Analysis of factors involved in the choice of rising or non-rising intonation in question utterances appearing in conversational speech

Hiroaki Hatano 1, Miyako Kiso 1, Carlos T. Ishi 1

1 ATR Intelligent Robotics and Communication Labs.
hatano.hiroaki@atr.jp, miyakokiso@atr.jp, carlos@atr.jp

Abstract
In general, the end of question utterances is accompanied by a rising intonation. However, non-rising intonation is commonly observed in question utterances appearing in conversational speech. In order to clarify the factors involved in the choice of rising or non-rising intonation, in the present work, we analyzed question utterances extracted from Japanese conversational dialogue speech data of multiple speakers. Each utterance was categorized in terms of the question type, the phrase final intonation and the phrase final morpheme. Analysis results revealed that (1) about 20% of question utterances were neither accompanied by a rising intonation nor by a pitch reset; (2) among the question types, non-rising intonation appeared in more than 50% of “request for agreement”, “open-type questions”, “backchannel-type questions”, and “self-directed doubt-type questions”; (3) regarding morphemes, non-rising intonation appeared in more than 50% of utterances ending with question-related final particles; in contrast, more than 80% of the utterances ending with morphemes other than final particles were accompanied by a rising intonation or a pitch reset.

Index Terms: question, intonation, speech act, final particle, Japanese natural conversation

1. Introduction
In general, the intonation of question sentences differs from that of declarative sentences. Typically, the F0 (fundamental frequency) of an utterance falls in the end of declarative sentences, while it rises in the end of question sentences. The sentence final intonation is also focused in Japanese teaching, so that Japanese learners are instructed to raise the sentence final intonation in interrogative sentences and request for confirmation [1][2]. It is also reported that rising intonation appears in questions where the interlocutor’s reaction is required [3]. However, most of past works for Japanese are based on written language data or introspection, and not on spoken language (spontaneous speech) data.

In [4], natural conversations were analyzed in Swedish, showing that prosody differs for different question types. They classified questions in four types: yes-no question, wh question, alternative question, and multi-question. They found that yes-no questions generally have a falling intonation, whereas wh questions generally have a rising intonation. In [5], the terminal intonation of question utterances extracted from conversations between a doctor and a patient in Dutch was analyzed by classifying the utterances in yes/no-questions, wh-questions and declarative questions. It has been found that terminal F0 of final rises assume higher values in the predicted order wh questions < yes/no-questions < declarative questions, only in male speech.

In Japanese daily conversations, it is also common to observe question utterances ending with non-rising intonation. In [6], spontaneous speech of one Japanese speaker was analyzed. It was found that the distributions of phrase final tones differ depending on the dialogue acts (including questions) and the phrase final morphemes. However, the past works cited above do not clarify in which situations the question utterances are accompanied by rising or non-rising intonation.

Thus, in the present work, we investigated the factors involved in the choice of rising or non-rising intonation in question utterances extracted from Japanese conversational dialogue speech data of several speakers, by considering linguistic and non-linguistic cues. The findings of the present work can be applied to improve the quality/naturalness of speech synthesis, as well as to improve speech understanding in dialogue systems (e.g. for deciding if the system is really required to provide an answer).

2. Materials and methods
2.1. Speech data
For analysis data, we used the Japanese conversational database containing 58 dialogue sessions of 28 speakers (12 male and 16 female speakers). Each dialogue session has 10 to 15 minutes of free conversations. The total length of speech data is about 600 minutes. The dialogue partners include family members, friends and people meeting for the first time. Simultaneous recording of speech and EEG (electro-glottograph) signals are available. Sampling rates are 16 kHz/16 bits. Audio data is recorded using directional microphones (Sanken CS-1). In part of the sessions, headset microphone data is also available.

The speech utterances are segmented in phrase units based on pauses and clear pitch resets between phrases, including a total of 55,000 utterances. The database contains dialogue act labels (including turn-keeping categories, turn-yielding categories like statements and questions, and interjectional categories like backchannels and fillers) for each utterance, annotated by 2 ~ 4 annotators. In the present work, we selected the turn-yielding utterances where 2 or more annotators attributed dialogue act labels other than declarative sentences, corresponding to utterances where the speaker expects a response from the interlocutor. In the present work, we call these utterances as “question utterances”. A total of 3,661 question utterances were obtained.

2.2. Classification of question types
For each of the question utterances, the question types were labeled by 3 annotators, based on the categories shown in Table 1. This list was elaborated by the authors after extensive discussion about the possible types and ambiguities in definition. The agreement rates (in terms of kappa coefficients) among the question type labels of the 3 annotators were 0.63, 0.76 and 0.79. The 2,660 utterances
The phrase final is split in two segments of equal length, and representative F0 values are extracted for each segment. Several candidates for the representative F0 values have been tested in [7]. Here, we use the ones that best matched with perceptual scores of the F0 movements. For the first segment, an average value is estimated using F0 values within the segment (F0avg2a). And for the second segment, a target value is estimated as the F0 value at the end of the segment of a first order regression line of F0 values within the segment (F0tgt2b). A variable called F0move is defined as the difference between F0tgt2b and F0avg2a, quantifying the amount and direction of F0 movement within the syllable. F0move is positive for rising F0 movements, and negative for falling movements.

\[ F0move = F0tgt2b - F0avg2a \]  

Phrase finals are categorized as rise pitch movements (Rt) when F0move > 1 semitone, fall pitch movements (Fa) when F0move < -2 semitones, and flat pitch movements (Ft) otherwise. These thresholds are based on pitch movement perception experiments [7].

For phrase final duration, an automatic procedure was first realized, by using power and spectral change constraints [7]. And then the errors in the automatic segmentation were manually corrected. The newly segmented boundary intervals are used as segmental duration of the phrase finals. As for pitch movement categories, duration categories are also defined as short (S) when duration < 200 ms, and long (L) otherwise.

A parameter called F0reset is another important factor in categorizing the phrase final tones. This parameter indicates the presence or absence (or degree) of pitch reset between the phrase final and the syllable prior to the phrase final. The degree of pitch reset is defined as follows:

\[ F0reset = F0avg_p - F0avg_2a, \]  

where F0avg_2a is an average F0 value of the final portion of the syllable preceding the phrase final. F0avg_p is estimated from four reliable F0 values obtained by back-tracking and searching from the phrase final start point. Positive F0reset values indicate presence of F0 reset (Rt).

A tone is then described by the combination of the pitch movement, pitch reset, and duration categories, as shown in Table 2. From the utterances where agreement was obtained for the question type annotations, the ones where F0 could not be extracted were removed from the analysis, resulting in a total of 2,016 utterances. The automatically classified tones were checked and corrected by a subject with experience on tone classification. The agreement rate between the automatic description of tones and the subject perceptual decisions was 0.84 for F0move and 0.97 for F0reset.

<table>
<thead>
<tr>
<th>F0move &gt; 1 semitone: Rs</th>
<th>dur &lt; 200 ms S</th>
<th>dur &gt; 200 ms L</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 &lt; F0move &lt; 1 semitone: Ft</td>
<td>F0reset &gt; 0; Ft</td>
<td>Sft (423)</td>
</tr>
<tr>
<td>F0move &lt; -2 semitone: Fa</td>
<td>F0reset &lt; 0</td>
<td>Sfa (316)</td>
</tr>
</tbody>
</table>

Table 2. Phrase final tones. The numbers within parentheses are the number of occurrences.
It is worth to mention that the tone types $Rs$, $Rt$, and $RtFa$ are equivalent to J-ToBI boundary pitch movements $LH\%$, $H\%$, and $HL\%$ respectively [8].

### 2.4. Classification of phrase final morphemes

Linguistic information about the part of speech attributed to morphemes appearing at phrase finals is taken into account when verifying the influence of tones. For example, phrases ending with no final particles are expected to have rising tone.

All phrases were first arranged according to the morpheme appearing at their phrase finals. The identification of the morpheme was conducted by a Japanese linguist, based on the text transcriptions of the speech data. Table 3 shows the list of morphemes arranged by their corresponding parts of speech. The morphemes with number of occurrences lower that 10 were removed from the table.

<table>
<thead>
<tr>
<th>Part of speech</th>
<th>Morpheme</th>
</tr>
</thead>
<tbody>
<tr>
<td>final particles</td>
<td>/no, N/ (315), /ka/ (230), /na/ (162), /yone/ (93), /ne/ (90), /kana/ (65), /kane/ (22), /yonae/ (17), /Qke/ (13)</td>
</tr>
<tr>
<td>auxiliary verbs</td>
<td>/yaro, daro/ etc. (148), /nai, heN/ etc. (60), /yaN, jaN/ (45), /te/ (17)</td>
</tr>
<tr>
<td>postpositional particles</td>
<td>/wa/ (38), /de/ (21), /mi/ (19), /ga/ (17), /Qte/ (14)</td>
</tr>
<tr>
<td>(others)</td>
<td>(518)</td>
</tr>
</tbody>
</table>

### 3. Results and discussions

#### 3.1. Analysis of the question types

Fig. 1 shows the distributions of the phrase final tones (obtained from $F0move$ and $F0reset$ measures) for the utterances where inter-labeler agreement was obtained for the question types. The categories where the total number of utterances was lower than 30 (Quiz-type questions, Counter-questions, Reflexive repetition, Repetition request, Authorization request, Action request) are removed from Fig. 1.

![Figure 1: Distributions of phrase final tones for each question type. The total number of utterances is shown under each label.](image)

The question types whose rising tones are higher than 50% are ordered as follows:

- Ambiguous > Yes-no questions > Subjective feedback request > Information request

Regarding the non-rising ones, the categories are ordered as follows,

- Open-type questions > Agreement request > Self-questions

Open-type questions and Agreement request show high percentage of reset-fall plus reset-flat tones compared with other types. Self-questions show the highest percentage of falling tones (without reset), which is the farthest tone to the rising tones.

The tone distributions above can be interpreted as corresponding to “where the answer information is belonging to”. For example, in Yes-No question, Subjective feedback request, and Information request, the speaker assumes that the interlocutor has the answer information to that question. On the other hand, in Open-type question, Agreement request, and Self-questioning, the speaker assumes that the interlocutor has not necessarily knowledge about the information. Therefore, the speaker tends to raise the phrase final tone when he/she assumes the interlocutor has the answer information to the question.

One-way ANOVA revealed that $F0move$ is dependent on the question type ($F (8, 1963) = 21.72; p < .01$). (The categories whose total number of utterances was lower than 30 were removed.) The results of multiple comparisons of Tukey’s honestly significant difference test are shown below (MSe: 17.24; $p < .05$). Classifications according to the answer information affect not only the occurrence rates of rising or non-rising tones, but also the degree of rising.

- Ambiguous, Yes-no questions, Subjective feedback request, Information request > Open-type questions, Agreement request
- Yes-no questions, Information request > Backchannel-type questions

One-way ANOVA revealed that $duration$ is also dependent on the question type ($F (8, 1963) = 7.13; p < .01$). The results of multiple comparisons of Tukey’s HSD test are shown below (MSe: 0.0067; $p < .05$).

- Backchannel-type questions > Yes-no questions, Information request, Agreement request, Proposal
- Subjective feedback request > Yes-no questions

Backchannel-type questions can be interpreted as having lower motivation to expect an answer from the interlocutor. Thus, these significant differences on duration possibly indicate speaker attitudes of interest.

#### 3.2. Analysis of the phrase final morphemes

Fig. 2 shows the distributions of the phrase final tones according to the last morphemes of the utterances. The morphemes with total number of utterances lower than 30 are removed from the table.

The morphemes whose rising tones are higher than 50% are ordered below.

- /wa/ > (others) > /nai/ group > /yaro/ group

All other morpheme categories had predominance of non-rising tones.

- /yone/, /ne/, /kana/, /na/, /ka/

![Figure 2: Distributions of phrase final tones for each morpheme. The total number of utterances is shown under each label.](image)
The question utterances were classified according to three indices representing interpersonal relationship: familiarity (casual ~ formal), the relative social status (lower ~ higher) and a “concern” factor (concern, no concern) between the dialogue partners. The “concern” label is attributed when the speaker talks with concern (care) to the interlocutor, for example, in situations where the interlocutor is older, or has a higher status (such as boss), or is meeting for the first time.

The labels were annotated by two annotators for each pair of dialogue partners, resulting in agreement rate of 0.7.

From the three indices, differences in the distributions of the phrase final tones were found only in the “concern” factor, as shown in Fig. 3.

One-way ANOVA revealed that F0move is dependent on the “concern” factor, and F0move in “no concern” was significantly higher than in “concern” (F (1, 2014) = 11.25; p < .01). Statistical significance tests on duration indicated no significant differences (F (1, 2014) = 0; ns).

These results can be interpreted as the speaker being restraining the use of rising tones when concerning for the interlocutor. Thus, rising intonation also has the effects of getting further into the interlocutor’s information territory.

4. Conclusions

In order to clarify the factors involved in the choice of rising or non-rising intonation at the end of sentences, we analyzed question utterances extracted from Japanese conversational dialogue speech data of multiple speakers. For analysis of the question utterances, we focused on the question types, the phrase final morphemes, and the interpersonal relationship.

Analysis results indicated that the occurrence rates of non-rising tones increase in question types where the speaker assumes that the answer information does not necessarily belong to the interlocutor, such as Open-type questions, Agreement request and Self-directed questions. Regarding the morpheme types, it was found that the occurrence rates of non-rising tones increase in /yone/, /ne/, /na/ and /kana/, which are final particles expressing intrinsic modalities with respect to the interlocutor. Further, from the analysis of interpersonal relationship, it was found that the degree of F0 values were significantly lower when the speaker expresses concerns for the interlocutor, i.e., in situations where the interlocutor is older, or has a higher status (such as boss), or is meeting for the first time.

In conclusion, we found that the intonation of question utterances in Japanese is influenced by both linguistic factors (sentence final morphemes) and non-linguistic factors (question types and inter-personal relationship).

Finally, although the results in the present work are for Japanese conversational speech, the same strategies can be used to extend the analysis of other languages.

5. Acknowledgements

This study was supported by JSPS KAKENHI (No. 23680019). We also thank all research assistants of the ATR group who contributed with the annotations.
6. References


