic patterns

From the observations of VQ clusters as shown in Figure 2, it was confirmed that the F0 dynamic patterns could be classified into four categories (rise, gradual fall, fall, and rise&fall) that had been employed in our previous study, though there existed marginal patterns. For F0 heights, they were continuously distributed. The analysis of F0 generation parameters revealed category specific control tendencies of the amplitude of phrase command A_p and accent command A_a.

As shown in Figure 3, phrase command A_p increased in the order of dynamic patterns of rise, gradual fall, rise&fall and fall. On the contrary, accent command A_a decreased in this order. These distribution differences indicate that F0 height control is not carried out uniformly. F0 height in rising patterns is mainly controlled by the change of accent command A_a. On the contrary, F0 height in gradual fall patterns is mainly controlled by the change of phrase command A_p. Other F0 height in two categories is controlled by both of them.

2.4. Correspondences between perceptual impressions and prosodic characteristics

As the previous study [6] showed the possibilities of perceptual impressions as input to specify prosody of output speech in conversational speech synthesis, we have confirmed the relationship between the prosodic patterns and perceptual impressions [5] using the current data. We asked seventeen native Japanese adults to put scores ranging from 0 (not at all) to 7 (very much) to categorize samples using twenty six impression words which were referred to our previous study (*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).

The results nicely coincided with our previous findings. Namely, F0 dynamics corresponds to the perceptual impressions described as distinctions of *doubtful-confident* and *unacceptable–allowable*. On the other hand, F0 average height gives the distinction between high group (*confident*, *allowable* and *positive*) and low group (*doubtful*, *unacceptable* and *negative*)

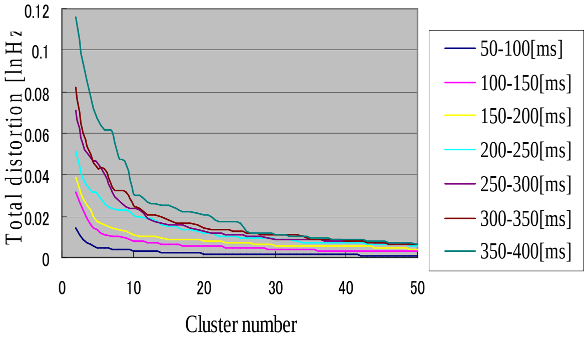

Figure1 Total distortion accompanying with the increase of cluster number

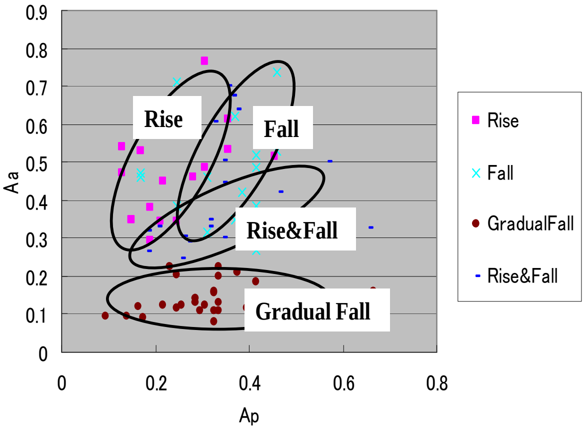

Figure3 Aa and Ap values for each F0 pattern category

In addition to F0 characteristics, total utterance duration also determined perceptual impressions. Higher scores were obtained for longer duration samples as *doubtful* and *unacceptable* while the shorter ones as *confident* and *allowable*.

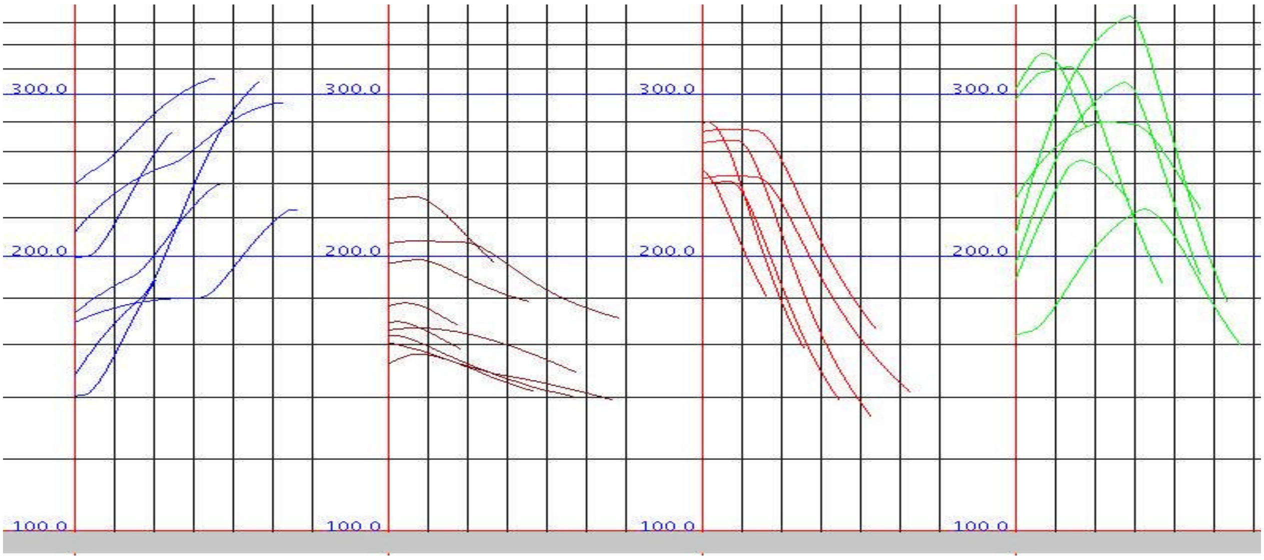

Figure2 Typical F0 dynamic patterns of “n” observed in real daily conversations

3. Communicative prosody generation based on control characteristics of one-word utterance “n”

3.1. Communicative prosody generation scheme

As we could have observed that the words directly expressing perceptual impressions had the same prosodic characteristics of one word utterance “n” with the same impressions [6], we can think of communicative speech generation scheme shown in Figure 4. As shown in the Figure, input words are not used to compute conventional prosody control characteristics used in text-to-speech systems but also their attributes can be used to estimate communicative prosody expressing *confident-doubtful, allowable-unacceptable* and *positive-negative.* Though the current modeling is carried out for each prototypical one-dimensional attribute only, we expect that this modeling can be expanded to combinations expressed as an impression vector determined by output lexicons.

These prosodic variations are newly generated in communicative prosody module using perceptual impression vectors. These perceptual impression vectors are obtained for each lexicon from the dictionary where multi-dimensional subjective impression scores are given to each lexicon. The communicative F0 dynamic patterns, durations and F0 average height are calculated from the input impression vectors.

3.2. Trials of communicative prosody generation

To confirm the validity of the proposed prosody generation scheme, we generated communicative prosody using control characteristics of “n”. As a first trial, we used twelve single phrase utterances consisting of lexicons expressing prototypical six impressions (*confident-doubtful, allowable-unacceptable* and *positive-negative*). These lexicons have already been confirmed to have an expected impression by a subjective evaluation test for lexicons by themselves (i.e. without speech). They were uttered in a reading style by two Japanese native speakers (one male and one female) and recorded in a quiet environment. The prosody of a read speech was used as an ideal prosody of conventional (i.e. read) component and was modified according to prosodic characteristics of “n” that has the same impression.

Table 1 shows the modification of F0 control parameters and total utterance duration. These values were obtained as differences between read-style “n” utterances and prototypical communicative “n” utterances for each impression by averaging several tens of corresponding speech samples. As for the values of phrase command A_p and accent command A_a, two values were given as two perceptual impressions *confident-doubtful* and *allowable-unacceptable* are corresponded to the degree of the difference of the four categories of F0 dynamic patterns. The numbers in case arc are obtained from the second closest F0 dynamic pattern to correspond to the perceptual impressions.

3.3. Perceptual naturalness evaluation

In order to see how the communicative prosody generation scheme reflecting the prosody of one-word utterance “n” would work for phrases expressing prototypical six impressions, perceptual evaluation tests were carried out. For the perceptual evaluation tests of naturalness, five different patterns of speech corresponding to two dimensional perceptual impressions; *doubtful-confident unacceptable–allowable* were prepared. Besides the neutral speech, prosody of each sample was converted to four different patterns that

Table 1 Modification for communicative prosody generation

	Confident	Doubtful	Allowable	Unacceptable	Positive	Negative
Fmin	+0.3	-0.25	+0.3	-0.25	+0.3	-0.25
Ap	*1.99(2.08)	*1.42(1.89)	*1.99(1.89)	*2.08(1.42)	*1	*1
Aa	*2.26(1.86)	*2.19(0.64)	*2.26(0.64)	*1.86(2.19)	*1	*1
duration	*0.75	*1.3	*0.75	*1.3	*1	*1

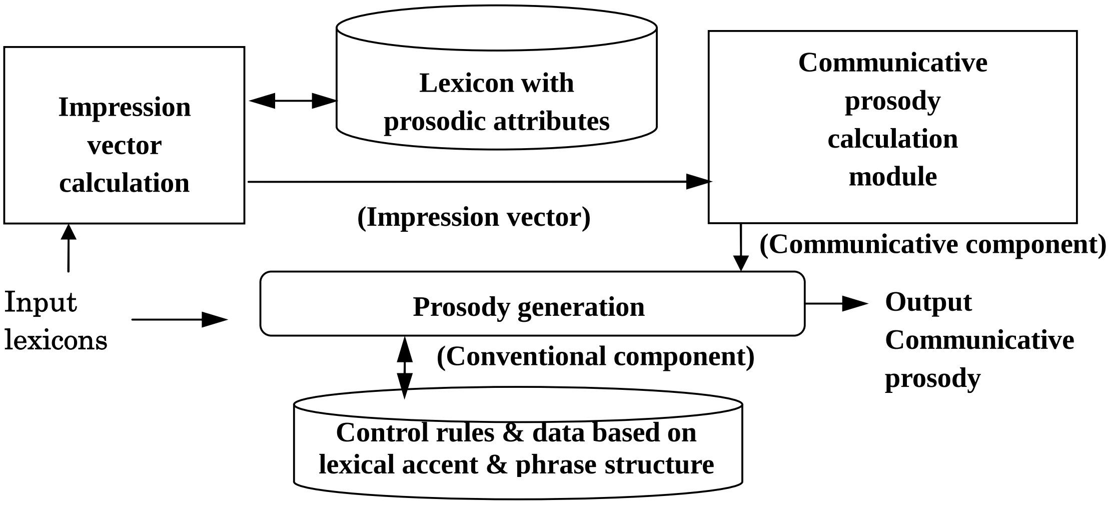
Impression vector calculation
Lexicon with prosodic attributes
Communicative prosody calculation module
(Impression vector)
(Communicative component)
Input lexicons
Prosody generation
Output Communicative prosody
(Conventional component)
Control rules & data based on lexical accent & phrase structure

Figure 4 Communicative prosody generation using impression prediction by input lexicons

modified only F_{min}, A_a and A_p, and duration, and all of the four prosodic characteristics using STRAIGHT speech synthesis [8]. As for the speech samples expressing *positive-negative*, the F_{min} modified version was generated. Accordingly, forty eight speech samples were prepared in total.

We asked a group of subjects to put scores ranging from 0 (reading style) to 7 (conversational style) to each sample phrase. The subjects were consisted of four native Japanese speakers (two male and two female). As shown in Figure 7, the results of the subjective perceptual evaluation test showed the superiority in naturalness of speech with communicative prosody by the proposed control scheme.

4. Conclusions

Using large amounts of one-word utterance "n" in ordinary daily conversations, we could have confirmed that communicative prosodic variations of "n" can be characterized by four types of F0 dynamics (rise, gradual fall, fall, and rise&fall), F0 height and duration, though there exist considerable amount of marginal patterns. Analysis using F0 generation model revealed F0 pattern specific control characteristics. Using high correlation between perceptual impressions and prosodic control characteristics, communicative prosody control scheme was proposed.

In the proposed control scheme, the prosodic characteristics of one-word utterance "n" were adopted to synthesize the communicative prosody of short phrases with the same impression as a lexical attribute. For prototypical perceptual impressions *positive-negative*, *confident-doubtful* and *allowable-unacceptable*, conversational speech was synthesized. Subjective naturalness evaluation confirmed the appropriateness of the proposed communicative prosody generation scheme. Though the current study showed the possibility of communicative prosody generation using lexical attributes in very restricted samples, the proposed approach is quite promising to generalize this control scheme to much wider word classes and word strings using corpus-based approach.

Acknowledgments

This work was in part conducted under the Waseda University RISE research project of "Analysis and modeling of human mechanism in speech and language processing".

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

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] Kokenawa, Y., Tsuzaki, M., Kato, K., and Sagisaka, Y., "F0 control characterization by perceptual impressions on speaking attitudes using Multiple Dimensional Scaling analysis", Proc. ICASSP, SP-P1.3.(1-273), Mar.2005

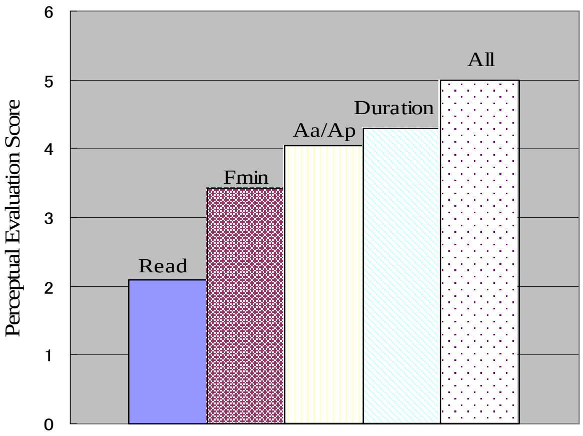

Figure5 Naturalness increase by adding communicative prosodic characteristics

[6] Greenberg, Y., Tsuzaki, M., Kato, K., and Sagisaka, Y., "Communicative speech synthesis using constituent word attributes", Proc. Interspeech2005, pp.517-520, Sep.2005

[7] Fujisaki, H. and Hirose, K., "Analysis of voice fundamental frequency contours for declarative sentences of Japanese," J. Acoust. Soc. Japan (E), Vol.5, No.4, 233-242, 1984

[8] Kawahara, H., Masuda-Katsuse, I. and Cheveign´e, 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," Speech Communication 27, 187-207, 1999.

[9] JST/CREST Expressive Speech Processing project.
http://feast.atr.jp/esp/esp-web/