Sentence-Final Prosody Analysis of Japanese Communicative Speech Based on the Command-Response Model

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

Hideharu Nakajima
NTT Communication Science Labs.
NTT Corporation
Kyoto, Japan

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

Abstract

Aiming at communicative speech synthesis, we analyzed sentence-final prosody characteristics through subjective impression on constituting lexicons. Since Japanese sentence-final particles and postpositionals are expected to be employed to generate communicative prosody showing speaker's intention and attitudes, we designed 52 single-phrase utterances showing different strength of the speaker's impressions about judgment. These impressions were quantified in Semantic Differential (SD) scales. F0 contour characteristics were analyzed by using the command-response model. To cope with sentence final F0 characteristics, an additional accent command was introduced for F0 rise and drop of sentence-final particles. The analysis showed systematic communicative prosody control by the accent command reflecting effect of judgment impressions which can be obtained from constituting lexicons. These results indicate possibility of sentence-final prosody control using impression obtained from lexicons constituting output sentences.

Keywords— *speech synthesis, communicative speech, Semantic Differential (SD), Command-Response Model, linguistic modality*

I. INTRODUCTION

Prosodies in communicative speech have wider variations than those in reading-style speech. In speech engineering field, some of them have been studied as para-linguistic or expressive speech prosody including emotional one [1,2,3,4]. Most of these studies have focused on predetermined speech categories such as emotional ones for speech analysis and synthesis. In real-field communications, there exist much wider variations which cannot be treated by pre-determined variation categories, which requires detailed analyses between constituting lexicons and communicative prosody [5].

Recently, communicative prosody variations have been analyzed for Japanese human interactions from data analysis viewpoint and sentence-final prosodic characteristics have been analyzed [6,7]. These sentence-final prosody variations in Japanese speech have already been studied by phoneticians relating to their linguistic attributes [8]. These studies in speech engineering and phonetics support the control possibilities of communicative prosody using lexicons constituting sentence-final parts.

For sentence-final prosody, we have analyzed single phrase utterances consisting of a verb, postpositionals working as a modality of judgment, and final particles working as a modality of utterance [9] as shown in the top row (phrase form) of Table I. Throughout this analysis, we could have found the correlation between sentence-final lexicons and F0 drops in the final mora.

In this paper, we have analyzed sentence-final prosody of these single phrase utterances using the command-response model [10] to understand the prosody control characteristics more clearly and to confirm quantitative control possibilities from constituting lexicons. In the following sections, in Section II we introduce our previous studies on the correspondence between sentence-final prosody in communicative speech and impressions obtained from constituting particles and postpositionals. Section III describes the analysis method we employed; the introduction of the command-response model (what is called Fujisaki Model), newly introduced parameters for sentence-final F0 control, and the impression measurement criteria. Section IV describes the experimental results of model parameter characteristics in communicative speech. Finally, we sum up the findings in Section V.

II. BACKGROUNG

Aiming at communicative speech synthesis, we have been studying the differences between communicative prosody and reading-style one based on constituting lexicons of the utterance [11,12,13]. Through these analyses, we have found strong correlation between F0 height and the strength of degree adverbs [11]. Furthermore, based on the correlations between F0 shape and impressions found in short utterances [12], we have shown a possibility of communicative speech computation using multi-dimensional impressions obtained from constituting lexicons [13].

In the most recent study, we found the correlation between Japanese sentence-final communicative prosody and impressions given by constituting lexicons [9]. In Japanese sentences, speaker's subjective information what is called linguistic modality appears in grammar structure. Masuoka pointed out speaker's belief or assertion of what he says appears at the end of the sentence [14]. For example, Japanese particles and auxiliaries at the sentence-final positions show speaker's judgment on what he says expressing how probable (i.e. /kamoshirenai/ ("may")) or how obligatory (i.e. /nakyadameda/ ("must")). The final particle shows speaker's attitude expressing confirmation or emphasis (/ne/ or /yo/). Using utterances containing these postpositionals and final particles referred in [14] (Table I), we analyzed the correlation between communicative prosody and subjective impressions given by the lexicons. Especially, to measure impression about speaker's judgment given by the modality of judgment, we selected 3 axes related to speaker's judgment, "convinced", "assertive", and "advising". The analyses showed weak negative correlations between F0 rising at phrase-final mora and the magnitude of impressions expressing the speaker's judgment, convinced, assertive, and advising, obtained from particles and auxiliaries [9]. The correlations were found only in communicative speech but not in reading speech. In this study, to understand sentence-final prosody scientifically, we employed the command-response model proposed by Fujisaki [10] to represent observed F0 contour as parameters associated with linguistical factors. Employing this F0 generation model, we tried to find F0 control characteristics through its model parameters in communicative speech.

III. CHARACTERIZATION OF JAPANESE COMMUNICATIVE FO CONTOUR

In this section, we first briefly explain well-known command-response model [10] and an additional introduced accent command for quantitative analysis of sentence-final prosody variation together with speech data employed for the analysis.

A. F0 Contour Model

The command-response model is known as F0 contour generation model relating to linguistical factors [10]. This model generates F0 contour as a sum of phrase component, accent component, and base F0 parameter F_{min}, shown in (1). α, β, and γ are constants typically $\alpha = 3.0$, $\beta = 20.0$, and $\gamma = 0.9$ respectively.

$$\ln F_0(t) = \ln F_{\min} + \sum_{i=1}^{I} A_{pi} G_p(t - T_{0i}) + \sum_{j=1}^{J} A_{aj} (G_a(t - T_{1j}) - G_a(t - T_{2j})) \quad (1)$$

$$G_p(t) = \begin{cases} \alpha^2 t\, e^{-\alpha t} & \text{for } t \geq 0 \\ 0 & \text{for } t < 0 \end{cases} \quad (2)$$

$$G_a(t) = \begin{cases} \min(1-(1+\beta t)\, e^{-\beta t}, \gamma) & \text{for } t \geq 0 \\ 0 & \text{for } t < 0 \end{cases} \quad (3)$$

In these equations, F_{min} represents F0 baseline. A_{pi} and T_{0i} represents magnitude and occurred time of phrase commands. A_{aj}, T_{1j}, T_{2j} represents magnitude, onset time, and offset time of accent commands. As shown in these equations, F0 contour is generated by quite small number of parameters which can be directly associated with prosody control factors given by the input sentence. For this reason, we adopted this model for our scientific F0 analysis.

B. Applying Additional Commands for Phrase Final F0 Control

Previous study showed sentence-final communicative prosody shape varies depending on lexical impressions [9]. From the analysis, weak negative correlation was observed between final particle (i.e. sentence-final mora) F0 rising in Japanese communicative prosody and lexical impression values obtained from postpositionals. To see and quantify the variation of sentence-final F0 shape affected by lexical impression, we introduced an additional command corresponding to sentence-final mora F0 contour. Fujisaki et al. showed final particle prosody can be

described as accent commands by dialogue prosody analysis [15].

Therefore, in this study, we represented sentence-final mora command as an accent command (hereinafter called Aalast, T1last, T2last), which shows local F0 control as shown in bottom layer of Figure 1. In Japanese utterances, accent commands are usually positive [16]. On the other hand, though the commands are usually positive as well in English, negative commands were used when speaker exaggerates para-linguistic information [16]. Since final mora F0 drop found in communicative prosody [9], we allow sentence-final command A_{alast} to be negative in the following analysis.

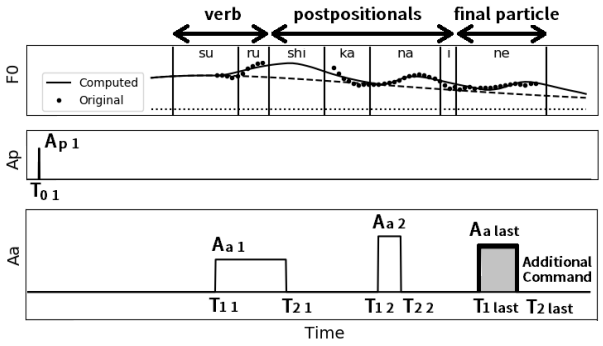


Figure 1. An additional accent command to control sentence-final prosody

TABLE I. PHRASES USED FOR ANALYSES (ENGLISH TRANSLATIONS OR EXPLANATION IN "")

Phrase form: Verb + Postpositionals + Final particle (*i.e. /suru shikanai ne/*)	
Verb (2 words)	
suru (type 0; accentless) "do"	
toru (type 1; head-accented) "take"	
Postpositional (13 words)	
working as modality of judgment; speaker's judgment about contents	
kamoshirenai "may"	*nichigainai* "must" (very likely)
mitaida "look like"	*rashii* "sound"
hazuda "should"	*bekida* "ought to"
rebaii "only have to"	*hougaii* "had better"
nakyadameda "must" (obligation)	*shikanai* "just have to"
temoii "can" (permission)	*nakutemoii* "not have to"
chadameda "must not" (prohibition)	
Final particle (2 words)	
working as modality of utterance; speaker's attitude or intention	
yo	*ne*

C. Data Sets for Experiments

Communicative speech utterances of 52 phrases with postpositionals showing different level of speaker's impression about judgment [9] were employed for the analysis. Table I shows the words employed in the phrases. These phrases consist of a verb, postpositionals working as modality of judgment, showing speaker's judgment, and a final particle working as modality of utterance, showing speaker's attitude [14]. Impression given by text of the constituting words is measured by Semantic Differential (SD) scale method [17]. Communicative speech samples were the ones uttered as the speaker talks to their friends. Their F0 contours were extracted by WaveSurfer and smoothened by simple moving average. Reading-style speech samples were also collected to compare with communicative speech. F_{min} parameters were treated as constant depending on speakers and speech styles (communicative or reading-style).

As natural communicative speech recording is quite difficult, we carefully asked all speakers to imagine real situation. Despite these considerations, as it is difficult to utter communicative speech naturally, it tends to be similar to reading speech. For that reason, we selected utterances of 10 speakers which show clear difference between communicative and reading prosody for the analysis. The difference of these prosodies was measured by F0 fluctuation range, which is residual from regression line of F0 contour.

IV. EXPERIMENTAL RESULTS

A. Controllability of Sentence-Final Communicative Prosody Based on Constituting Lexicons

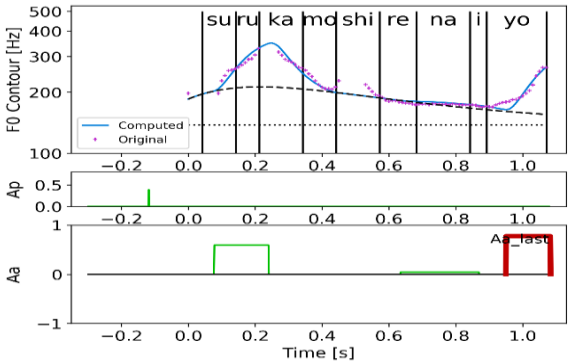
The analysis was carried out on the additional sentence-final commands in communicative prosody. The mean errors between measured F0 contours and computed ones were less than a semitone (communicative: 92 cents, reading: 80 cents). Not only positive but negative sentence-final commands were observed. As shown in the examples of Figure. 2a and 2b, sentence-final rising and falling F0 contours have been nicely approximated by the adopted sentence-final accent commands. These accent commands suggest local F0 control at the phrase-final mora occurs in communicative prosody. Also, these characteristics seem to be resulted from strong manifestation of speaker's judgment represented by "kamosirenai" (maybe) in Figure 2a and "nichigainai" (must be) in Figure 2b.

The sentence-final command magnitude A_{alast} values in communicative speech turned out to be significantly smaller than those in reading speech ($p < 0.05$). Especially, as shown in Figure 3, sentence-final

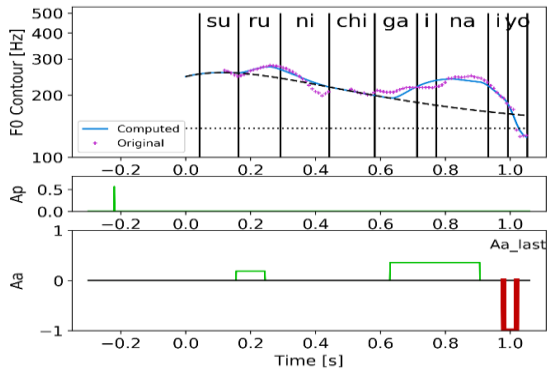
F0 drop characteristic was observed more frequently in the communicative speech (28.7%) than the reading one (17.7%).

B. Effect of Lexical Impression to Sentence-Final Communicative Prosody

In this analysis, we tried to find the possibilities to use impression obtained from lexicons for the control of communicative prosody. To focus on the lexicons constituting sentence-final parts showing judgment, we measured the magnitude of impressions about judgment ("convinced", "assertive" and "advising") obtained only from input lexicons. As we did not ask speakers to produce these speech samples with strict instructions, it is expected that individual sample and speaker may reflect factors other than impressions directly obtained from lexicons.



(a) An example of large F0 rising in speech with weak judgment (*/surukamoshirenaiyo/ (may do)*)



(b) An example of large F0 drop in speech with strong judgment (*/surunichigainaiyo/ (must do)*)

**Figure 2 .Model parameters of communicative utterances with sentence-final F0 rising (a) and lowering (b) (*bold line in accent command: sentence-*

Table II shows the correlation scores between A_{alast} and impression magnitudes about judgment obtained from constituting lexicons. As shown in the Table, weak correlations were observed in communicative prosody compared with the uncorrelated reading ones between A_{alast} and impression magnitude of constituting lexicons for the samples with sentence-final particle "yo". As Japanese final particle "yo" emphasizes speaker's judgment, these correlations imply these lexical impressions affect communicative prosody.

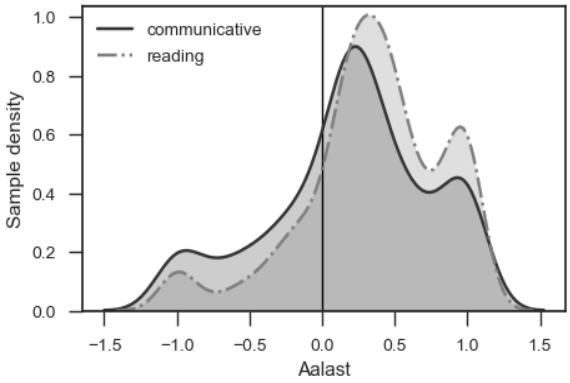


Figure 3. Distribution of Final mora command magnitude (A_{alast}) in communicative/reading speech samples

The smallness of the correlations even in communicative speech looks due to high freedom of judgment magnitude. To directly measure only the constituting lexical effect, we should strictly control the production context or apply listening-based experiments we have used in the previous study [11]. However, all correlations were minus and those in communicative prosody were larger than reading ones, which suggests the involvement of input lexicons in communicative prosody control.

TABLE II. CORRELATION COEFFICIENTS BETWEEN AALAST AND IMPRESSIONS OBTAINED FROM CONSTITUTING LEXICONS (WITH PARTICLE "YO", WHICH STRENGTHENS THE JUDGMENT IMPRESSIONS)

Correlation coefficients	Impression about judgment		
	Convinced	*assertive*	*advising*
Communicative	-0.201	-0.171	-0.241
Reading	-0.072	-0.076	-0.041

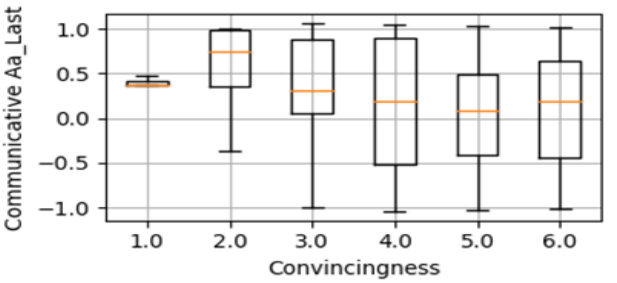


Figure .4 A_{alast} parameter within each impression value "convinced" for communicative speech samples with final particle "yo"

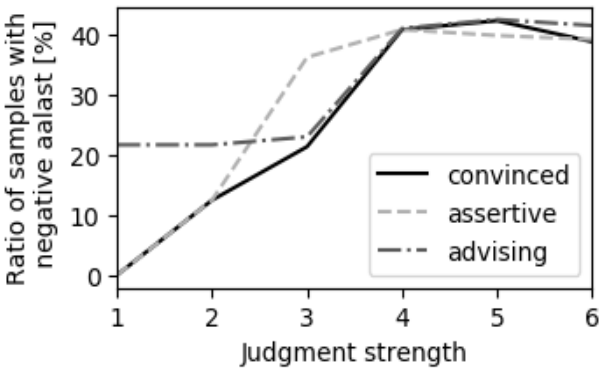


Figure 5. Ratio of samples whose A_{alast} are negative (*communicative speech samples with final particle "vo"*)

Figure. 4 shows values of A_{alast} parameters for each strength of judgment impression "convinced". Although majority of samples show positive sentence-final accent command, negative A_{alast} are also seen in utterances with strong convincingness. Figure 5 shows ratio of speech samples with negative A_{alast} command for each judgment impression strength. As shown in the figure, the negative A_{alast} command as exemplified in Figure. 2b were mostly observed in speech samples whose constituting lexicons show high scores for these three judgment impressions. In other words, the negative commands are only restricted to communicative speech samples with lexicons showing strong impression of judgment.

V. CONCLUSION

Aiming at the control of communicative prosody reflecting information obtained from input lexicons, we have analyzed F0 contours of communicative prosody using the command-response model by contrasting with reading-style prosody. For communicative speech samples consisting of single phrase with Japanese particles showing judgment magnitude ("convinced", "assertive", and "advising"), we could have observed their consistent control characteristics reflecting constituting lexical effects.

The observed control characteristics are summarized as follows.

- An additional sentence-final accent command can work systematically to express sentence-final prosody variety.
- Negative correlations were observed between final command magnitude and the all judgment impressions obtained from constituent lexicons.
- Negative control of final prosody is restricted to speech samples constituting lexicons with strong judgment.

We expect that these control characteristics enable to generate communicative speech prosody using lexical impressions. The smallness of correlations between communicative prosody and lexical impressions suggests more strict control is necessary for further analysis for computational modeling.

REFERENCES

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

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

[3] 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, 2006, pp. 1097-1098.

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

[5] Y. Sagisaka and Y. Greenberg, "Communicative Speech Synthesis as Pan-Linguistic Prosody Control", In Speech Prosody in Speech Synthesis: Modeling and generation of prosody edited by K. Hirose and J. Tao, Springer, 2015, pp.73-82.

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

[7] J. Venditti, K. Maeda and J. P.H. van Santen, "Modeling Japanese Boundary Pitch Movements for Speech Synthesis", ESCA/COCOSDA Workshop on Speech Synthesis, 1998, pp. 317-322.

[8] T. Koyama, "Bunmatsushi to Bunmatsu Intonation" [Sentence Final Particle and Its Intonation], In Bunpo to Onsei [Grammar and Speech], Kuroshio, 1997, pp. 97-119 (in Japanese).

[9] K. Takada, H. Nakajima, and Y. Sagisaka: "Analysis of communicative phrase prosody based on linguistic modalities of constituent words", Proc. iSAI-NLP, 2018, pp. 217-221.

[10] H. Fujisaki and K. Hirose: "Analysis of voice fundamental frequency contours for declarative sentences of Japanese", J. Acoust. Soc. Jpn. (E), 5, 1984, 233-242.

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

[12] 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 (CDROM), 2010.

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

[14] T. Masuoka, Nihongo Modariti Tankyuu [Investigations of Japanese Modality], Kuroshio, 2007 (in Japanese), in press.

[15] H. Fujisaki, S. Ohno, M. Osame, and M. Sakata, "Prosodic characteristics of a spoken dialogue for information query", ICSLP, 1994, pp. 1103-1106.

[16] H. Fujisaki, S. Ohno, and C. Wang, "A command-response model for F0 contour generation in multilingual speech synthesis", the 3rd ESCA/COCOSDA Workshop (ETRW) on Speech Synthesis, 1998, pp. 299-304.

[17] Osgood C.E., "The nature and measurement of meaning", Psychological Bulletin, Vol. 49, No. 3, 1952, pp. 197-237. G. Eason, B. Noble, and I. N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April.