Interpersonal factors affecting tones of question-type utterances in Japanese

Hiroaki Hatano [1], *Carlos T. Ishi* [1], *Miyako Kiso* [1]

[1] ATR Intelligent Robotics and Communication Labs.

{hatano.hiroaki, carlos, miyakokiso}@atr.jp

10.21437/SpeechProsody.2014-189

Abstract

The purpose of this paper is to clarify the interpersonal factors affecting phrase final tones of question-type utterances in daily conversations. We extracted question-type utterances ending with final particles from our Japanese dialogue speech database and classified them into two categories according to the degree of information request. Prosodic features were then analyzed by focusing on phrase final F0 movement and pitch reset. Analysis results indicated that F0 rising and falling degrees increase when the speaker expresses an attitude of intimacy to the dialogue partner, such as in conversations among family members and infant-directed speech. In addition, the presence of pitch reset in the phrase final was found to have functions of relieving the speaker's tension, when the dialogue partners have distant relationship.

Index Terms: question, intonation, interpersonal relationship

1. Introduction

In daily conversations, question is one of the most commonly-used speech acts. In general, a question is premised on the presence of an interlocutor. Therefore, it is a speech act peculiar of dialogues, being a key to understand the details in dialogue communication.

In general, it is important for raising the intonation of question-type utterances in Japanese. For example, it is stated that if the pitch of phrase final is lowered, it sounds like a cross-examination [1]. Therefore, an incorrect use of the phrase final intonation may cause misunderstandings in communication. Further, Japanese learners should acquire the phrase final intonation of questions [1].

There are many researches on prosodic analysis of questions in natural conversations. For example, in [2] the prosody of questions has been investigated in natural Swedish conversations, revealing that prosody differs according to the different types of questions. It was reported that yes-no questions generally have a falling intonation, whereas wh questions generally have a rising intonation in Swedish. In [3], terminal intonations of questions extracted from conversations between a doctor and a patient in Dutch have been examined. It was found that terminal F0 rises have higher values in the order of wh questions < yes/no-questions < declarative questions, in male speech. We also have investigated questioning prosody extracted from daily conversations in Japanese [4][5][6]. In our previous studies, we pointed out that the occurrence rates of non-rising tones increase in question types where the speaker assumes that the interlocutor does not necessarily hold the answer information.

These researches are mainly focused on the types of questions (e.g. yes-no questions, wh questions and so on). Although these approaches are effective to reveal prosodic variability of questions, one also should concern with another factor, namely interpersonal relationship between the dialogue partners. It is well-known that infant-directed speech (**IDS**) has some characteristic prosody relative to adult-directed speech (**ADS**) [7]. In [7], the prosodic modifications (such as higher mean-f0, f0-minimum and f0-maximum, greater f0-variability, shorter utterances and longer pauses) of IDS have been analyzed. In [8], it has been revealed that the voice quality parameter "NAQ" varies according to dialogue partner. As seen above, it is clear that the interpersonal relationships affect the speech prosody in various ways.

We have reported that phrase final tones of questions are varied according to the hierarchical relations or familiarity [4][5][6]. For example, if the interlocutor has higher status relative to speakers (e.g. senior staff), F0 of phrase final syllables are significantly lower than that of lower status (e.g. junior staff) or equal footings (e.g. colleague). These previous studies, however, were not conditioned about question types and phrase final part-of-speech information.

In this study, we aim to clarify the interpersonal factors which affect the prosodic modification of question-type utterances. Our previous research had made the result with combined all phrase finals, such as final particles, auxiliary verbs and without suffixes. Under that condition, it was unclear whether the difference in intonation was caused by morphological or interpersonal factors. Thus, in the present work, we constraint our analyses to questions ending with final particles.

2. Materials and methods

2.1. Speech data

We used a Japanese conversational speech database recorded in ATR/IRC labs, which was also employed in our previous studies [4][5][6]. The database has 69 dialogue sessions including 31 speakers (12 adult males, 15 adult females, 2 young child males and 2 young child females). "Young child" means pre-elementary school (age under 6) in this study. Each dialogue session has 10 to 15 minutes, face-to-face communication and no specific tasks (topic-free conversations). The total duration of speech data is about 900 minutes. Different types of interpersonal relationships between the dialogue partners are included, for example, mother/father-son/daughter, superior-subordinate, friends, first meets, and so on. Some speakers participated in multiple sessions talking with different interlocutors.

Simultaneous recordings of speech and EGG (electro-glottograph) signals are available in this database. Sampling rates are 16 kHz/16 bits. Audio data is recorded using directional microphone (Sanken CS-1). In part of the dialogue sessions, headset microphone data is also available.

2.2. Extraction of question-type utterances

The speech utterances in the database are segmented in phrase units based on pauses and clear pitch resets between phrases. 2~4 native Japanese speakers evaluated each utterance from the standpoint of turn-taking function (turn-keeping, turn-yielding, backchannel, and fillers). In the present work, we employed the turn-yielding utterances where 2 or

more annotator's judgments agreed. As a result, we got 4231 utterances from the database.

For each of the utterances, question types were annotated by 3 native speakers (research assistants). We used the same label set of question types used in our previous research [5][6]. The agreement rates (in terms of Kappa coefficients) among the question type labels by each pair of annotators were .77, .75 and .68. We use the utterances where 2 or more annotators agreed for the subsequent analysis.

We then separated these question types according to the degree of information request. The category of question types expressing higher degree of information request includes: *Yes-No questions* (n=1258), *Information request* (451), *Subjective feedback request* (317), *Repetition request* (39) and *Counter-questions* (16). The category of question types expressing lower degree of information request include: *Quiz-type questions* (n=8), *Agreement request* (979), *Open-type questions* (159), *Backchannel-type questions* (377), and *Self-questions* (161). We call the former category as "**HIR**: Higher degree of Information Request" and the latter category as "**LIR**: Lower degree of Information Request" hereafter. By the above procedure, we got 2081 utterance in HIR and 1684 utterance in LIR (total size is 3765).

2.3. Parameterization of phrase final tones

There are many acoustical candidates and categorizing method for phrase final tones. For example, the difference between the average pitch of the first half and the second half of the question is used in [1] for rough estimate of the rising or falling intonation. J_ToBI and X-JToBI are well-known prosodic labeling schemes [11][12]. The latter is particularly adjusted for spontaneous speech descriptions and used in the "*Corpus of Spontaneous Japanese*". In the present study, however, we employ a set of parameters proposed in [9][10] for description of phrase final tones, on the grounds that these parameters are based on human's perception. Five tone categories are used in the present work, namely Rise, Flat, Fall, Reset-Flat and Reset-Fall. Added to this, parameters can be automatically extracted by the procedures described below.

For phrase final duration, an automatic procedure was first realized, by using power and spectral change constraints [9]. 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.

For the pitch-related parameters, F0 values are first estimated based on a conventional method of picking peaks in the normalized autocorrelation function. F0 was extracted for both speech and EGG signals available in the database. However, the speech signals are used only when the EGG signals are not available in the database. For speech signal, the auto-correlation of the LPC inverse-filtered residue of the pre-emphasized signal was used, while for the EGG signal, the autocorrelation of a high-pass filtered signal with 70Hz cutting frequency (for removing DC and low frequency movements) was used. All estimated F0 values are then converted to a musical (log) scale.

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 [9]. 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 = F0tgt2b - F0avg2a

In this study, phrase finals are categorized as rise pitch movements when *F0move* > 1 semitone & duration > 100 ms, fall pitch movements when *F0move* < -1.3 semitone & duration > 100 ms, and flat pitch movements 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 = F0avg2a - F0avg_p

F0avg_p 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. A pitch reset is judged to be present when *F0reset* > 1 semitone.

The thresholds above were based on pitch movement perception experiments. From the utterances where agreement was obtained for the question type annotations, the ones where F0 could not be extracted and is unclear were removed from the analysis, resulting in a total of 2,226 utterances. We randomly picked 758 utterances for perception experiments. 3 annotators labeled each utterance according to the 5 tone categories (Rise, Fall, Flat, Reset-Fall and Reset-Flat). The agreement rates (in terms of weighted Kappa coefficients: Rise > Reset-Flat > Flat > Fall > Reset-Fall) among the tone category labels of the 3 annotators were 0.75, 0.61 and 0.57. The automatically classified tones under the above criteria were also checked for agreement with annotator's decision. The agreement rates (in terms of weighted Kappa coefficients) between above criteria and the 3 annotators were 0.68, 0.65 and 0.52.

Extracted phrase final tones according to the degree of information request, HIR and LIR, are shown in Table 1.

Table 1. Distribution of the phrase final tones classified according to degree of information request (HIR/LIR: Higher/Lower degree of Information Request). The numbers indicates the occurrences.

Tone	*HIR*	*LIR*	*Sum*
Rise	749 (62.8%)	319 (30.9%)	1068 (48.0 %)
Reset-Flat	88 (7.4%)	136 (13.2%)	224 (10.1 %)
Flat	25 (21.0%)	257 (24.9 %)	508 (22.8 %)
Fall	76 (6.4%)	152 (14.7 %)	228 (10.2 %)
Reset-Fall	29 (2.4%)	159 (16.4 %)	198 (8.9 %)
Sum	1193	1033	2226

2.4. Classification of phrase final morphemes

Linguistic information about the part-of-speech and morphemes appearing at phrase finals was taken into account when verifying the influence of tones. For example, the occurrence rate of rising tones at the last syllable of questions

was about 30 % in phrases ending with final particles, while it was over 60% in phrases not ending with final particles (e.g. nouns) [5].

For classification of phrase final morpheme and part-of-speech, we used the free part-of-speech and morphological analyzer software "MeCab" [13] (the dictionary used was UniDic [14]). Because of the spoken style of the utterances in the database, we checked and corrected the output from MeCab which mainly attempt to analyze written language. In this study, we selected the phrases ending with final particles.

Table 2 shows the distributions of the phrase final particles (top 3) according to the degree of information request, HIR and LIR. The total numbers of final particles were 391 in HIR and 660 in LIR.

Table 2. Distributions of the phrase final particles (top 3) according to the degree of information request (HIR and LIR). The number indicates the occurrences.

HIR		*LIR*	
/no (N)/	235 (60.1%)	/yone/	117 (17.7 %)
/ka/	115 (29.4%)	/ne/	110 (16.7 %)
/Qke/	14 (3.6%)	/na/	98 (14.8 %)

2.5. Criteria for analysis of interpersonal relationship

The aim of this section is to clarify if phrase final prosody could change according to the distance in the interpersonal relationship between the dialogue partners. For the subsequent analysis, we select the speakers according to the interpersonal relationship under the following criteria.

1. Speakers who have dialogue sessions with both young child and adult speakers.
2. Speakers who have dialogue sessions with his/her family member and others (only adult speakers).
3. Speakers who have dialogue sessions with acquaintances and with someone else for the first meeting (only adult speakers).

In case 1, we attempt to verify how *F0move* of phrase finals vary at questioning utterances of IDS and ADS in HIR and LIR. Under this condition, we got 5 speakers in HIR and 4 speakers in LIR. In cases 2 and 3, we attempt to verify how *F0move* or *F0reset* vary according to the degree of closeness (intimacy) in their interpersonal relationships. The purpose of case 2 is to verify whether the prosodic features of IDS appear in similar situations or not. Generally speaking, talking with a young child is a stress-free situation and the speaker's attitude tends to be friendly. This situation is thought to be similar to when talking with family members. Under this condition, we got 4 speakers in HIR and 5 speakers in LIR. The purpose of case 3 is to consider whether or not the speaker changes the prosody of questions when talking with a completely unknown person (for the first meeting). Under this condition, we got 7 speakers in HIR and LIR.

3. Analysis Results and Discussions

3.1. *F0move*: for young children or adults

Fig. 1 shows the average *F0move* in semitones when the interlocutor is young child or adult, in HIR and LIR conditions.

As **Fig. 1** indicates, when the interlocutor is a young child, *F0move* in the final particle is significantly higher than that of adult at HIR (Welch Two Sample t-test; t (48.1) = -2.8, p < .01). On the other hand, *F0move* for young child is significantly lower than that for adult at LIR (t (58.1) = 2.1, p < .05). No significant differences were found in *F0reset* in both conditions (HIR: *F0reset* for young child is -0.46, for adult is 0.98, t (37.9) = -0.4, *ns*; LIR: for young child is 1.97, for adult is 2.02, t (62.6) = 0.1, *ns*).

These results mean that the typical prosody to young child at HIR questions is exaggerated in rising tones. At LIR, on the other hand, we can notice that the falling tones are exaggerated for young child. The typical question prosody of LIR ending with final particles is a reset-flat (F0move > -1.3) when the interlocutor is an adult, while it becomes a reset-fall (F0move < -1.3) when it is a young child.

These results agree with a past study on Infant-directed speech (IDS), where it was mentioned that IDS has greater F0 variability than adult-directed speech (ADS) [7].

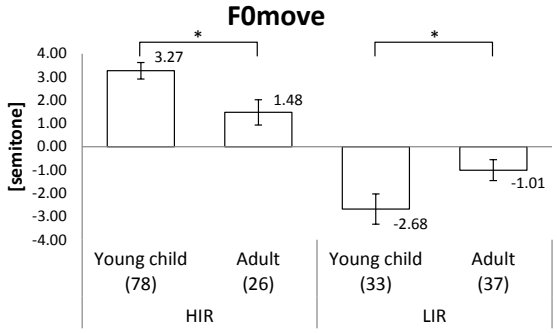

Fig. 1 Average *F0move* for young child/adult at HIR and LIR questions. The numbers in parenthesis indicate the numbers of occurrences.

3.2. *F0move*: for family members or others

Fig. 2 shows the average *F0move* in semitones when the interlocutor is family member or others, in HIR and LIR conditions.

As can be seen in **Fig. 2**, when the interlocutor is a family member, *F0move* of the final particle is significantly higher than that in others, at HIR (t (66.8) = 2.7, p < .05). On the other hand, *F0move* for family member is significantly lower than that for others at LIR (t (120.4) = 2.0, p < .05). *F0reset* is significantly different at LIR but not at HIR (HIR: *F0reset* for family member is -0.73, for others is 1.64, t (57.4) = -1.6, *ns*; LIR: for family member is 3.70, for others is 1.59, t (113.0) = 2.0, p < .05).

Fig. 2 indicates a similar tendency to **Fig. 1**. F0move is exaggerated for family members at HIR and LIR conditions. It is interesting to note that tone categories at both conditions for others are flat (i.e. -1.3 < *F0move* < 1). This result suggests that questions for others have the tendency to become a reset-flat tone. In other words, flattened F0 movements (i.e., repressed rising and falling degrees) in final particles appear in a relatively formal dialogue situation. In our previous study about questioning tones [4], it has been reported that if the interlocutor has higher status relative to the speaker (e.g. senior staff), *F0move* of phrase final syllables were significantly lower than that of lower status (e.g. junior staff)

or equal footings (e.g. colleague). And in [6], we showed that the occurrence rate of rise tone of question is lower when the speaker feels concern for the interlocutor (e.g. elderly) than that of "no concern" situation. These previous studies, however, were not conditioned on the part-of-speech of phrase finals. Generally speaking, speakers are required a formal speech style when talking with someone who has higher status or when feeling concern for the interlocutor. The results of the present work, where the part-of-speech of phrase finals is constrained to final particles, also support the results above.

It has been pointed out in [15] that a rising intonation at phrase finals in Japanese indicates a relatively heavy attitude of speakers. In other words, rising tone has a strong attitude of requesting an answer to the interlocutor. From this point of view, it seems reasonable to suppose that it is easier to speakers exposing their attitudes for family members. Consequently, the degree of *F0move* at HIR for family members becomes higher than that for others. In addition to the features of rising tones, the degree of falling can also be interpreted as showing strong attitudes at LIR condition (i.e. requesting agreement).

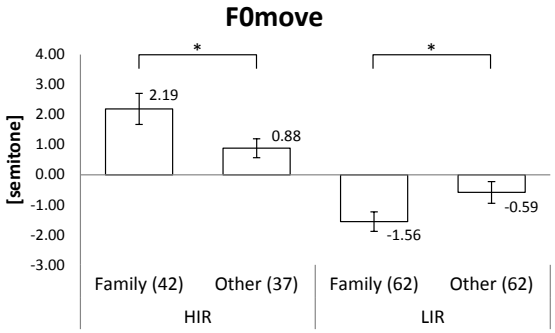

Fig. 2 Average *F0move* for family-member/others at HIR and LIR questions. The numbers in parenthesis indicate the numbers of occurrences.

3.3. *F0reset*: for acquaintances or first meets

Fig. 3 shows the average *F0reset* when the interlocutor is an acquaintance or a person of first meeting, in HIR and LIR conditions.

As **Fig. 3** indicates, when the interlocutor is a person of first meeting, *F0reset* is over 1 semitone (i.e. pitch reset is present) at both conditions (HIR and LIR). On the other hand, *F0reset* is lower than 1 semitone for acquaintance at both conditions. These differences are significant (HIR: t (91.5) = 2.1, $p < .05$; LIR: t (212.2) = 2.0, $p < .05$). In contrast, *F0move* didn't show significant differences in both conditions (HIR: *F0move* for acquaintance is 0.77, person of first meeting is 0.70, t (82.4) = -0.1, *ns*; LIR: for acquaintance it is -0.62, and for person of first meeting it is -1.23, t (213.1) = -1.7, *ns*).

As we mentioned in the previous section, speakers tend to be mindful of the questioning manner, according to the interpersonal relationship (to raise or drop in phrase final tone at HIR or LIR). Although significant differences did not appear, values of *F0move* are in the range of flat tones at HIR and LIR conditions (i.e. -1.3 < *F0move* < 1). It is thought that the effect of repressing the degree of rising or falling tones in relatively formal situations is reflected in this result.

In addition to this observation, **Fig. 3** showed another aspect of tone functions. Speakers clearly use pitch reset in questions at HIR and LIR conditions for completely unknown people (*F0reset* > 1). It has been described in [15] that pitch reset ("ukiagari-tyo" in their term) at phrase finals means speaker's light attitude. Although the participants of recordings can stop the dialogue at any time, they usually try to be friendly by continuing the conversation in smooth manner. Considering this situation and the description in [14], pitch reset at phrase final particles in question utterances can be interpreted as having functions of relieving the tension or showing a friendly attitude of the speaker.

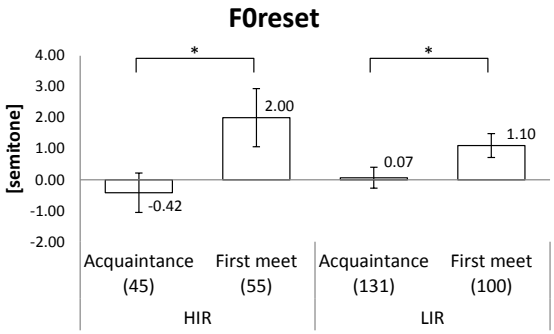

Fig. 3 Average *F0reset* for acquaintance/first meet at HIR and LIR questions. The numbers in parenthesis indicate the numbers of occurrences.

4. Conclusions

We conducted quantitative analysis on free conversational dialogue database, for clarifying the functions of phrase final tones in questions, from an interpersonal relationship viewpoint. Question-type utterances were classified into two categories on the basis of degree of information request (HIR and LIR). The analysis results indicated that:

1. The degree of F0 rising at HIR questions and the degree of F0 falling at LIR questions are exaggerated when talking with a young child, in comparison to when talking with an adult.
2. Similarly, the degree of F0 rising at HIR questions and the degree of F0 falling at LIR questions are exaggerated when talking with a family member, in comparison to when talking to others.
3. The degree of F0 reset at the phrase final syllable is higher when talking with a person of first meeting at both HIR and LIR questions, in comparison to when talking with an acquaintance.

It is inferred from these results that an increase of raising degree at HIR question and lowering degree at LIR question can be regarded as an attitude of intimacy to the interlocutor. In addition, the presence of pitch reset at the final particles can be interpreted as having functions of relieving the speaker's tension.

5. Acknowledgements

This work was supported by JSPS KAKENHI Grant Numbers 23680019, 25884099. We also thank all research assistants of the ATR group who contributed with the annotations.

6. References

[1] Ayusawa, T., "Acquisition of Japanese accent and intonation by foreign learners", *Journal of the Phonetic Society of Japan*, Vol.7 No.2, 47-58, 2003 (in Japanese).

[2] Strömbergsson, S., Edlund, Jens., and Hous, David., "Prosodic measurements and question types in the Spontal corpus of Swedish dialogues", Proceedings of *INTERSPEECH* 2012., 2012.

[3] Heuven, V. J., van, Hann, J. and Kirsner, R. S., "Phonetic correlates of sentence type in Dutch: Statement, question and command", Proc. ESCA International Workshop on Dialogue and prosody, 35-40, 1999.

[4] Hatano, H., Arai, J. and Ishi, C. T., "Analysis of factors which contribute to choice of questioning prosody in natural conversation", Proceedings of *The Spring Meeting of the Acoustical Society of Japan*., 429-430, 2013 (in Japanese).

[5] Hatano, H., Kiso, M. and Ishi, C. T., "On the factors which contribute to decision of phrase final intonation of questioning utterance in natural conversation", Proceedings of *The Twenty-Seventh General Meeting of the Phonetic Society of Japan*., 59-64, 2013 (in Japanese).

[6] Hatano, H., Kiso, M. and Ishi, C. T., "Analysis of factors involved in the choice of rising or non-rising intonation in question utterances appearing in conversational speech", Proceedings of *INTERSPEECH 2013*., 2013.

[7] Fernald, A., Taeschner, T., Dunn, J., Papousek, M., de Boysson-Bardies. and Fukui, I., "A cross-language study of prosodic modifications in mothers' and fathers' speech to preverbal infants", J. Child. Lang., 16:477-501, 1989.

[8] Campbell, N. and Mokhtari, P., "Voice Quality: the 4th prosodic dimension", Proceedings of the *ICPhS'03*, 2417-2420, 2003.

[9] Ishi, C. T., "Perceptually-related F0 parameters for automatic classification of phrase final tones", *IEICE Trans. Inf. & Syst.*, Vol. E88-D, No.3, 481-488, 2005.

[10] Ishi, C. T., "The function of phrase final tones in Japanese: Focus on turn-taking", *Journal of the Phonetics Society of Japan*., Vol.10 No.3, 18-28, 2006.

[11] Venditti, J., "The J_ToBI model of Japanese intonation", in Jun, S. A [Ed] *Prosodic typology: The phonology of intonation and phrasing*, 172-200, New York: Oxford University Press, 2005.

[12] Maekawa, K., Kikuchi, H., Igarashi, Y. and Venditti, J., "X-JToBI: An extended J_ToBI for spontaneous speech", Proceedings of *ICSLP2002*, 1545-1548, 2002.

[13] Downloadable: http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html, accessed on 12 Dec 2013.

[14] Downloadable: http://sourceforge.jp/projects/unidic/, accessed on 12 Dec 2013.

[15] Kawakami, S., "Bunmatsu nado no joosyootyoo ni tsuite", Kokugo-kenkyuu, 16, 2-46, 1963 (in Japanese).