Analysis of communicative phrase prosody based on linguistic modalities of constituent words

Kazuma Takada
Pure&Applied Math
Waseda University
Tokyo, Japan
kazuma.takada2020@gmail.com

Hideharu Nakajima
NTT Media Intelligence Labs.
NTT Corporation
Kanagawa, Japan

Yoshinori Sagisaka
Pure&Applied Math
Waseda University
Tokyo, Japan
ysagisaka@gmail.com

***Abstract*—In this paper, phrase prosody is analyzed based on linguistic modalities of constituent words for communicative speech synthesis. Since Japanese final particles and auxiliaries play crucial roles to indicate speaker's intention and attitudes as modality differences, Japanese phrase sets showing different degree of the speaker's judgment were employed. Communicative/reading speech data were compared over 5 kinds of modality of epistemic judgment (uncertainty of what the speaker said) and 8 kinds of modality of evaluative judgment (what the speaker wishes listeners to be). These modality differences were quantified in 6-point Semantic Differential (SD) scales. The corresponding phrase communicative/reading prosody differences were measured by the F0 rising in the phrase final mora. Statistical analysis showed negative correlation value between F0 rising in the phrase final mora and SD about judgment only in communicative prosody but not in reading prosody. These results support the communicative prosody control possibilities from the modality information embedded in constituent words.**

***Index Terms*—communicative speech prosody, modality, Semantic Differential (SD)**

I. Introduction

Natural speech contains rich information which has not yet been thoroughly analyzed in conventional studies. Following the traditional typological classification [1], information embedded in speech is classified in three different categories referred as linguistic, paralinguistic, and nonlinguistic information. Boosted by the increasing needs of speech data analyses in the real fields, research efforts have been devoted to so called "expressive speech" or "emotional speech" to study paralinguistic characteristics [2]–[6]. Those efforts have been supported by many researchers and many related workshops and research challenges have been carried out [7], [8]. However, till now, there is neither an established description scheme nor systematic analysis for scientific understanding of the differences between the reading-style speech and communicative speech. For this reason, it is impossible to synthesize even simple communicative speech such as "Thank you very much" or "I am terribly sorry" with natural prosody observed in ordinary conversations, not with reading-prosody current text-to-speech synthesis systems can provide.

In the conventional study on expressive speech, as overall speaking characteristics have been mainly studied [2]–[8], utterance-dependent communicative prosody as shown in the above examples has not yet been studied intensively. To cope with this problem, we have been analyzing the differences between the reading style prosody and communicative one based on words constituting each utterance [9]–[11] instead of overall para-linguistic factors. Through these studies, we have found strong correlations between communicative prosody and impressions derived from constituent words [10], [11]. Furthermore, we could have shown the possibility of communicative prosody calculation using multi-dimensional impressions based on constituent words for some sentence sets [11].

In this paper, aiming at further generalization of communicative prosody computation based on linguistically well-defined factors, we have analyzed communicative prosody by focusing on linguistic modality. By focusing on modality showing judgement, systematic communicative prosody control has been observed in 52 Japanese short phrases uttered by 25 speakers. In the following sections, in Section II, we explain Japanese phrase final prosody and linguistic modality studied in phonetics and linguistics as background knowledge of this study. In Section III, we explain experimental setups for communicative prosody and their impression measurement. Section IV describes the experimental results showing consistent control characteristics of communicative prosody based on their impressions supplied by constituent words. Finally, the obtained results are summarized in Section V as conclusion.

II. PHONETIC AND LINGUISTIC BACKGROUND FOR THE SELECTION OF JAPANESE PHRASES AND PROSODY MEASUREMENTS

A. Information appeared in the prosody of final mora

In Japanese dialogue, it is well known that many variations of information appear at the final mora prosody. In particular, it is pointed out that the prosody of the final particle (discourse particle at the end of the phrase) is related to the meaning and function of the particle. Iwata et al. classified the F_0 shape of the final particle by clustering and showed that the utterance intention changes by each shape by perceptual experiment [12]. Oshima roughly classified the final particle /yo/ and /ne/ into four types by its intonation shape and considered the correspondence relationship between each intonation shape and the function of the final particle [13].

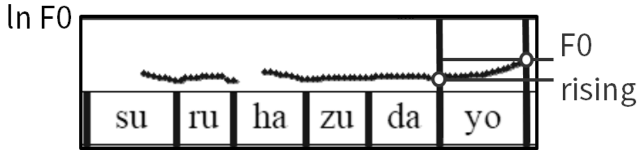

Fig. 1. F_0 rising at the interval of phrase final mora (calculated by $\ln F_0(t_e) - \ln F_0(t_s)$ [ln Hz], t_s[s]: when last mora started, t_e[s]: when last mora ended)

B. Linguistic modality

It is known that a speaker's subjective information, "linguistic modality", appears in Japanese sentence structure. Masuoka pointed out that Japanese sentences have nesting structures, consisting of grammatical layers: proposition, modality of judgment and modality of utterance in the order from inside [14].

(nesting structures)
[[[proposition] modality of judgment] modality of utterance]
(example: "(He) seems to go (there).")
[[[/ik/] /urashii/] /yo/]
[[(go)] (sound)] (enhances the speaker's opinion)]

This means the speaker's subjective information appears at constituent words at the end of a phrase. Also, the prosody of final mora changes corresponding to discourse.

Therefore, it is expected that communicative speech prosody in the phrase final mora can be estimated by subjective information appeared in phrase final constituent words. This expectation has not been pointed out nor confirmed in previous studies. In this study, to confirm the expectation, firstly we analyzed correlation value between impressions given by modality of judgment and F_0 at the phrase final mora.

III. Experimental Setups for Communicative Prosody and Their Impression Measurement

A. Overview of methods

We analyzed the correlation value between lexical impressions given by modality of judgment and F_0 at the phrase final mora. F_0 rising at the phrase final mora was employed as a prosody feature. Lexical impressions were measured by Semantic Differential (SD) scale method [16].

B. Phrases used for analysis

Table I shows 52 phrases employed for the analysis. They consist of a verb, postpositional words working as a modality of judgment, and words working as a modality of utterance. All modalities of judgment shown in [15] which can be followed by last particles were chosen for the phrases.

To confirm whether other elements influence prosody and impressions, control experiments were also carried out. First, two last particles, /yo/ and /ne/, were used to confirm whether modality of utterance influence the impressions of the whole phrase. Second, two different accent-typed verbs, /sur/ (type 0, "do") and /tor/ (type 1, "take"), were used to confirm whether the accent type affects the F_0 rising.

TABLE I
Phrases used for analyses (English translations or explanation in " " and pronunciation in / /)

phrase form: verb + modality of judgment + modality of utterance	
verb (2 words)	
"do"	/sur/ (type 0)
"take"	/tor/ (type 1)
modality of judgment (13 words)[a]	
modality of epistemic judgment (uncertainty of what the speaker said)	
"may"	/ukamoshirenai/
"must"	/unichigainai/
"look like"	/umitaida/
"sound"	/urashii/
"should"	/uhazuda/
modality of evaluative judgment (ideal of the speaker)	
"ought to"	/ubekida/
"only have to"	/ebaii/
"had better"	/uhougaii/
"must"	/anakyadameda/ [b]
"just have to"	/ushikanai/
"can"	/itemoii/ [c,d]
"not have to"	/anakutemoii/ [b]
"must not"	/ichadameda/ [c,d]
modality of utterance (2 words)	
"a particle which enhances the speaker's opinion [17]" /yo/	
"a particle which shows speaker's respect for listeners [17]" /ne/	

[a] All modalities of judgment shown in [15] which can be followed by last particles were chosen for the phrases.
[b] if /sur/ precedes this, /a/ in it is dropped out and /sur/ changes to /shi/.
[c] if /sur/ precedes this, /i/ in it is dropped out and /sur/ changes to /shi/.
[d] if /tor/ precedes this, /icha/ in it changes to /Qcha/ and /tor/ changes to /to/.

C. Measurement of prosody characterization

F_0 rising at the phrase final mora was measured as a prosody difference as shown in Figure 1. It was calculated by the formula below.

$$\ln F_0(t_e) - \ln F_0(t_s) \quad \text{[ln Hz]} \tag{1}$$

(t_s[s]: when last mora started, t_e[s]: when last mora ended)

Speech samples of Table I were collected from 25 participants (19 males and 6 females) who were native Japanese speakers from 18 to 24 years old. As a control experiment, the same analysis was carried out with reading speech. Communicative speech samples were recorded by asking the speakers to utter as if they were talking to their friends. While reading speech ones were collected by asking to speak as a newscaster quoting someone's words. F_0 was extracted by Praat, and voice samples whose last mora were unvoiced were removed by hand.

D. Evaluation of lexical impressions of phrases

Subjective impressions of the phrases shown in Table I were quantified by 6-point SD scale. To measure the subjective impressions about judgment, three impression pairs

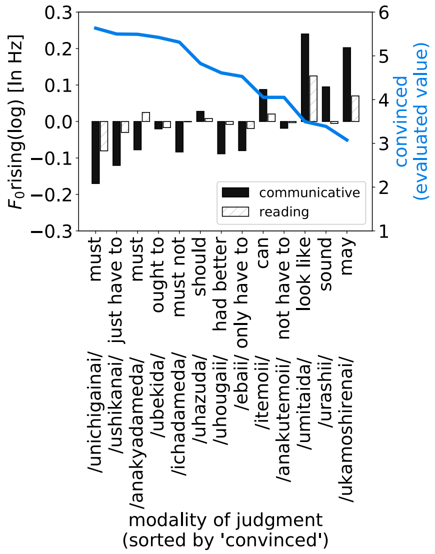

Fig. 2. The evaluated SD value of the subjective impression (line) and F_0 rising at the phrase final mora (bar) (modality of utterance: /yo/, participants: 25, scale of SD values: 6-point scale, impression: "convinced". Impression "assertive" and "advising" also show similar tendency.)

about speaker's judgment, "convinced-doubtful", "assertive-unassertive", "advising-not advising" were employed. Each impression of each phrase was evaluated by 6-point SD values. Participants were same as experiment III-C. To obtain only the impressions given by linguistic modalities, participants evaluated SD by looking at phrases written on a screen so that impressions given by the corresponding speech prosody did not affect the results. In addition to the above three dimensions relating to judgment impressions, we have also employed subjective impressions widely used for SD method (10 kinds: dull-keen, thick-thin, round-sharp, heavy-light, sober-flashy, soft-solid, dark-bright, rough-smooth, dirty-beautiful, muddy-pure, which were used for monosyllabic impression evaluation by Isonaka et al. [18]) to find the possibility of general prosody control. Twelve Japanese native speakers from 18 to 24 years old participated in this lexical impression experiment.

IV. Experimental Results

A. Correlation value between F_0 rising in the phrase final mora and the SD values of subjective impressions

Fig. 2 shows the comparison between communicative prosody and reading one in relation to the modality of judgment. As shown in the figure, F_0 rising at the phrase final mora of communicative speech varies consistently depending on the modality of judgment. The negative correlation value is observed between F_0 rising in the phrase final mora of communicative speech and the SD values of impression "convinced". On the other hand, the variations of the F_0 rising of the corresponding reading speech is small. The correlation value between the F_0 rising in communicative speech and the SD values of the impression "convinced" is significantly higher than that of reading speech value (at the 5% significant level).

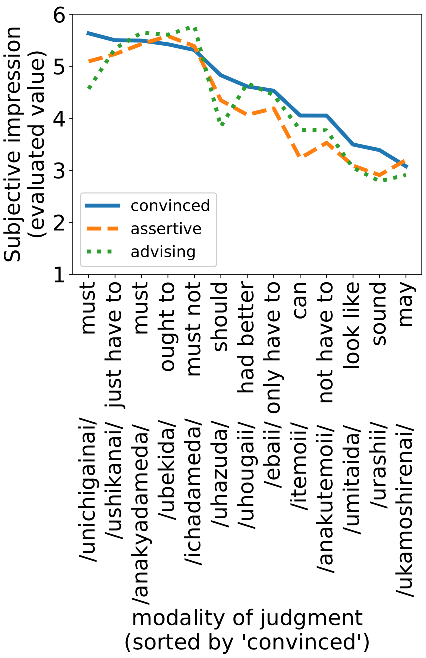

Fig. 3. The evaluated SD values of the subjective impressions about judgment (modality of utterance: /yo/, participants: 25, scale of SD values: 6-point scale)

The observed F_0 rising characteristics consistent to impressions of judgment suggest the possibilities of communicative speech prosody control based on the modality of constituent words in the same way as our previous research [9]–[11].

B. Correlation value between subjective impressions about judgment and F_0 rising in the phrase final mora

Fig. 3 shows SD values of 3 subjective impressions about judgment given by each modality of judgment. As shown in this figure, the other subjective impressions about judgment, "assertive", "advising" had similar SD value to that of "convinced". As shown in Table II, negative correlation values are observed between F_0 rising in the phrase final mora and all subjective impressions about judgment. Same as impression 'convinced', the correlation values between the F_0 rising in communicative speech and the SD values of impression 'assertive' and 'advising' are significantly higher than those of reading speech values (at the 5% significant level). This indicates that impressions about judgment given by modality of judgment involve phrase final F_0 lowering.

As Fig. 2 shows, F_0 in the phrase final mora tends to rise when modality of epistemic judgment appears in the phrase. While it tends to fall when evaluative judgment appears. These prosody characteristics imply that F_0 lowering in the phrase final mora is given by strength of the speaker's opinion known from the modality of constituent words.

TABLE II

Correlation value between F_0 rising at the phrase final mora and evaluated SD values (6-point scale) of each subjective impression about judgment (com: communicative, read: reading, *: the correlation value is significant at the 5% significant level, participants=25)

condition	modality of utterance	/yo/				/ne/			
	verb	/sur/		/tor/		/sur/		/tor/	
	speaking style	com	read	com	read	com	read	com	read
subjective impressions about judgment	convinced	-0.27*	-0.03	-0.38*	-0.14*	-0.04	-0.12	-0.14*	-0.26*
	advising	-0.25*	-0.03	-0.41*	-0.10	0.06	-0.05	0.06	-0.02
	assertive	-0.24*	-0.05	-0.42*	-0.08	0.03	-0.03	-0.04	-0.11

TABLE III

Multiple regression model coefficient of evaluated SD values of subjective impressions about judgment using that of general impression [18] (participants=12)

impressions about judgment	general impressions widely used for SD method [18]										intercept	$\overline{R}^2$	MSE
	keen	pure	dark	dirty	sharp	heavy	solid	flashy	thick	rough			
convinced	0.33	0.07	0	0.06	0	0	0.67	0	0	-0.09	-0.53	0.74	0.026
assertive	0.67	0.05	0	0.15	0	0	0	0	0.10	0	-0.26	0.49	0.019
advising	0.54	0.15	0	-0.08	0	0.20	0.14	0	0.01	0	-0.41	0.61	0.051

C. Influence of modality of utterance on F_0 rising in phrase final mora

F_0 rising in the phrase final mora highly depends on the kind of modality of utterance. As shown in Fig. 4, when final particle /yo/ is used as modality of utterance, F_0 rising in the phrase final mora varies depending on modality of judgment. On the other hand, when /ne/ is used, F_0 rising barely changes. For that reason, the correlation values between the F_0 rising with final particle /yo/ and the SD values of 3 impressions about judgment are significantly higher than those with final particle /ne/ (at the 5% significant level). As pointed out in linguistic studies, /yo/ exaggerates the speaker's opinion and /ne/ shows the speaker's agreement with listeners [17], which nicely derives the prosody characteristics control of /ne/ weakens the F_0 change which modality of judgment affects.

D. Relation between accent type of verb and F_0 rising

Regardless of accent type, F_0 rising in phrase final mora and evaluated SD values of the 3 impressions about judgment show negative correlation values as shown in Table II. On the other hand, the correlation values between the F_0 rising with verb"/tor/" (accent type 1) and the SD values of impressions about "advising" and "assertive" are significantly higher than those with "/sur/" (accent type 0) at the 5% significant level. The correlation value of impression "convinced" shows no significant difference between the 2 verbs. The difference of correlation value was seemingly caused by features of accent types as follows; type 1 verb raises accent lowering so that the F_0 at phrase final mora tends to be lower than that of type 0 verb, which allows the F_0 at phrase final mora to move more dynamically.

E. Specification of impressions about judgment by general impressions

Table III shows the result of expressing the subjective impressions about judgment by a multiple regression model, using the general subjective impressions as explanatory variables. The explanatory variables were chosen so as to minimize Akaike's Information Criterion (AIC). As a result of performing the cross-validation method (k = 10) using the above explanatory variables, the mean square error was sufficiently small as about 0.05 in any impression about judgment. This indicates that impressions about judgment can be estimated from general impressions, and F_0 rising can be specified only by the general impressions.

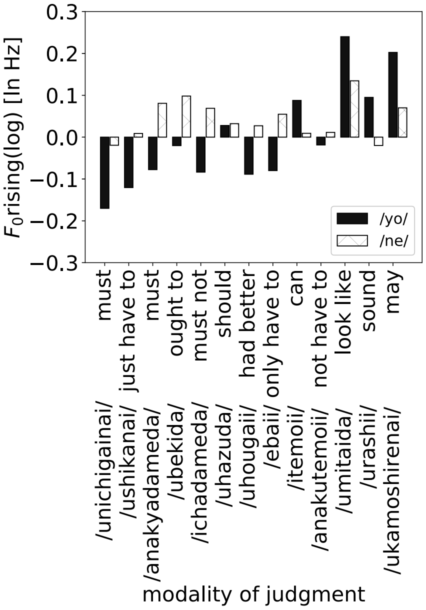

Fig. 4. F_0 rising in the phrase final mora by using last particle /yo/ or /ne/ (participants: 25)

V. Conclusions

Aiming at further generalization of communicative prosody computation based on linguistically well-defined factors, we have analyzed communicative prosody by focusing on linguistic modality. F_0 rising at the phrase final mora was employed as a prosody feature. SD values of subjective impressions given by constituent postpositional words working as modality of judgment were used as a feature of the lexical impression. The analyses showed the following characteristics of communicative prosody.

- There were negative correlation values between F_0 rising at the phrase final mora and three subjective lexical impressions about judgment given by modality of judgment (Table II).
- The correlation values between F_0 and the subjective impressions of judgement were observed only in communicative speech but not in reading speech (Fig. 2).
- Modality of utterance influences magnitude of F_0 rising at the phrase final mora (Fig. 4).
- The accent type of the verb in the phrase does not affect F_0 rising in the phrase final mora (Table II).
- Subjective impressions about judgment can be estimated from that of general impressions (Table III).

The above observations show the possibility of communicative prosody control using impressions of lexicons constituting a phrase. Based on these results, we are planning to build a communicative speech prosody control model together with a dictionary with lexicons having subjective impression information.

References

[1] H. Fujisaki, "Information, prosody, and modeling – with emphasis on tonal features of speech –," in *Proc. Speech Prosody*, 2004, pp. 1–10.

[2] M. Schröder, "Emotional speech synthesis: A review," in *Proc. EUROSPEECH*, 2001, pp. 561–564.

[3] D. Erickson, "Expressive speech: Production, perception and application to speech synthesis," *Acoust. Sci. & Tech.*, Vol. 26, pp. 317–325, 2005.

[4] N. Campbell, W. Mamza, H. Höge, J. Tao and G. Bailly, "Special section on expressive speech synthesis," *IEEE Trans. Audio Speech Lang. Process.*, Vol. 14, pp. 1097–1098, 2006.

[5] ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion, http://www.isca-speech.org/archive_open/speech_emotion/

[6] Y. Yamashita, "A review of paralinguistic information processing for natural speech communication," *Acoust. Sci. & Tech.*, Vol. 34, Issues 2, pp.73–79, 2013.

[7] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller and S. Narayanan, "The INTERSPEECH 2010 Paralinguistic Challenge," in *Proc. INTERSPEECH*, 2010, pp. 2758–2761. http://emotion-research.net/sigs/speech-sig/paralinguistic-challenge

[8] http://emotion-research.net/sigsligs/detail/2

[9] Y. Sagisaka, T. Yamashita, and Y. Kokenawa, "Generation and perception of F0 markedness for communicative speech synthesis," *Speech Communication*, Vol. 46, Issues 34, pp. 376–384, 2005.

[10] Y. Greenberg, M. Tsubaki, H. Kato, and Y. Sagisaka, "Analysis of impression-prosody mapping in communicative speech consisting of multiple lexicons with different impressions," *Oriental-COCOSDA* (CD-ROM), 2010.

[11] S. Lu, Y. Greenberg, and Y. Sagisaka, "Communicative F0 generation based on impressions," *5th IEEE Conference on Cognitive Infocommunications*, pp. 115–119, 2014.

[12] K. Iwata and T. Kobayashi, "Expression of speaker's intentions through sentence-final particle/ intonation combinations in Japanese conversational speech synthesis," *SSW*, pp. 235–240, 2013.

[13] D. Y. Oshima, "On the Relation Between the Intonation Types and the Functions of Discourse Particles in Japanese," *Forum of international development studies*, 43, pp. 47–63, 2013 (in Japanese).

[14] B. Pizziconi and M. Kizu, *Japanese Modality*: Exploring its Scope and Interpretation Palgrave Macmillan, 2009, pp. 36–55.

[15] T. Masuoka, "*Nihongo Modariti Tankyuu*" [Investigations of Japanese Modality]. Kuroshio Shuppan, 2007 (in Japanese).

[16] Osgood C.E., "The nature and measurement of meaning," *Psychological Bulletin*, Vol. 49, No. 3, pp. 197–237, 1952.

[17] DY. Lee, "Involvement and the Japanese interactive particles ne and yo Research article," *Journal of Pragmatics*, Vol. 39, Issues 2, pp. 363–388, 2007.

[18] Y. Isonaka, Y. Kanno, Y. Sagisaka, and K.watanabe, "Perceptual Impressions from Timbres of Japanese Single Syllables," *Reports of the spring meeting the Acoustical Society of Japan*, 2-5-9, 2015 (in Japanese).