Expression of Speaker’s Intentions through Sentence-Final Particle/Intonation Combinations in Japanese Conversational Speech Synthesis

Kazuhiro Iwata, Tetsunori Kobayashi

Perceptual Computing Laboratory, Waseda University, Japan

Abstract

Aiming to provide the synthetic speech with the ability to express speaker’s intentions and subtle nuances, we investigated the relationship between the speaker’s intentions that the listener perceived and sentence-final particle/intonation combinations in Japanese conversational speech. First, we classified F0 contours of sentence-final syllables in actual speech and found various distinctive contours, namely, not only simple rising and falling ones but also rise-and-fall and fall-and-rise ones. Next, we conducted subjective evaluations to clarify what kind of intentions the listeners perceived depending on the sentence-final particle/intonation combinations. Results showed that adequate sentence-final particle/intonation combinations should be used to convey the intention to the listeners precisely. Whether the sentence was positive or negative also affected the listeners’ perception. For example, a sentence-final particle ‘yo’ with a falling intonation conveyed the intention of an “order” in a positive sentence but “blame” in a negative sentence. Furthermore, it was found that some specific nuances could be added to some major intentions by subtle differences in intonation. The different intentions and nuances could be conveyed just by controlling the sentence-final intonation in synthetic speech.

Index Terms: speech synthesis, speaker’s intention, sentence-final particle, sentence-final intonation, conversational speech

1. Introduction

Speech synthesis technology has made remarkable progress in its synthetic voice quality, and the natural-sounding synthetic speech has recently become available. Moreover, various approaches to synthesizing expressive and conversational speech have been also reported in the past decade [1, 2, 3, 4]. However, due to the diversity of conversational speech, numerous problems still need to be solved to build a useful speech synthesis system for robots and speech-enabled agents that communicate with humans through synthetic speech. We communicate with each other through linguistic and paralinguistic information [5]. An utterance is affected by the speaker’s intentions, attitudes, feelings, personal relationship with the listeners, and so forth. The paralinguistic features of the utterance vary a great deal.

In spoken Japanese, a speaker’s intention is usually conveyed at the end of a sentence by sentence-final particles or auxiliary verbs [6]. The functions of sentence-final particles have been extensively studied in the field of linguistics [7, 8, 9, 10, 11, 12, 13]. For example, a sentence-final particle ‘yo’ indicates a strong assertion, and ‘ne’ indicates a request for listeners’ agreement. In addition, the intonation of the sentence-final particle, namely, the sentence-final intonation plays a significant role in expressing the intention and has also been studied over the years [14, 15, 16, 17, 18, 19]. For instance, a sentence with a sentence-final particle ‘ku’ becomes a declarative sentence with a falling intonation, whereas it becomes an interrogative sentence with a rising intonation. Moreover, it can express various additional nuances such as surprise, admiration, concern, doubt, or anger through different intonations. The functions of the sentence-final and phrase-final intonation have been discussed also in terms of turn-taking [20, 21]. However, not only the sentence-final intonation but also the intonation of the whole sentence varies depending on the speaker’s intention or attitude. In some languages other than Japanese, it has been reported that the listeners could identify the speaker’s attitudes before the end of the sentences [22, 23]. In contrast, the experiment that used two utterances of the same sentence uttered with different intentions demonstrated the importance of the sentence-final intonation in Japanese. When the speech segments of the sentence-final particle were cut off and swapped, the listeners perceived the intention that was expressed by the segment of sentence-final particle though the overall F0 contours of the utterances differed from each other [24].

On the other hand, there are not as yet many approaches to expressing the speaker’s intention by controlling the sentence-final intonation in the field of speech synthesis technology. Boundary pitch movements in Tokyo Japanese have been analyzed and modeled [25]. Five boundary pitch movements were chosen and their meanings were examined through association with eight semantic scales. A prosody control method has been proposed [26], which was based on the analysis of intonations of the one-word utterance ‘n’. This study revealed that speaking attitudes were expressed by the F0 average height and dynamic pattern of the word ‘n’, since it has no particular lexical meaning. 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. Although we investigated the relationship between the speaker’s intentions and sentence-final intonation contours in a previous study [27], we did not consider the relevance of the sentence-final particles.

In this study, we focus on the listeners’ perception of speaker’s intention, associating with sentence-final particles and their intonation contours in order to enable synthetic speech to express various intentions. In Section 2, we classify sentence-final intonation contours to find what kinds of sentence-final intonation were used in actual speech. In Section 3, we select several distinctive intonation contours from among the classification results and conduct a subjective evaluation to find suitable sentence-final particle/intonation combinations for conveying some specific intentions (hereafter “major intentions”).
Furthermore, in Section 4, we investigate sentence-final particle/intonation combinations that can add subtle nuances to the major intentions aiming to provide the synthetic speech with wide expressiveness. In Section 5, we conclude the paper with some remarks.

2. Classification of sentence-final intonation

2.1. Previous research in linguistics

As mentioned above, the functions of the sentence-final intonation have been discussed and various views have been proposed by the linguists. Table 1 shows some examples for the categorization of the sentence-final intonation in terms of the contours. In the majority of them, the sentence-final intonation contours were classified into two categories or up to five categories at most. However, there seems to be no accepted notion.

Table 1: Examples for categorization of sentence-final intonation in linguistics.

<table>
<thead>
<tr>
<th>Number of categories</th>
<th>Categories</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>Rise, Fall [15]</td>
</tr>
<tr>
<td>4</td>
<td>Rise, Fall-and-rise, Fall, Rise-and-fall [16]</td>
</tr>
<tr>
<td>5</td>
<td>Interrogative rise, Prominent rise, Fall, Rise-and-fall, Flat [17]</td>
</tr>
</tbody>
</table>

2.2. Speech data

We used speech data that were created with the aim of developing an HMM-based speech synthesis system that had multiple HMMs depending on situations of conversation [4]. To build the HMMs, we designed several situations and more than 2000 sentences derived from dialogues that our communication robot [28] performed. These sentences were uttered by a voice actress, to whom we did not indicated any specific intentions for each sentence but explained the situations in which the robot was supposed to utter each sentence. The F0 contours were extracted using STRAIGHT [29], and the phonemes were manually segmented. The intonation contours at the end of these utterances varied a great deal and expressed subtle nuances and connotations. Of these data, 2092 utterances whose sentence-final vowel was not devoiced were used for the analysis.

The sentence-final intonation contours, that is, the F0 contour in the sentence-final syllable was extracted by referring to the phoneme boundaries. Because the actual F0 values of the utterances differed from each other and were difficult to classify, the time and frequency axes were normalized. To remove F0 perturbation caused by jitter and microprosody, the logarithmic F0 contour was approximated by a third-order least squares curve. The approximated curve was sampled at 11 points that equally divided the duration into 10. Finally, the starting point of the sampled curve was parallel-translated to the origin [27].

2.3. Classification of sentence-final intonation contours

The normalized F0 contours obtained by the above process were classified by Ward’s clustering algorithm [30], which is one of the hierarchical clustering algorithms and merges clusters so as to minimize the within-cluster variance.

Figure 1 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 intonation contours were found, including not only simple rising and falling intonations but also rise-and-fall and fall-and-rise intonations.

2.4. Perceptual discrimination of intonation contours by centroids

We found the sentence-final intonation 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 ‘haua’, which had no specific linguistic meaning, was resynthesized by the STRAIGHT vocoder [29], and its F0 contour was replaced with 127 centroid pairs including a reverse order for each pair of centroids and then asked whether they perceived the two intonations to be the same or different. The results of the evaluation are shown in Figure 2. 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 experiments.

3. Major intentions conveyed by sentence-final particle/intonation combinations

3.1. Selection of representative intonation 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 intonation contours. Referring to the results of the preliminary evaluation (Figure 2), we stopped dividing clusters whose child clusters received 28 or fewer perceptions that their intonations were different. The selected clusters were C5, C29, C39, C5, C16, and C19.

Compared with the categorization in the previous research shown in Table 1, the centroid of the cluster C5 seems to correspond to the interrogative rise intonation, C20 to the fall-and-rise, C6 to the fall, C15 to the rise-and-fall, C16 to the prominent rise, and C19 to the flat. These results indicate that specific intonation contours corresponding to the categories in the previous research could be obtained from the speech database.

3.2. Experimental setup

A subjective evaluation was conducted to clarify what kind of intentions the listener could perceive through the sentence-final intonations produced by the selected six 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 (‘yon’, ‘yon’, ‘datteyo’, ‘daroone’, etc.). Synthetic voices of these sentences were generated by our HMM-based speech synthesis system [4]. 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
when we need to express "blame" in a negative sentence with a sentence-final particle 'yo', we should use C6 or C15.

When we need to express "blame" in a negative sentence with a sentence-final particle 'yo', we should use C6 or C15.

Sentence-final particle 'ne'
Generally, the use of 'ne' signals a polite "request". This was endorsed with the rising intonations C5 and C20 (Figures 3(a) and 3(f)). In the positive sentence 'Tabete ne', a "request" was also conveyed with the rising C16 intonation. On the other hand, in the negative sentence 'Tabenaide ne' with the rising C16 and flat C19 intonations, an "order" was perceived more clearly than in the positive sentence.

Sentence-final particle 'yo'
In the positive sentences 'Tabete yo' (Figure 3(b)) with the falling intonations C6 and C15 and 'Taberuna yo' (Figure 3(d)) with C6, an "order" ("Don't eat.") was conveyed. In contrast, in the negative sentences 'Tabenaide yo' (Figure 3(g)) and 'Taberuna yo' (Figure 3(i)), which meant prohibition, "blame" ("Why did you eat even though I told you not to?") was strongly conveyed with the falling intonations C6 and C15. When with the rising and flat intonations C5, C16, and C19, they caused an "order" impression.

Sentence-final particles 'yone' and 'yona'
'Yone' is known to have different lexical functions from 'ne' and 'yo'. "Blame", which was not much perceived in the sentences with 'ne', was conveyed with the rising C16 and flat C19 intonations (Figures 3(c) and 3(h)). This tendency differs from the case with 'yo', where "blame" was conveyed with the falling intonations C6 and C15. "Blame" could be conveyed also in 'yona' with C16 and C19 (Figures 3(e) and 3(j)).

To summarize the results, we can obtain Table 2, which shows the highest scored intentions for each combination. Then we can consult this table when we generate synthetic speech. For example, we can express "request" in a positive sentence with a sentence-final particle 'ne' by using C5 or C20 as the sentence-final intonation. When we need to express "blame" in a negative sentence with a sentence-final particle 'yo', we should use C6 or C15.

3.3. Results and discussion
Figure 3 shows the key results of the subjective evaluation, with a particular focus on a "request", an "order", and "blame".
Figure 3: Subjective evaluation results of major intentions depending on sentence-final particle/intonation combinations. The sentences in (a), (b), (c), (d), and (e) are positive (roughly, “Please eat.”), and the others are negative (“Please don’t eat.”).

Table 2: Speaker’s intention conveyed by sentence-final particle/intonation combination. The intentions (“request”, “order”, and “blame”) that received the highest positive score are shown by the initial letters and underlined when their scores are higher than or equal to 1.0 (rather suitable). *** p<0.001, ** p<0.01, * p<0.05, + p<0.1, where p is the maximum p-value among two comparisons.

### Additional nuances conveyed by sentence-final particle/intonation combinations

As the next step of this study, we investigated whether the listeners could perceive additional intentions, attitudes, or feelings (hereinafter collectively called “additional nuances”) through the sentence-final intonations.

#### Selection of representative intonation contours

We increased the types of representative intonation contours to be used for the subjective evaluation of additional nuances. The selected six clusters in 3.1 were divided into several subclusters. This time, we merged two clusters from the 128 leaf nodes to the six clusters. When two clusters received 27 or more out of 30 (more than or equal to 90%) perceptions that their intonations were different, they were not merged. Additionally, their parent cluster was not merged with its paired cluster either. Thus, the cluster C5 was ultimately divided into 8 subclusters, C20 into 3, C6 into 11, C15 into 5, C16 into 7, and C19 into 7, as listed in Table 3.

#### Experimental setup

We defined the intentions of a “request”, an “order”, and “blame” as the major intentions for this evaluation. Considering the results of the previous section (Table 2), we chose sentence-final particle/intonation combinations from among those which received the significantly different (p<0.05) score: ‘Tabete ne’ with the sentence-final intonations by the subclusters of C5, C20, and C16 for a “request”; ‘Tabenai de yo’ with the subclusters of C16 and C19 for an “order”; and ‘Tabenaide yo’ with the subclusters of C6 and C15 for “blame”. We generated 18 synthetic utterances with different intonations for the major intentions “request”, “order”, and “blame” as the stimuli.

Three additional nuances for each major intention (Table 4) were designed, which one of the authors perceived from these stimuli in advance. The stimuli were presented to 22 listeners,
along with the situation of the dialogue and the additional nuances for each of the three major intentions. The listeners were asked to evaluate whether they felt the additional nuances on a four-level scale: 3 (strongly felt), 2 (felt), 1 (slightly felt), and 0 (not felt). In addition, they were asked to freely describe any other nuances they felt and evaluate the stimuli in the same way.

4.3. Results and discussion

The results are shown in Figure 4. Several intonation contours were found to be able to convey some additional nuances.

- “Request” (Figure 4(a))

The additional nuances (r1) and (r2) have mutually exclusive meanings. C61, C23, C43, and C131 conveyed the nuance (r1) well. They all rise toward a considerably high frequency at the end of a sentence, which is a characteristic that can be considered to convey “sincerity” or “cordiality”. In contrast, C248 and C123, which were slightly rising and rather flat overall contours, implied (r2). C135 and C41, which were fall-and-rise contours with the extreme lowering, insinuated (r3). As for the free description, some listeners noted that C248 and C346 created a sense of familiarity and caused an impression that the speaker was just like a senior (such as a parent of a friend). However, whether others similarly perceive it or not needs to be investigated further.

- “Order” (Figure 4(b))

The additional nuances (o1) and (o2) are mutually exclusive. C269, C101, and C229, which were undulated, conveyed (o1). C71, which were slightly rising and undulated, also conveyed (o1). On the other hand, (o2) was not very clearly conveyed, but by C248. C229, C214, and C101 conveyed (o3) in addition to (o1).

- “Blame” (Figure 4(c))

The additional nuance (b1) seemed to be expressed by a large rise-and-fall movement such as C117 and C155. C24 and

Table 3: Selected subclusters as representative intonation contours to be used for the evaluation of additional nuances. Their centroids are shown in Figure 4.

Table 4: Evaluated additional nuances for each major intention.
C329, which had a steep fall after slight rise, expressed (b2) clearly. C151 and C233, which had a slight fall, expressed (b3).

5. Conclusions

We investigated the expression of speaker’s intentions through sentence-final particle/intonation combinations in Japanese conversational speech. Results showed that the sentence-final intonation contours varied a great deal and that adequate sentence-final particle/intonation combinations should be used to convey the intention to the listeners precisely. Furthermore, it was found that some specific nuances could be added to some major intentions by subtle differences in intonation.

We indicated that not only major intentions but also subtle nuances could be expressed by sentence-final particle/intonation combinations. However, other prosodic features such as the duration, power, and voice quality of the sentence-final syllable must be changed in actual speech to contribute to conveying the intentions. They need to be adequately controlled in order to express the intentions more precisely and more expressively. We intend to elucidate the relationship between these features and the intentions in our future work.

6. Acknowledgements

This work was supported in part by Global COE Program “Global Robot Academia”.

7. References


