Expressing Speaker’s Intentions through Sentence-Final Intonations for Japanese Conversational Speech Synthesis

Kazuhiko Iwata, Tetsunori Kobayashi

Perceptual Computing Laboratory, Waseda University, Tokyo, Japan

Abstract

In this study, we investigated speaker’s intentions that the listeners perceive through subtly different sentence-final intonations. Approximately 2,000 sentence utterances were recorded and the fundamental frequency (F0) contours at the last vowel of those sentences were classified through one of the standard clustering algorithms. There found various F0 contours, namely, not only simple rising and falling intonations but also rise-fall and fall-rise intonations. In order to reveal the relationship between the intonation and the intentions, 10 representative contours were selected on the basis of the results of the clustering. Using the selected contours, a subjective evaluation was conducted. Six Japanese sentences that could have different meanings according to the sentence-final intonations were synthesized and the F0 contour at the last vowel of each sentence was replaced with the contours. The results of the evaluation by nine listeners showed that, for example, a certain falling intonation could express the intention of the “conviction” and another one that slightly differ in the shape could convey “doubt.” It was found that the subtle difference in the sentence-final F0 shape conveyed various nuances and connotations.

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

1. Introduction

Text-to-speech systems have been put to practical use in reading out texts such as emails, web pages and weather forecasts, and have achieved steady success. There are, on the other hand, high expectations for conversational speech synthesis technology from spoken dialogue systems (e.g., humanoid robot systems, speech-oriented guidance systems, etc.), and several approaches to synthesizing expressive and emotional speech have been reported [1]. However, numerous problems still need to be solved to establish the conversational speech synthesis technology. Various expressions in speech are used in human communications depending on not only emotions but also speaker’s intentions, attitudes, and so on.

In the Japanese language, a speaker’s intention is generally represented at the end of the sentence by a specific linguistic expression such as sentence-final particles or auxiliary verbs. In consequence, the intonation shape at the end of the sentence also varies widely depending on the speaker’s intentions and attitudes. These sentence-final intonations have been extensively researched in the field of Japanese linguistics [2], since they play an important role in conveying the speaker’s intentions. There are, however, as yet few approaches to express the speaker’s intentions with subtle connotations by controlling intonations [3].

In this study, we focus on the relationship between the subtle expression in the sentence-final fundamental frequency (F0) contours and the speaker’s intentions and attitudes that the listeners perceive, in order to introduce the ability of expressing speaker’s intentions to the conversational speech synthesis. This paper is organized as follows. In Section 2, the classification of sentence-final intonations is described. The method for normalizing sentence-final intonations and the results of the classification are discussed. In Section 3, the subjective evaluation to investigate the relationship between the sentence-final intonations and the speaker’s intentions that the listeners perceive through those intonations is explained. Furthermore, the results of the evaluation are discussed in detail. Finally, our conclusions of this paper are given in Section 4.

2. Classification of sentence-final intonations

2.1. Speech Data

We have developed a hidden Markov model (HMM) based speech synthesis system, which has multiple HMMs depending on communication situations [4]. To build the HMMs, we designed several communication situations and 2,111 sentences derived from the dialogues that our communication robot [5] was supposed to perform. These sentences were uttered by a voice actress assuming the situations. The F0 contours were extracted by using STRAIGHT [6] and the phoneme segmentation was done manually.

It was found that the intonations of these utterances were varied and played the role to express subtle nuances
and connotations. Hence, these speech data were applied for classifying the sentence-final intonations. Of these data, 2,092 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 vowel in the sentence-final syllable were classified by a standard clustering algorithm (Ward’s algorithm [7]). Voiced consonants /l/, /m/, /y/, /r/ and /w/ were included in the section where the F0 contour was extracted for classification, since the powers of those consonants were relatively strong and their F0 contours had the continuous movement to that of the following vowel. Because the actual F0 values of the utterances differed from each other and were difficult to be classified, the time and frequency axes were normalized.

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

2.3. Classification by clustering

The normalized F0 contours thus obtained by the above-mentioned process were classified by Ward’s clustering using the normalized F0 values and their differences as feature vectors. Fig. 2 shows an example of the results of the classification where the number of clusters was set to 16. The F0 contours denoted by the red circles are the centroids of each cluster.

There were found various sentence-final F0 contours, namely intonations: not only simple rising and falling intonations but also rise-fall and fall-rise intonations.

3. Analysis of relationship between intonations and intentions

3.1. Selection of representative F0 contours

It was found that the sentence-final F0 contours could be classified into several distinctive clusters. However, this clustering was based on the shapes of the F0 contours. Thus, it was prefigured that some clusters could have no notable perceptual difference from the others. Within a cluster, on the other hand, some contours could potentially be perceived differently and could express subtle nuances.

Then, a preliminary evaluation was conducted and 10 discriminative F0 contours were singled out in order to investigate the relationship between the intonation and the intentions through the subjective evaluation. In this
Table 1: Sentences for subjective evaluation. Each sentence has multiple meanings depending on different sentence-final intonations.

<table>
<thead>
<tr>
<th>Sentence (in Japanese)</th>
<th>Example of Meaning (in English)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(S1) Hontodesuka</td>
<td>Is it true? I see.</td>
</tr>
<tr>
<td>(S2) Kotae wakattandesuka</td>
<td>Did you get the answer? I've got the answer.</td>
</tr>
<tr>
<td>(S3) Sounandesuka</td>
<td>Really? I see.</td>
</tr>
<tr>
<td>(S4) Ashitamo kitekudasaiyo</td>
<td>I do hope you can come again tomorrow. You have to come again tomorrow.</td>
</tr>
<tr>
<td>(S5) Tanoshikattandatte</td>
<td>You seemed to have had fun, didn’t you? I heard that it was fun.</td>
</tr>
<tr>
<td>(S6) Korede tideshou</td>
<td>You accept it, don’t you? OK, I accept it.</td>
</tr>
</tbody>
</table>

study, the number of the selected contours was limited to 10 in consideration of the load for the participants in the evaluation. The selected contours are shown in Fig. 3.

3.2. Subjective evaluation

To clarify what kind of speaker’s intentions would be perceived by the listener depending on the sentence-final intonations, six Japanese sentences that could convey various connotations with different intonations were prepared as shown in Table 1. Synthetic speech for each sentence was generated by the speech synthesis system we have developed [4] and the F0 contour at the last vowel of each sentence was replaced with the selected contours shown in Fig. 3. For each sentence, we designed situations in which the sentence could be uttered and several intentions and attitudes that the sentence could express. Nine listeners were presented with the synthetic voices and were asked to select one or more speaker’s intentions listed in the answer sheet or to describe the intentions freely.

3.3. Results and discussions

The relationship between the F0 contours and the perceived intentions is shown in Table. 2. The figures denote the number of responses for each intonation and each intention. For each intonation, the intention that had the most responses is underlined. The intonations shown by “–” were not evaluated, because the preliminary experimental results indicated that there was no notable perceptual difference from the other contours.

3.3.1. Sentences (S1), (S2) and (S3)

In general, when sentences such as (S1), (S2) and (S3) are uttered with a rising intonation, they are perceived as interrogative sentences. In contrast, those with a falling intonation represent that the speaker is convinced.

Although the F0 contours C1 and C2 were both falling intonations, their perceived intentions prominently differed: C1 mainly conveyed the intention of “conviction” while C2 conveyed “doubt.” It could be presumed that, in C2, the degree of descent in F0 became gradual from the middle of the duration, which caused the impression of intonational rising.

As for “doubt” and “surprise,” although the listeners were asked their degrees, namely, “strong” or “slight,” no clear distinction between them was found. However, for C9, 11 responses out of 13 for “doubt” were “slight doubt.” This implied that C9, which was rising intonation, seemed to be a general intonation for “inquiry.” On the other hand, C8 and C10, which were also rising intonations, gave the impression of “doubt” rather than “inquiry.” They conveyed a nuance different from mere inquiry. Furthermore, due to the steep increase in F0, C7 conveyed “surprise” rather than “doubt.”

3.3.2. Sentence (S4)

In this sentence, the degree of the speaker’s requirement could be expected to be expressed through different intonations. The flat and rather flat intonations, C6 and C2, expressed “disinterest.” In contrast, C10, which showed the deep fall-rise, conveyed “imperative.” “Desire” was conveyed with not only C7 and C9, which were rising intonations, but also C5, which was falling intonation.

3.3.3. Sentence (S5)

Sentence (S5) had two major meanings depending on if it was uttered with a rising or falling intonation. As expected, C7, C8, C9 and C10, which were rising intonations, represented “inquiry” and from C1 to C6, which were falling intonations, “hearsay” was the majority.

3.3.4. Sentence (S6)

Sentence (S6) had two major meanings depending on the intonation, the same as Sentence (S5). The results were
almost the same as those of (S5), except C2, in which the anterior half was falling intonation and the posterior half was rather flat, which indicated a “request for acceptance.” This could be caused by the same reason as the case in which C2 conveyed “doubt” rather than “conviction” in sentences (S1), (S2) and (S3). However, the reason for the difference in the results between (S5) and (S6) should be discussed from the point of view of the relation with Japanese sentence-final particles and auxiliary verbs.

4. Conclusions

The relationship between sentence-final intonations and speaker’s intentions that the listeners perceived through the intonations was investigated. The results of the classification of sentence-final F0 contours indicated that there were various F0 contours, namely, not only simple rising and falling but also rise-fall and fall-rise ones. The subjective evaluation revealed that the sentence-final intonations played an important role in conveying different intentions and connotations. In this paper, we focused on the F0 contours. Other prosodic features such as duration and power, however, must contribute to expressing the speaker’s intentions. F0 average heights in one-word and short utterances were reported to express speaker’s attitudes [3]. Revealing the relationship between those features and the intentions is one of our future works.

In this study, some sentence-final F0 contours were found that expressed almost the same intentions as the others but obviously gave different impressions to the listeners. The F0 contours that could give the different impressions to the listeners could be expected to enhance the expressiveness in speech synthesis, even though the intentions that were conveyed through those contours would be identical. To make the conversation between humans and robots more natural, spoken dialog systems require speech synthesis systems to be equipped with the ability to express different intonations and subtle nuances. We also intend to find useful F0 contours in our future works.

5. References


Table 2: Relationship between representative F0 contours and perceived speaker’s intentions. The figures denote the number of the responses for each intonation and each intention. For each intonation, the intention that had the most responses is underlined.

<table>
<thead>
<tr>
<th>Sentence</th>
<th>Intention</th>
<th>C1</th>
<th>C2</th>
<th>C3</th>
<th>C4</th>
<th>C5</th>
<th>C6</th>
<th>C7</th>
<th>C8</th>
<th>C9</th>
<th>C10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total of</td>
<td>Conviction</td>
<td>18</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td></td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>(S1)</td>
<td>Inquiry</td>
<td>6</td>
<td>6</td>
<td>7</td>
<td>5</td>
<td></td>
<td>3</td>
<td>5</td>
<td>7</td>
<td>11</td>
<td>7</td>
</tr>
<tr>
<td>(S2)</td>
<td>Doubt</td>
<td>2</td>
<td>15</td>
<td>8</td>
<td>8</td>
<td></td>
<td>8</td>
<td>12</td>
<td>18</td>
<td>13</td>
<td>20</td>
</tr>
<tr>
<td>(S3)</td>
<td>Surprise</td>
<td>2</td>
<td>2</td>
<td>9</td>
<td>7</td>
<td></td>
<td>2</td>
<td>20</td>
<td>10</td>
<td>8</td>
<td>5</td>
</tr>
<tr>
<td>(S4)</td>
<td>Imperative</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td></td>
<td>0</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Desire</td>
<td>4</td>
<td>0</td>
<td>4</td>
<td>5</td>
<td></td>
<td>7</td>
<td>0</td>
<td>6</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>Request</td>
<td>5</td>
<td>1</td>
<td>3</td>
<td>4</td>
<td></td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Disinterest</td>
<td>2</td>
<td>7</td>
<td>2</td>
<td>0</td>
<td></td>
<td>1</td>
<td>8</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>(S5)</td>
<td>Hearsay</td>
<td>9</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td></td>
<td>7</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Inquiry</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td></td>
<td>2</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>7</td>
</tr>
<tr>
<td>(S6)</td>
<td>Acceptance</td>
<td>8</td>
<td>0</td>
<td>9</td>
<td>7</td>
<td></td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Request</td>
<td>for acceptance</td>
<td>1</td>
<td>9</td>
<td>0</td>
<td>2</td>
<td></td>
<td>8</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
</tr>
</tbody>
</table>

INTERSPEECH 2012