Japanese Politeness and Suprasegmentals - a Study based on Natural Speech Materials

Mika Ito

Department of Theoretical and Applied Linguistics
University of Edinburgh, United Kingdom
mika@ling.ed.ac.uk

Abstract

This paper discusses some of the problems regarding the unnaturalness of speech data currently used in research on oral Japanese politeness and proposes improved techniques. Two experiments were carried out. First, to extract natural unscripted utterances, within a specific vocabulary and context, the experimental design of the Map Task was employed with the population controlled for social status. Second, to elicit a perception of the level of politeness of spoken Japanese, a rating experiment of formality was conducted. Raters scored the degree of formality of lexically similar but non-identical tokens without any manipulation. Magnitude Estimation (ME) was employed to reflect the perceived formality. Results from the production side showed that raising the fundamental frequency (F_0) or changing speech rate are not always correlated with increasing formality. In the perception experiments, listeners did not show good agreement in their judgement of formality for most stimuli, though some trends were observed in stimuli. Although the speakers produced the utterances with appropriate honorifics, neither F_0 nor speech rate seem to be dominant factors for judging formality. Therefore there is a need to explore the interaction between lexical cues and acoustic cues for conveying formality.

1. Introduction

There have been some studies on the acoustic contribution to paralinguistics, however this area has not been well explored. Brown and Levinson [4] stated that the usage of paralinguistics seems to share a number of universal characteristics among cultures and language systems, but the strategies to express politeness are specified according to social background. They also suggested that sustaining high pitch may implicate self-humbling and thus deference, as a part of "negative politeness"(the expression of restraint). Ohala [10] [11] also associated a high F_0 usage with politeness, in the context of frequency code, as a near-universal pattern. Therefore, analyzing Japanese politeness in speech may contribute to the study of universal features of politeness. Recent studies (e.g. Hirose et al. [6] [13], and Ofuka et al. [9]) associate spoken Japanese politeness with speech rate. Since they collected polite/non-polite utterances with acting, and manipulated stimuli were used for perception experiments, naturalness of the utterances was not guaranteed. This problem should be considered seriously to achieve fair evaluation on expressive speech, as was pointed out by Campbell [5]. Also, as previous studies used a linear scale rating with a limited scale, scores given may not reflect the perceived degree of politeness. In this study, to reveal the role of suprasegmentals expressions of politeness in a natural manner, formality (between people of different social status) was focused on as a controllable target, and the following methods of data collection were used. In order to collect natural speech, a Map Task was used in which the status relationship between participants was controlled. This was intended to elicit the adoption of politeness strategies in dialogue, according to the social status of the participants. The higher status participant is predicted to use either a formal or an informal style, but the lower one is predicted to use only formal, honorific language [14]. In a perception experiment, stimuli were carefully chosen so as to be lexically similar, but not identical and were presented without manipulation. Magnitude Estimation was then employed for the rating method so that distances between rated stimuli were maintained.

2. Speech Data Collection

The Map Task was used to elicit spontaneous speech data containing politeness and role information in a controlled setting. The Map Task was originally conceived at the University of Edinburgh, HCRC [1], and a Japanese corpus based on the Map Task was collected at Chiba University [2]. The Map Task works as follows: the two participants in the conversation each have a map showing a variety of named landmarks. The maps may differ slightly in detail. Neither speaker can see the other's map. One map (the "Instruction Giver's" map) has a route marked on it. The task is for the Instruction Giver to explain to the Instruction Follower where the route passes, referring to the various landmarks along the way - accurately enough for the Instruction Follower to reproduce the route on his or her own map. The benefits of using the Map Task include the following. First, the formality can be maintained in a dialogue between participants. Since the relationship between participants alternates, the effect of relative status change is hoped to elicit the production of different suprasegmental features. Second, the effects of role change (Giver vs. Follower) on suprasegmentals can also be observed. Finally, the Map Task enables us to compare suprasegmentals of lexically similar utterances. Since one aim of the Map Task is to make participants concentrate on their task, their vocabulary and intentions for every utterance fall within a certain range. By analysis of these utterances, which occur frequently in the dialogues, it is possible to compare suprasegmental features.

2.1. Subjects and Corpus

Four male adults in the same research group but with different social status were recruited for this recording, to maintain familiarity but to control for the status relationship. They were asked to participate in the task in pairs, with a higher status subject and

Table 1: F_0 and Articulatory Rate of the speakers.

Speaker	Status	F_0(Hz)(s.d.)	Articulatory Rate(mora/s) (s.d.)
1	higher	135.58 (34.97)	8.503 (1.940)
	lower	136.45 (35.12)	8.495 (2.389)
2	higher	118.88 (29.79)	7.812 (1.734)
	lower	129.90 (29.78)	8.178 (1.927)

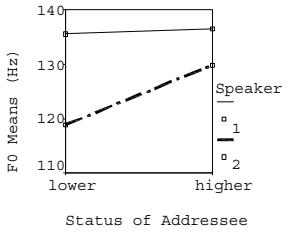

Figure 1: F_0 means of the Tokyo-dialect speakers.

a lower status subject taking the role of an Instruction Giver or Instruction Follower, and vice versa. All materials were digitally recorded (16bit, sampling frequency = 48kHz, stereo) on DAT with a close-talking microphone and a DAT channel for each participant, and down-sampling to 16kHz for further analysis. Two speakers out of four satisfied the following conditions: 1) Both spoke to both higher status and lower status, 2) both were native speakers of the Tokyo dialect. Target utterances from these two speakers were extracted successfully, and were used for further analysis and experiment.

2.2. Method of Analysis

In this study, as in previous investigations, the overall tendency of F_0 and speech rate were measured. For F_0 estimation, a pitch tracker using a normalized cross correlation function algorithm [16] was employed with a 10ms interval, and extracted F_0 values were statistically processed. For speech rate, articulatory rate (mora/sec) [8], which excludes not only silent pauses but also filler pauses (e.g. disfluencies like /e-to/), were calculated.

2.3. Results

First, the speakers successfully used honorifics appropriate to the relationship between participants. Figure 1 and Table 1 show the statistics of the F_0 of utterances for each speaker, in the Instruction Giver role. Their overall F_0 behaviors were completely different when affected by status change of the addressee. Speaker 2 showed his formality by raising his F_0 in both roles, but speaker 1 did not show his formality by raising his F_0. Figure 2 and Table 1 show the statistics of the articulatory rate of utterances for each speaker, in the Instruction Giver role. Means of speaker 2 show difference between the two statuses. Overall, the data suggest that speaker 2 showed his formality with the suprasegmental cues proposed in previous studies, but speaker 1 showed his formality in a different way.

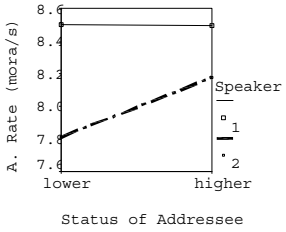

Figure 2: Articulatory Rate of the Tokyo-dialect speakers.

3. Perception Experiment

The purpose of the perception experiment was to determine if listeners can detect acoustic differences in formality as predicted by experimental design. Because of the results of the production experiments, we expect listeners to detect formality differences for speaker 2 since he showed a difference in F_0 means and speech rate. The perception experiment will also determine if listeners detect differences in formality that could be due to acoustic cues besides F_0 and speech rate. These formality judgement experiments were carried out using tokens of the phrase "/wakarimasita/", extracted from the Map Task data set, as stimuli. Since many researchers have pointed out the effect of dialect on the perceived Japanese speech (e.g.,[12]), native speakers of the Tokyo dialect were recruited as raters for the formality perception experiments.

3.1. Method

3.1.1. Rating - Magnitude Estimation

To measure formality in perceived speech, the Magnitude Estimation method (ME) was employed. Magnitude Estimation was originally developed for psychophysics [15], to associate physical amount with perception of acoustics. In psycholinguistics, Bard et al. [3] introduced ME to measure linguistic acceptability. The advantages of ME follow. One common rating method is Linear Scaling: using a shown scale and plotting a point. But there is a problem with expressing distinctiveness between stimuli. For example, if we give the maximum number to one stimulus, and a later stimulus is stronger, there will be no available value to show distinctiveness. With ME, there is no restriction of values, so all differences can be expressed quantitatively, and become measurable, in contrast with the Linear Scaling Method. Secondly, if we wish to investigate distinctiveness, one of the popular methods is the comparative judgement method, involving presentation of stimuli in pairs. But comparative judgement needs all possible pair-wise combinations of stimuli to be presented. Another problem with the comparative judgement method is that it does not allow for measurement of the degree of distinctiveness. Using ME, subjects compare each stimulus with one modulus, and therefore the number of judgements for subjects will be reduced, compared with comparative judgement. Thirdly, ME may be able to explain the correlation between acoustical amounts and the degree of formality in a quantitative manner. According to Stevens [15], this correlation is well explained using a Power Function. In this section, estimated magnitude and other acoustical amounts are evaluated on a logarithmic scale, regarding this power function correlation. The main disadvantage of ME is that subjects are used to Linear Scaling methods. Instructions need to be given carefully

so as to allow subjects to rate the comparative magnitude of the stimuli and the modulus easily.

3.1.2. Subjects

To avoid the effects on prosodic features from dialect and regional cultural background, native speakers of the Tokyo dialect were recruited to participate in this experiment. A total of 23 people participated in this experiment. A group of 18 subjects were students without any work experience, and the other 5 subjects had work experience. The majority of them were born and brought up in the Tokyo area, and all participants have been residents in the Tokyo area for more than four years.

3.1.3. Materials

In this experiment, the aim was to compare listeners' reactions to differences in acoustic features, so the utterances, which are lexically similar but not identical, were carefully chosen from the previous speech collection, so as to avoid semantic contextual influences and disfluency. Thus, lexically similar but not identical utterances, which contain a phrase of /wakarimasita/ (6-mora word of accent type 4: which has accentual rise between the first and the second mora, and has accentual fall between the fourth and the fifth mora), were selected from the previous corpus. Since the meaning of this phrase is "I understand", and participants produce this phrase to confirm of receipt of information, so intentions were restricted. These utterances did not have disfluency. Therefore, a total of 18 tokens consisting of nine identical tokens for each of two speakers were extracted as test stimuli. In addition to the Map Task recordings, those two speakers were also recorded reading the phrase /wakarimasita/. These recordings of read speach were used as the modulus for the sets of magnitude estimation.

3.2. Procedure

Subjects were tested individually. They heard a set of stimuli from the collection of materials after a modulus, and were required to estimate the magnitude of formality of each stimulus. They were instructed to give a number greater than one if a stimulus sounds more formal than the modulus, or to give a number between zero and one if a stimulus sounds more informal than the modulus, so as to express its relationship with the modulus as a ratio. For example if the stimulus sounded twice as formal as the modulus, the subjects were told to answer "2", while if the stimulus sounded twice as informal as the modulus, they were told to answer "0.5 (=1/2)". Each input field was displayed on the PC, after a stimulus, and the subjects were asked to give the estimated score. A set of stimuli consisted of nine phrases of /wakarimasita/ from a speaker. Each session started with a speaker's modulus followed by a set of nine stimuli. The subjects were presented with the sets, which alternated between speaker 1 and speaker 2. Each set of nine stimuli was presented in two different randomized orders, so as to avoid the influence of presenting order. Finally, the target phrases in whole utterances from both speakers were presented once. Thus a total of 60 stimuli were presented. After completing ratings, the subjects were asked to submit the data, and the data were collected automatically online.

3.3. Results from Magnitude Estimation

For each subject, correlation coefficients were computed between the set of ratings from the first session and the set of ratings from the second session. Twenty out of twenty-three subjects showed correlation coefficients ($r \geq 0.60$) with a significance level of ($p < 0.01$) between the two sessions. These twenty subjects' data were employed for further statistical analysis. The correlation coefficient for judging all stimuli when they were presented alone is not significantly high (r=0.774, $p < 0.01$), and the coefficient for the speaker 1 stimuli (r=0.792, $p < 0.01$) is higher than the coefficient for the speaker 2 stimuli (r=0.742, $p < 0.01$). These results show that it is difficult to conclude that the subjects used acoustic features as their cues for the formality judgement. And, as is shown in these correlation coefficients, the listeners did not always judge the status of addressee successfully, especially for speaker 2 (Figure 3).

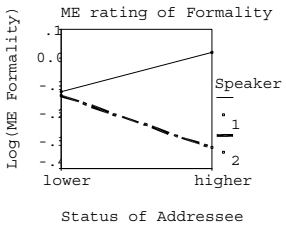

Figure 3: Formality Ratings and Status of Addressee

3.4. Analysis -correlation of acoustics and formality-

The correlation coefficients between acoustic features and formality ratings within subjects were computed. No acoustic features measured in the previous section showed a good correlation with listeners' judgement ($p < 0.05$). Six stimuli for speaker 1, and eight for speaker 2 showed a significant correlation between subjects ($p < 0.05$). The acoustic features of the stimuli subset of 2 stimuli for each speaker which were rated as most formal/informal were analyzed. F_0 means and first formant bandwidth (BW1), showed a good correlation with ratings for both speakers, though observed tendencies were opposite to the trends of previous studies (Figure 4, Figure 5, Figure 6). The F_0 mean was lower and BW1 was wider when rated formal, though previous studies said that higher F_0 was associated with giving deference and lax voice with wider BW1 was associated with showing intimacy [7]. An examination of the stimuli which were judged as having intermediate formality showed, however, that the relationships beteeen formality and both F_0 and BW1 is non-linear. Overall it is questionable to say that any particular prosodic feature successfully predicts their ratings, even though the results from the most formal/informal stimuli showed that certain acoustic features might have been employed in judging their formality.

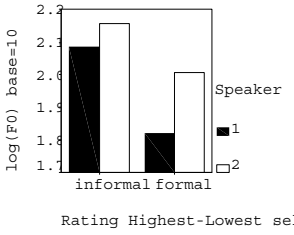

Figure 4: F_0 means of the most formal/informal stimuli

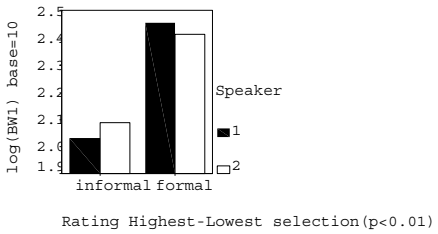

Figure 5: Bandwidth of F_1 of the most formal/informal stimuli

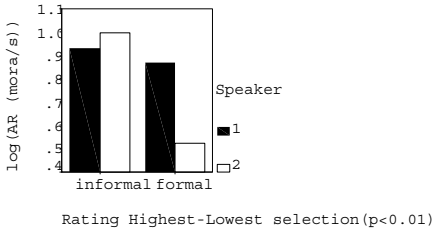

Figure 6: Articulatory Rate of the most formal/informal stimuli

4. Discussion

In speech data collection, two tendencies were observed. First, the speakers managed to use honorific expressions properly and utterances reflected the social relationship between participants successfully. Second, raising F_0 or controlling speech rate may be one strategy to indicate formality to a higher status addressee, though this control of suprasegmentals may not always occur. From rating experiments, it is questionable to say that the subjects rated the formality of stimuli using acoustic cues without contexts. However, the results from the most formal/informal stimuli, with agreement of listeners, suggest that F_0 or BW1 might be a cue for conveying formality, but it needs more investigation. One of the problems is that the number of speakers is not great enough to show tendencies of formality apart from speaker's voice characteristics. Therefore observation of data collected from a greater number of speakers is necessary. At the same time, a further exploration of the measurement method of voice quality is needed. Because of transitions in each segment, observation of BW1 in spontaneous speech is difficult. Another problem to be explored is the characteristics of politeness in Japanese speech from sociolinguitic point of view, because for giving deference, there might be another strategy to be taken for showing politeness in spoken Japanese. Overall, the interaction between lexical cues and paralinguistic cues is important, and there may be individual differences in paralinguistic strategy. Using acted stimuli, these differences will not be revealed. It is important to use natural stimuli from speech with a certain situation control for investigating these phenomena.

5. Acknowledgements

This research is based on research completed as a pilot study of my PhD work at the University of Edinburgh. I would like to thank my supervisors, Prof. Ladd and Dr. Turk for their helpful comments from time to time. I would like to thank Prof. Tsuchiya and Dr. Horiuchi (Chiba University), for allowing me access to the materials of their Map Task corpus, and giving many helpful technical advices before the recordings. I'm also grateful for Prof. K. Hirose (the University of Tokyo) to allow me access to his laboratory's facilities and staff through these experiments. However, any mistakes that remain are my own.

6. References

[1] Anderson, A.H.; Bader, M.; Bard, E.G.; Boyle, E.; Doherty, G.; Garrod, S.; Isard, S.; Kowtko, J.; McAllister, J.; Miller, J.; Sotillo, C.; Thompson, H.S.; Weinert, R.; 1991. The HCRC Map Task Corpus. *Language and Speech.*, 34, 351-366.

[2] Aono, M.; Ichikawa, A.; Koiso, H.; Satoh, S.; Naka, M.; Tutiya, S.; Yagi, K.; Watanabe, N.; Ishizaki, M.; Okada, M.; Suzuki, H.; Nakano, Y.; and Nonaka, K ; 1994. The Japanesse Map Task Corpus: An interim report (in Japanese). *Spoken language understanding and discourse processing, Japanese Society for Artificial Intelligence.*, SIG-SLUD-9402, 25-30.

[3] Bard, E.G.; Robertson D.; Sorace, A.; 1996. Magnitude Estimation of Linguistic Acceptability. *Language.*, 72, 32-68.

[4] Brown, P.; Levinson, S.; 1987. *Politeness: Some universals in language usage.* Cambridge: Cambridge university Press.

[5] Campbell, W.N.; 2000. Databases of Emotional Speech. *ESCA Workshop on Speech and Emotion*, Belfast, 34-37.

[6] Hirose, K.; Kawanami, H.; Ihara N; 1997. Analysis of Intonation in Emotional Speech. *ESCA Workshop on Intonation: Theory, Models and Applications*, Athens Greece, 185-188.

[7] Laver, J.; 1980. *The phonetic description of voice quality.* Cambridge: Cambridge university Press.

[8] Laver, J.; 1994. *Principles of Phonetics.* Cambridge: Cambridge university Press.

[9] Ofuka, E.; McKeown, J.D.; Waterman, M.G.; Roach, P.J.;2000. Prosodic cue for rated politeness in Japanese speech. *Speech Communication*, 32, 199-217.

[10] Ohara, J.J. ;1984. An ethlogical percpective on common cross-language utilization of F_0 of voice *Phonetica*, 41, 1-16.

[11] Ohara, J.J. ;1996. Ethlogical Theory and the Expression of Emotion in the Voice *Proc. ICSLP*, Philadelphia, 3, 1812-1815.

[12] Otake, T.; Cutler, A. ;1999. Perception of suprasegmental structure in a non-native dialect. *Journal of Phonetics*, 25, 229-253.

[13] Sakata, M.; Hirose, K; 1995. Analysis and synthesis of prosodic features in spoken dialogue of Japanese. *Proc. 4th EUROSPEECH*, Madrid, 9, 1007-1010.

[14] Shibatani, M.; 1990. *The languages of Japan*. Cambridge: Cambridge university Press.

[15] Stevens, S.S.; 1969. On predicting exponents for cross-modality matches. *Perception and Psychophysics*, 6, 251-256.

[16] Talkin, D.; 1995. A Robust Algorithm for Pitch Tracking (RAPT). In *Speech Coding and Synthesis*, Kleijn, W. B. and Paliwal, K. K. (Eds.) New York: Elsevier.