SPEAKER'S INTENTIONS CONVEYED TO LISTENERS BY SENTENCE-FINAL PARTICLES AND THEIR INTONATIONS IN JAPANESE CONVERSATIONAL SPEECH

Kazuhiko Iwata, Tetsunori Kobayashi

Perceptual Computing Laboratory, Waseda University, Tokyo, Japan

ABSTRACT

We investigated listeners' perception of speaker's intention depending on sentence-final particles and their intonations in Japanese conversational speech in order to build a speech synthesis system that can express different intentions and subtle nuances. First, we clustered F0 contours derived from approximately 2000 sentence-final syllables and found the sentence-final F0 contours varied a great deal. Next, we selected six distinctive F0 contours that gave perceptually different intonations from among the cluster centroids, and subjectively evaluated synthesized sentence utterances that had various sentence-final particles and their intonations. Results showed that suitable combinations of a sentence-final particle and its intonation should be used to precisely convey the intention to the listeners, and whether the sentence was positive or negative also affected the listeners' perception of the intention.

Index Terms— speech synthesis, speaker's intention, sentence-final intonation, sentence-final particle, conversational speech

1. INTRODUCTION

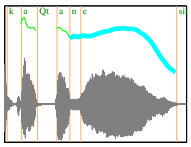
The demand for conversational synthetic speech has been increasing as spoken dialogue systems, such as humanoid robots and speech-enabled agents, are becoming more commonly used. Several approaches for synthesizing conversational speech have been reported to date [1, 2, 3]. However, due to the diversity of conversational speech, there are still many problems that need to be solved to establish effective conversational speech synthesis technology. It is important to take note of the various common expressions in human speech that depend on not only the emotion but also the intention, attitude, relationship with the listeners, and so on.

In the Japanese spoken language, a speaker's intention is represented at the end of a sentence by sentence-final particles or auxiliary verbs [4]. For example, *'Iku yo'*, a verb *'Iku'* ("go") followed by a particle *'yo'*, indicates a strong assertion, and *'Iku ne'* (*'Iku'* followed by *'ne'*) indicates a request for listeners' agreement. The functions of sentence-final particles have been extensively studied in the field of linguistics [5, 6, 7, 8]. Intonation shapes at the end of a sentence are known to vary depending on the speaker's intention [9, 10, 11], and also known to have an impact on perceptions of politeness [12]. The sentence-final intonations have significant roles in Japanese conversational speech. However, in the field of speech synthesis technology, there are not as yet many approaches to expressing the speaker's intention by controlling the sentence-final intonation.

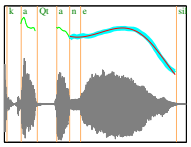
In order to produce conversational prosody, an F0 modeling technique based on prosodic labels using an X-JToBI labeling scheme has been developed [13]. Although this model could successfully produce intonations in conversational speech using the X-JToBI labels, the relationship between the X-JToBI labels and the speaker's intentions was not discussed. As for the expression of the speaker's intention or attitude, boundary pitch movements in Tokyo Japanese have been analyzed and modeled in relationship to semantic scales [14]. A prosody control method based on the analysis of intonations of the one-word utterance *'n'* is also well known [15]. However, none of these models above considered the association of the expressions of the speaker's intention with the sentence-final particles despite the fact that the intention is verbally expressed by the sentence-final particle in Japanese spoken dialogue. In a previous study, we investigated the relationship between the speaker's intention and sentence-final intonation [16]. However, we also did not considered the relationship with sentence-final particles.

In this study, we focus on the listeners' perception of speaker's intention associated with sentence-final particles and their intonations in order to enable conversational speech synthesis technology to express intention. In Section 2, we classified normalized sentence-final F0 contours extracted from sentence utterances to determine what kind of sentence-final F0 contours are used in actual speech. In Section 3, we selected several distinctive intonations from among the classification results. A subjective evaluation was conducted to reveal what kind of speaker's intentions could be conveyed by those intonations depending on the sentence-final particles. In Section 4, we conclude the paper with some remarks.

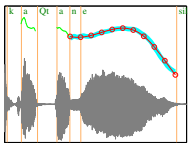
This work was supported in part by Global COE Program "Global Robot Academia" from the Ministry of Education, Culture, Sports, Science and Technology of Japan.



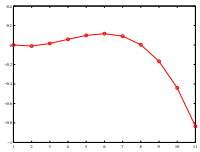
(a) F0 contour in the sentence-final syllable 'ne'(thick line)



(b) approximated F0 contour (thin line)



(c) sampled F0 contour (circles)



(d) normalized F0 contour

Fig. 1. Process of extracting normalized F0 contours.

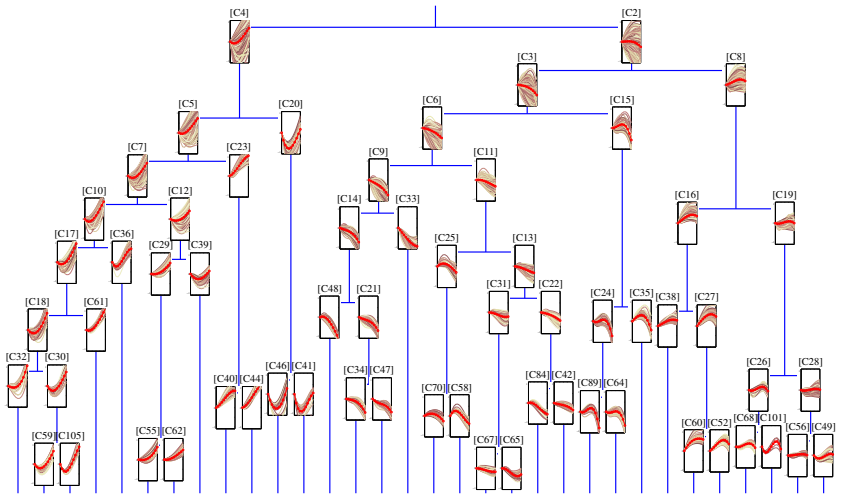


Fig. 2. Result of clustering sentence-final F0 contours when the number of clusters was set to 32.

2. CLASSIFICATION OF SENTENCE-FINAL INTONATIONS

2.1. Speech data

In this study, we used speech data that were created with the aim of developing an HMM-based speech synthesis system that had multiple HMMs depending on communication situations [3]. To build the HMMs, we designed several communication situations and more than 2000 sentences derived from dialogues that our communication robot [17] performed. These sentences were uttered by a voice actress assuming the situations. The F0 contours were extracted using STRAIGHT [18]. The intonations of these utterances varied and expressed subtle nuances and connotations. Of these data, 2092 utterances whose sentence-final vowel was not devoiced were used for the analysis.

2.2. Extraction of sentence-final F0 contours

The F0 contours of the sentence-final syllable were classified by one of the hierarchical clustering algorithms, Ward's method [19], which merges clusters so as to minimize the within-cluster variance. Because the actual F0 values of the utterances differed from each other and were difficult to classify, the time and frequency axes were normalized.

The normalization process is shown in Fig. 1. The vertical lines denote the phoneme boundaries. First, the F0 contour in the sentence-final syllable was extracted (the thick line in Fig. 1(a)) by referring to the phoneme boundaries. To remove F0 perturbation caused by jitter and microprosody, the logarithmic F0 contour was approximated by a third-order least squares curve (the thin line in Fig. 1(b)). The approximated curve was sampled at 11 points that equally divided the duration into 10 (the circles in Fig. 1(c)). Finally, the starting point of the sampled curve was parallel-translated to the origin, as shown in Fig. 1(d).

2.3. Classification by clustering

The normalized F0 contours obtained by the above process were classified by Ward's clustering using the difference values of the normalized F0 as 10th order feature vectors. Fig. 2 shows an example of the clustering results when the number of clusters was set to 32. The F0 contours denoted by thick circles and a thick line are the centroids of each cluster. The numbers in square brackets (e.g., [C2]) are expedient cluster IDs corresponding to the clustering sequence. Note that the lengths of the vertical lines do not represent the distance between the clusters due to the limitation of the page layout. Various sentence-final F0 contours, namely intonations, were found, including not only simple rising and falling intonations but also rise-fall and fall-rise intonations.

2.4. Perceptual discrimination of intonations by centroids

We found the sentence-final F0 contours were classified into distinctive clusters. However, we predicted not all the pairs of cluster centroids would have a notable perceptual difference from each other because the clustering was based only on the shapes of the F0 contours. Therefore, a preliminary evaluation was conducted. A back-channel utterance *'haa'*, which had no specific linguistic meaning, was resynthesized by the STRAIGHT vocoder [18], and its F0 contour was replaced with 127 centroid pairs obtained in the process of classifying the F0 contours into 128 clusters. 15 listeners were randomly presented with 254 *'haa'* pairs including a reverse order for each pair of centroids and then asked whether they perceived the two intonations as the same or different. The results of the evaluation are shown in Fig. 3. The numbers in parentheses

indicate the number of responses when the intonations by the centroids on both sides were perceived as different. This is how we obtained the criteria for sifting through and selecting the F0 contours that would be used in the next experiment.

3. ANALYSIS OF SPEAKER'S INTENTIONS CONVEYED BY SENTENCE-FINAL PARTICLES AND THEIR INTONATIONS

3.1. Selection of representative F0 contours

We consulted previous studies prior to conducting the subjective evaluation to investigate what kind of speaker's intentions could be conveyed by sentence-final particles and their intonations. In these studies, the sentence-final intonation contours were classified into two categories (rise and fall) [11, 20] or up to at most five categories (interrogative rise, prominent rise, fall, rise-and-fall, and flat) [10]. Therefore we decided to sift through and select around five F0 contours to be used among the cluster centroids. Referring to the results of the preliminary evaluation (Fig. 3), we stopped dividing clusters whose child clusters received 28 or fewer perceptions that their intonations were different. Then, the six centroids shown in Fig. 4 were ultimately selected. The centroids in Figs. 4(a) and 4(b) seem to correspond to the interrogative rise intonations, Fig. 4(c) to the fall, Fig. 4(d) to the rise-and-fall, Fig. 4(e) to the prominent rise, and Fig. 4(f) to the flat in the previous research [10]. These results indicate that specific intonation contours corresponding to the categories of sentence-final intonations in the previous research could be obtained from the speech data.

3.2. Subjective evaluation

A subjective evaluation was conducted to clarify what kind of intentions the listener could perceive through the sentence-final intonations produced by the selected centroids.

We prepared 31 short sentences consisting of a verb *'taberu'* ("eat") followed by a sentence-final particle (*'yo'*, *'na'*, *'ne'*, *'datte'*, etc.), an auxiliary verb (*'daroo'*), or one of their concatenations (*'yone'*, *'yona'*, *'datteyo'*, *'daroone'*, etc.). Synthetic voices of these sentences were generated by our HMM-based speech synthesis system [3]. The duration of the last vowel of each sentence was fixed to 313 ms, which was the mean duration of the last vowels of the sentences in the speech data. Then, the sentence-final F0 contour was replaced with the six centroids. We also designed 11 speaker's intentions ("request", "order", "blame", "hearsay", "guess", "question", etc.) and dialogue situations where these intentions could be indicated. We informed 20 listeners of the situations and speaker's intentions and asked them to evaluate on a five-level scale from –2 (unsuitable; suitable for a different intention) to +2 (suitable) whether or not both the lexical and intonational expressions of the stimulus were suitable for conveying the intention.

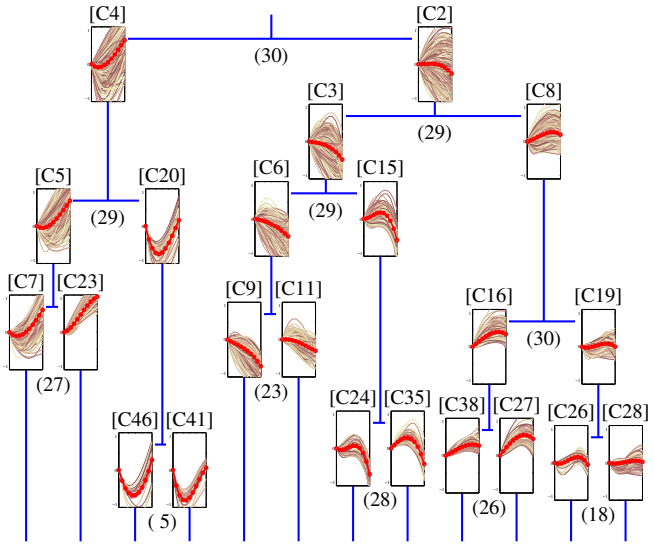
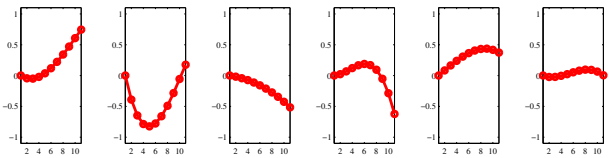


Fig. 3. Preliminary evaluation results of intonations generated by cluster centroids.



(a) C5 (b) C20 (c) C6 (d) C15 (e) C16 (f) C19

Fig. 4. Selected sentence-final F0 contours.

3.3. Results and discussion

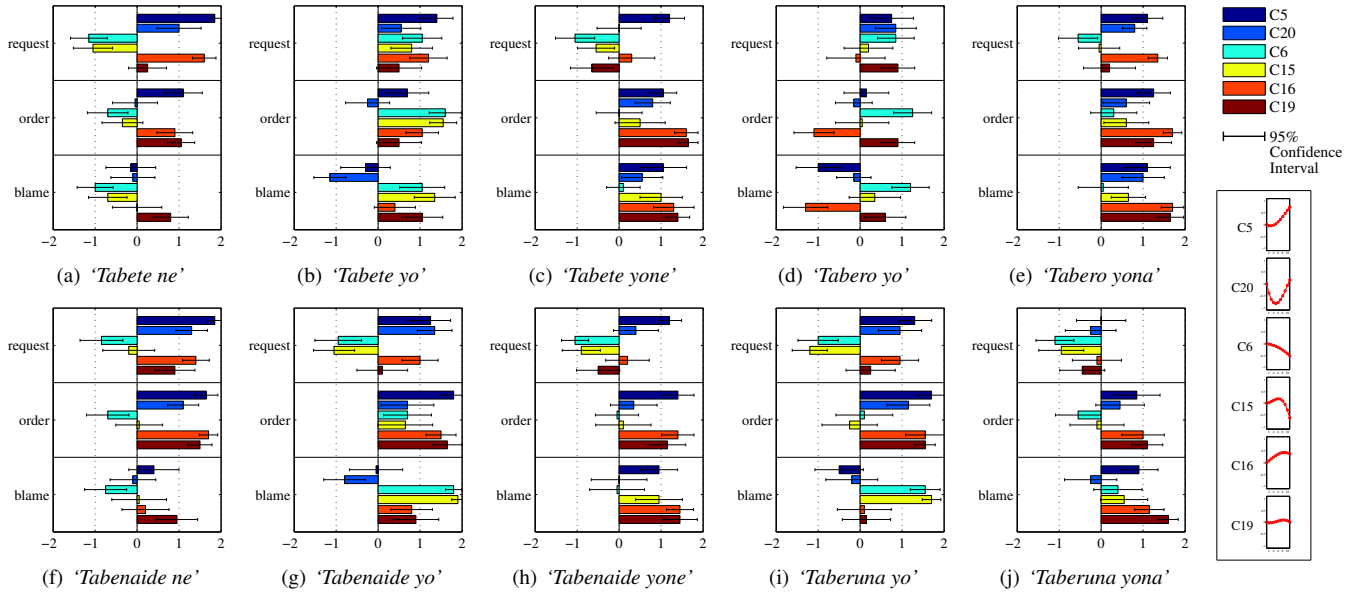
Fig. 5 shows the key results of the subjective evaluation, with a particular focus on a "request", an "order", and "blame".

- *'Tabete ne'* (Fig. 5(a)), *'Tabenaide ne'* (Fig. 5(f))

  Generally, the use of *'ne'* in the sentences *'Tabete'* and *'Tabenaide'* signals a polite "request". This was endorsed with the rising intonations C5, C20, and C16 in Fig. 4. In contrast, the flat intonation C19 conveyed an "order". In addition, in the negative sentence *'Tabenaide ne'*, an "order" was conveyed more clearly than in the positive sentence *'Tabete ne'* with the same intonation. Whether the sentence was positive or negative seemed to have a significant effect on the listeners' perception.

- *'Tabete yo'* (Fig. 5(b)), *'Tabenaide yo'* (Fig. 5(g))

  The sentence *'Tabenaide yo'*, which meant prohibition, strongly conveyed "blame" (*"Why did you eat even though I told you not to?"*) with the falling intonations C6 and C15, whereas it conveyed a "request" (*"Please don't eat."*) with the rising intonations C5 and C20. These results support previous findings about the function of *'yo'* [11]. In contrast, the falling intonations caused an "order"



(a) *'Tabete ne'* (b) *'Tabete yo'* (c) *'Tabete yone'* (d) *'Tabero yo'* (e) *'Tabero yona'*

(f) *'Tabenaide ne'* (g) *'Tabenaide yo'* (h) *'Tabenaide yone'* (i) *'Taberuna yo'* (j) *'Taberuna yona'*

Fig. 5. Subjective evaluation results of speaker's intentions depending on sentence-final particles and their intonations. The sentences in (a), (b), (c), (d), and (e) are positive (roughly, *"Please eat."*), and the others are negative (*"Please don't eat."*).

impression rather than "blame" in the positive sentence *'Tabete yo'*. Again, the listeners' perception seemed to be affected by whether the sentence was positive or negative.

- *'Tabete yone'* (Fig. 5(c)), *'Tabenaide yone'* (Fig. 5(h))

  'Yone' is known to have different lexical functions compared to *'ne'* and *'yo'*. "Blame", which was not much perceived in the sentences with *'ne'*, was conveyed with the rising C16 and the flat C19 intonations. This tendency differs from the case with *'yo'*, where "blame" was conveyed with the falling intonations C6 and C15.

- *'Tabero yo'* (Fig. 5(d)), *'Taberuna yo'* (Fig. 5(i))

  'Tabero' and *'Taberuna'* were imperative sentences. However, *'Tabero yo'* was perceived as an "order" only when with the falling intonation C6. On the other hand, for *'Taberuna yo'*, which meant prohibition, an "order" was perceived with the rising and flat intonations C5, C16, and C19, whereas "blame" was dominantly perceived with the falling intonations C6 and C15.

- *'Tabero yona'* (Fig. 5(e)), *'Taberuna yona'* (Fig. 5(j))

  The tendency in which the rising C16 and flat C19 intonations conveyed an "order" and "blame" was the same as that in *'yone'*. The negative sentence *'Taberuna yona'* was not perceived as a "request" with any intonations. This seems to be caused by the lexical functions that *'yona'* has in a negative sentence.

In addition to the above results, we want to point out here that *'Taberun datte'*, which lexically represented "hearsay" (*"I've heard that someone eats."*), was perceived as an "echo question" (*"Is it true that someone eats?"*) with the rising intonation C5. In contrast, *'Taberun datteyo'*, that is, *'Taberun datte'* followed by *'yo'*, was perceived as "hearsay" with C5 (in addition to C16 and C19). As for *'Taberu daroo'*, the falling intonation C6 gave an impression of a "guess" (*"I think that someone may eat."*), whereas the rising ones C5, C20, and C16 indicated a "question" (*"someone will eat, won't someone?"*). In *'Taberu daroone'* (*'Taberu daroo'* with *'ne'*), C16 (and C19) indicated a "guess".

4. CONCLUSION

We investigated listeners' perception of speaker's intention depending on sentence-final particles and their intonations in Japanese conversational speech. Results showed that the sentence-final intonations varied a great deal, and suitable combinations of a sentence-final particle and its intonation should be used to precisely convey the speaker's intention to the listeners

In this study, we used six representative centroids for analysis, even though we found various sentence-final F0 contours in the speech data. Furthermore, we found subtle differences in nuance, for example, between the rising intonations C5 and C16, which differed from each other in shape. Finding as many useful expressions for conveying subtle nuances and connotations as possible is a part of our future work. Another important issue is prosodic features such as the duration and power of the sentence-final syllable, which also contribute to expressing the speaker's intention. We intend to elucidate the relationship between these features and the intentions in our future work.

5. REFERENCES

[1] Y. Sagisaka, T. Yamashita, and Y. Kokenawa, "Generation and Perception of F0 Markedness for Communicative Speech Synthesis," *Speech Communication*, vol. 46, no. 3–4, pp. 376–384, July 2005.

[2] M. Schröder, "Expressive Speech Synthesis: Past, Present, and Possible Futures," in *Affective Information Processing*, J.H. Tao and T.N. Tan, Eds., pp. 111–126, Springer-Verlag, London, 2009.

[3] K. Iwata and T. Kobayashi, "Conversational Speech Synthesis System with Communication Situation Dependent HMMs," in *Proceedings of the Paralinguistic Information and its Integration in Spoken Dialogue Systems Workshop*, R.L.-C. Delgado and T. Kobayashi, Eds., pp. 113–123, Springer, New York, 2011.

[4] S. Makino and M. Tsutsui, *A Dictionary of Basic Japanese Grammar*, The Japan Times, Ltd., Tokyo, 1986.

[5] The National Language Research Institute, *Bound Forms ('Zyosi' and 'Zyodôsi') in Modern Japanese: Uses and Examples*, Shuei Shuppan, Tokyo, 1951 [in Japanese].

[6] A. Kamio, "The Theory of Territory of Information: The Case of Japanese," *J. Pragmatics*, vol. 21, no. 1, pp. 67–100, January 1994.

[7] S.K. Maynard, *Japanese Communication: Language and Thought in Context*, University of Hawai'i Press, Honolulu, 1997.

[8] H. Saigo, *The Japanese Sentence-Final Particles in Talk-in-Interaction*, John Benjamins Publishing Co., Amsterdam, 2011.

[9] N. Yoshizawa, "Intoneeshon (Intonation)," in *A Research for Making Sentence Patterns in Colloquial Japanese 1: On Materials in Conversation*, pp. 249–288, Shuei Shuppan, Tokyo, 1960 [in Japanese].

[10] S. Kori, "Intoneeshon (Intonation)," in *Asakura Nihongo Kooza 3: Onsei On'in (Asakura Japanese Series 3: Phonetics, Phonology)*, Z. Uwano, Ed., pp. 109–131, Asakura Publishing Co., Ltd., Tokyo, 2003 [in Japanese].

[11] Y. Katagiri, "Dialogue Functions of Japanese Sentence-Final Particles *'Yo'* and *'Ne'*," *J. Pragmatics*, vol. 39, no. 7, pp. 1313–1323, July 2007.

[12] E. Ofuka, J.D. McKeown, M.G. Waterman, and P.J. Roach, "Prosodic Cues for Rated Politeness in Japanese Speech," *Speech Communication*, vol. 32, no. 3, pp. 199–217, October 2000.

[13] T. Koriyama, T. Nose, and T. Kobayashi, "An F0 Modeling Technique Based on Prosodic Events for Spontaneous Speech Synthesis," in *Proc. ICASSP*, 2012, pp. 4589–4592.

[14] J.J. Venditti, K. Maeda, and J.P.H. van Santen, "Modeling Japanese Boundary Pitch Movements for Speech Synthesis," in *Proc. 3rd ESCA/COCOSDA Workshop on Speech Synthesis*, 1998, pp. 317–322.

[15] Y. Greenberg, N. Shibuya, M. Tsuzaki, H. Kato, and Y. Sagisaka, "A Trial of Communicative Prosody Generation Based on Control Characteristic of One Word Utterance Observed in Real Conversational Speech," in *Proc. Speech Prosody*. 2006, PS8–8–37.

[16] K. Iwata and T. Kobayashi, "Expressing Speaker's Intentions through Sentence-Final Intonations for Japanese Conversational Speech Synthesis," in *Proc. Interspeech*. 2012, Mon.P2b.03.

[17] S. Fujie, Y. Matsuyama, H. Taniyama, and T. Kobayashi, "Conversation Robot Participating in and Activating a Group Communication," in *Proc. Interspeech*, 2009, pp. 264–267.

[18] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné, "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*, vol. 27, no. 3–4, pp. 187–207, April 1999.

[19] J.H. Ward, Jr., "Hierarchical Grouping to Optimize an Objective Function," *J. Am. Statistical Assoc.*, vol. 58, no. 301, pp. 236–244, March 1963.

[20] T. Moriyama, "Bun-no Imi-to Intoneeshon (Sentence Meaning and Intonation)," in *Kooza Nihongo To Nihongo Kyooiku 1: Nihongogaku Yoosetsu*, Y. Miyaji, Ed., pp. 172–196, Meiji Shoin, Tokyo, 1989 [in Japanese].