



Inter-language prosodic style modification experiment using word impression vector for communicative speech generation

Ke Li, Yoko Greenberg and Yoshinori Sagisaka

GITI/Language and Speech Science Res. Lab Waseda University 1-3-10 Nishi-Waseda, Tokyo, 169-0051, Japan
rikoku@akane.waseda.jp, yoko.kokenawa@toki.waseda.jp, sagisaka@giti.waseda.jp

Abstract

To confirm the language independency of a communicative prosody generation from input word impression vector, we synthesized communicative Mandarin speech using prosodic characteristics of communicative Japanese speech. The fundamental frequency and duration characteristics of one-word “n” utterances of Japanese were copied to Mandarin through input word attributes. From the subjective impressions of an input word, a three-dimensional vector was calculated through Multi-Dimensional Scaling analysis. Three dimensions reflecting impressions of confident-doubtful, allowable-unacceptable and positive-negative correspond to systematic prosodic variations; F0 height, F0 dynamics and duration. Subjective evaluation of synthesized speech showed the possibility of communicative prosody generation from input word impression vector language independently.

1. Introduction

In the study of prosody control, up to now, linguistic control correlates have been mainly studied. Though it is quite natural to formulate prosody control from linguistic viewpoints, this tendency may restrict the understanding of prosody control only to written language. For better understanding of spoken language and the generation of communicative speech, we have been studying the F0 characteristics of one-word utterances of “n” [1]-[3]. Through these studies, we have tried to specify the input and the output for communicative prosody. By analyzing F0 of nonverbal utterances “n” in real communication, we found that their perceptual impression could be approximated by three-dimensional impressions corresponding to confident-doubtful, allowable-unacceptable and positive-negative [1].

Through the study on communicative speech synthesis, we found that word intrinsic lexical impression, for example, subjective scores on markedness for adverbs are effectively used in communicative speech synthesis [4]. In daily speech communication, there seem to exist high correlations between input words and their prosody. When talking unacceptable issues, people use words showing unacceptable meaning with prosody with unacceptable feeling. This correlation indicates communicative prosody can be specified by input words by themselves. Our previous studies on communicative prosody generation have given the evidences supporting this idea [1]-[6]. By naturally generalize this idea, it may be possible for us to share the same communicative prosody control characteristics language universally.

As a first step to confirm language universality, we carried out communicative prosody generation using speech data of two different languages. We employed communicative F0 control characteristics of Japanese one-word utterances of

“n” to generate communicative prosody of Mandarin via input word attributes.

In Section 2, we first describe a communicative prosody generation scheme that we have been proposing [3]-[6]. Next, we explain how we generate communicative prosody from input word attributes. This generation scheme enables communicative Mandarin prosody generation from communicative Japanese characteristics. The experimental setup of prosodic style modification is introduced in Section 3. In Section 4, Experimental results and discussions are described. Finally, conclusions and further works are stated in Section 5.

2. Communicative speech generation from input word impression vector

2.1. Word impression vector as input specification

As shown in synthesis of communicative speech, the specification of input information has been one of the most serious problems since we need some information to change read prosody to communicative one. Though we think that corpus-based approach is still useful, we cannot solve this problem merely by swapping read speech data to communicative one. As often seen in many studies treating expressive prosody or emotional prosody, simple training data swapping to the one with specific characteristics cannot cover all prosodic variations in communication systematically. We have been studying communicative prosody variations of single phrases dependent on input words [1]-[6].

Through a series of analyses and syntheses, we have confirmed that input words by themselves could explain differences between communicative prosody and read one. Based on these studies, we have proposed a communicative prosody generation scheme as shown in Figure 1. In this scheme, communicative prosody is generated by the conventional prosodic component and communicative component which is newly obtained from input word attributes. For this input word attributes, we proposed a *word impression vector* specified by word intrinsic properties in relation to words to describe perceptual impression of communicative speech. For this impression vector, we tentatively adopted multi-dimensional subjective scores of the lexicon by quantifying twenty-six components corresponding in seven scales for each.

By using Euclidian distance of these impression vectors as a metric, Multi-Dimensional Scaling (MDS) analysis showed that the impression vector can be approximated by three dimensions expressing *confident-doubtful, allowable-unacceptable and positive-negative* [1].

Furthermore, it has been observed that these three dimensions nicely correlated prosodic control

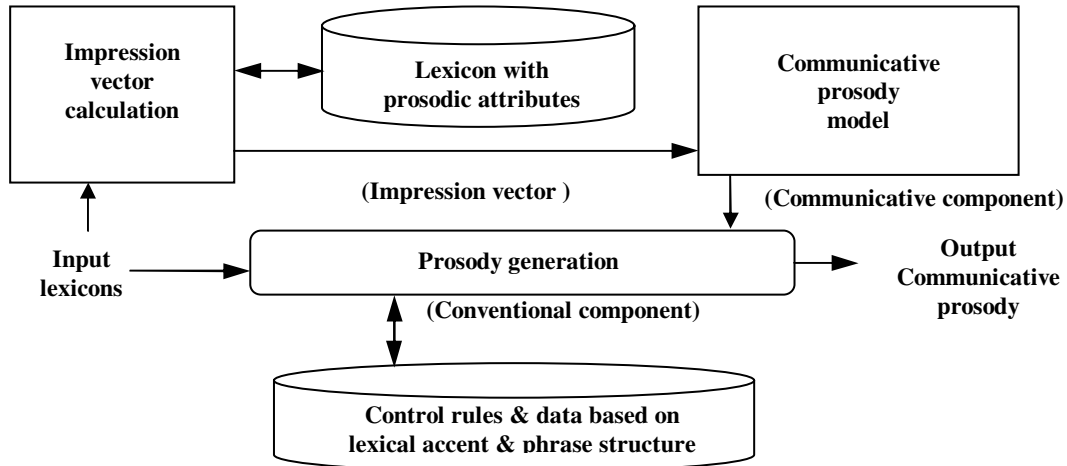


Figure1. Communicative prosody generation using impression prediction by input lexicons

characteristics not only one-word utterances of “n” but also many single phrases.

As we expect that communicative component has language independent characteristics and can be calculated independent to conventional language-dependent prosodic characteristics, we decided to confirm the possibility of communicative component using different language. In the following sections, we tried to synthesize one-word Mandarin speech with communicative prosody using communicative characteristics extracted Japanese speech. To keep the synthetic speech quality as much as natural, we employed read Mandarin speech instead of using conventional component. This means that each read Mandarin speech was modified using Japanese communicative prosodic characteristics. An input word impression vector was employed to determine communicative component of Japanese for modification explained in the next subsections.

2.2. Communicative F0 generation using a command-response model

To modify Mandarin read speech using Japanese communicative characteristics, we employed a command-response model proposed by Fujisaki [7]. As this model enables F0 modification constrained by F0 generation model in small number of freedoms, we expect natural modifications constrained by generation mechanism and clear understanding of modification.

As well known, in a command-response model, an F0 contour is expressed as equation (1) a sum of phrase components and accent components.

$$\begin{aligned} \ln F_0(t) = & \ln F_{\min} + A_p G_p(t - T_0) \\ & + [A_{a1} \{G_t(t - T_1) - G_t(t - T_2)\}] \\ & + A_{a2} \{G_t(t - T_3) - G_t(t - T_4)\} \end{aligned}$$

where G_p , G_t , F_{\min} , A_p , A_a , T_0 , T_1 , T_2 , T_3 and T_4 correspond to phrase component, accent component, bias level, the magnitude of phrase command, the magnitude of accent command, the onset time of phrase, the onset and the reset time of accent command. In the prosody modification of

this experiment, we extract all of these parameters from Mandarin speech and modify F_{\min} , A_p , A_a and duration only to add communicative component extracted from Japanese in these F0 control parameter values.

2.3. Communicative F0 component generation

In the proposed scheme of communicative prosody generation, communicative component is computed from input word impression vector. Though we have already proposed a training scheme for the mapping from impression vector to F0 control parameters using a neural-network for one-word utterance of “n”[6], it still needs further improvement for real-use. For the current study, we simply modify Mandarin speech by the average value of F0 control parameters and duration obtained from Japanese one-word utterances of “n” for the corresponding impression. We believe that this approximation by the average values would be acceptable for the current use though they are not ideal.

3. Experiments on communicative prosody generation and their perceptual evaluation

3.1. Mandarin speech samples

For the experiments of generation and perception test of communicative Mandarin speech, we recorded ten read-style Mandarin words uttered by one native Chinese. They consist of one to four syllables which gave impressions corresponding to three dimensions of two opposite directions (*confident-doubtful*, *allowable-unacceptable* and *positive-negative*). Table 1 shows the information of these ten word, targeted communicative prosody, Chinese pinyin with tone, English translation and a three-dimensional word impression vector together with a three-dimensional impression vector of synthetic speech obtained as perceptual experiment results. A three-dimensional word impression vector was obtained by dimension reduction through MDS for twenty-six-dimensional impression scores for input lexicon. To confirm that these phrases have the same impression that we originally intended, their lexical impression were scored from -3 (not at all) to +3 (very much) in twenty six impressions words. The average score of eight native Chinese adults (six male, two female) were calculated for each factor.

Table1: Comparison between input impression from lexicon and perceived impression
(Shown three-dimensional vector components were obtained through MDS analysis from 26 dimensional expressions)

No.	Prosody	Chinese phrase	English meaning	Word literal impression			Speech perceptual impression		
				dimension1	dimension2	dimension3	dimension1	dimension2	dimension3
1	Confident	jue2dui4	absolutely	2.91	-10.55	1.20	6.09	-10.75	14.62
2	Doubtful	shuo1bu2ding4	may be	-0.94	-4.89	-5.98	-1.03	-0.29	-1.92
3	Allowable	zan4cheng2	agree	2.42	4.37	-15.22	13.76	8.79	-6.27
4	Unacceptable	bu4hao3	not good	8.43	2.82	-6.70	2.31	0.70	-6.68
5	Positive	you3qu4	interesting	-10.47	13.50	-1.93	-1.83	1.36	2.63
6	Negative	ku1zao4	boring	21.45	-4.69	3.20	1.88	-0.19	0.50
7	Positive	shuang3	felling well	8.85	-8.88	9.35	-3.14	-6.86	8.72
8	Positive	shuang3a	felling very well	-6.52	-3.32	14.13	-5.53	-1.93	12.32
9	Positive	ting3gao1xing4a	sound fun	2.94	10.54	-8.29	-2.75	-2.59	-11.90
10	Negative	mei2jin4	uninteresting	-12.25	-2.88	17.17	-13.86	-2.17	-7.93

Table2: Modification for communicative prosody generation

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

3.2. Communicative Mandarin prosody generation

For communicative speech synthesis from the input impression vector, communicative prosodies of Japanese utterance “n” were added to Mandarin speech samples. In the modification of communicative prosody for Mandarin speech, the parameter of Fmin, Ap, Aa and duration were first extracted from read Mandarin speech samples and modified the prosodies for three dimensional impressions (*confident-doubtful, allowable-unacceptable and positive-negative*) with using the values in Table2. The original phrase utterances were prepared to be uttered in a reading style by a Chinese male and recorded in a quiet environment. The speech samples were synthesized using STRAIGHT speech synthesis [8] by changing prosody only.

3.3. Perceptual evaluation experiment

To evaluate how much the synthesized speech samples can provide communicatively adequate natural impression perceptual listening test was carried out. First, we asked a group of participants put scores ranging from -3 (very unnatural) - 3 (very natural) to each speech samples to judge if the synthesized speech were natural as utterance in daily conversation. Next, the participants were asked to put scores ranging from -3 (not at all)- 3(very much) to evaluate communicative impressions in twenty six dimensions for synthesized speech samples. The participants consisted of eight adult native Chinese speakers(six male and two female).

4. Experimental results

The result of the naturalness evaluation test shows that 85% speech sample synthesized with the prosodies derived from the input lexicon impression was judged as natural. And all of ten synthesized speech samples were used in communicative impression evaluation test. In order to see how the directions of impression vectors of input lexicon and synthesized speech sample in three-dimensional spaces are placed, each impression vector was obtained based on the subjective impression scores by twenty six impression words using MDS. The results are shown in Table 1.

To see the difference between the impression vectors of input word literal impression and perceptual impression of output synthesized speech, the correlation was calculated. As shown in Figure 2, the correlation values of unnatural speech samples judged in naturalness evaluation are lower than the naturalness samples. That is, the appropriate communicative prosodies could be decided depending on the degree of the subjective impression scores of input phrase. The inner product values were also calculated to see the similarity of the directions of impression vectors of input lexicon and perceptual impression of output speech samples. As shown in Figure 3, values appeared to be high for the speech sample. The inner product value turned out to be high for the speech sample where prosodies impressions are the same with the input phrases. Only few lexical impression vectors did not match to corresponding perceptual impression vectors because the lexicons of No.9 and No.10 have various prosodies to express the same impressions.

The comparison of lexical impression vector and perceptual impression vector shows the direct correspondences of input word impression to output speech prosodies. These distributions clearly show the adequacy of communicative prosody control with lexicon derived prosody.

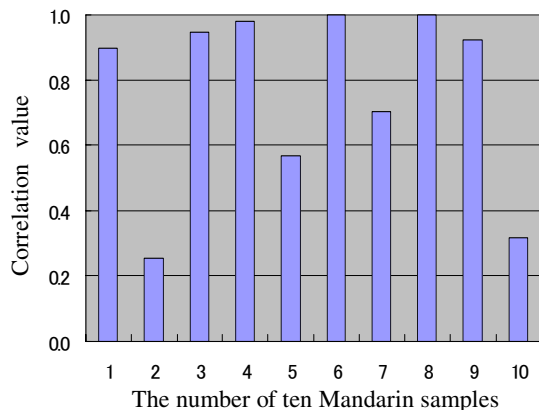


Figure2. Correlation between word lexicon impression and speech perceptual impression

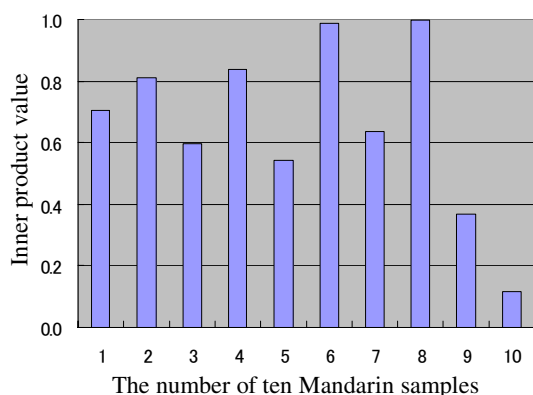


Figure3. Inner product between lexical impression vector and perceptual impression vector

5. Conclusions

We synthesized communicative Mandarin speech using prosodic characteristics of communicative Japanese speech. To evaluate communicative naturalness of prosody control, synthesized speech was perceptually evaluated in twenty-six dimensional impressions by native listeners. Perceptual evaluation test showed that the proposed communicative prosody generation scheme could provide natural Mandarin speech and high correlations between lexical impressions of Mandarin words and perceived prosodic impressions of corresponding synthesized speech. These experimental results confirmed the possibility of language universal control of communicative prosody. We would like to pursue complete generation scheme for longer and more complex sentences together with more language pairs to confirm the generality of the proposed scheme.

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 supported in part by the Grant-in-Aid for Scientific Research (B) No.18300063, JSPS.

6. References

- [1] Sagisaka, Y., Yamashita, T. and Kokenawa Y., "Generation and perception of F0 markedness for communicative speech synthesis" *Speech Communication* 46 pp. 376-384 2005
- [2] 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*, pp.273- 276, Mar.2005
- [3] Greenberg, Y., Tsuzaki, M., Kato, K., and Sagisaka, Y., "Communicative speech synthesis using constituent word attributes", *Proc. Interspeech2005*, pp.517-520, Sep.2005
- [4] Sagisaka, Y., Yamashita, T., and Kokenawa, Y., "Speech Synthesis with Attitude" *Proc. Speech Prosody 2004*, pp.401-404, 2004
- [5] Greenberg, Y., Tsuzaki, M., Kato, K., and Sagisaka, Y., "A trial of communicative prosody generation based on control characteristic of one word utterance observed in real conversational speech", *Proc. Speech Prosody 2006*, pp.37-40, May 2006
- [6] Li K., Greenberg, Y., Campbell N. and Sagisaka Y., "On the analysis of F0 control characteristics of nonverbal utterances and its application to communicative prosody generation" *Nato book* (to appear)
- [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, pp.233-242, 1984.
- [8] Kawahara, H., Masuda-Katsuse, I. and Cheveigné, 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