Proceedings of the Postgraduate Conference 2001-Department of Theoretical and Applied Linguistics, The University of Edinburgh

Rating experiments of Spoken Japanese Politeness

Mika Ito
(mika@ling.ed.ac.uk)

ABSTRUCT

To observe perception side of spoken Japanese politeness, a rating experiment of formality was conducted. Lexically similar tokens, addressed to higher/lower status, produced by two speakers of Tokyo dialect were presented as stimulus, so that various utterances in degree of formality, as a kind of politeness, were to be rated. Considering dialects' influences, native listeners of two dialect groups, one from Tokyo area and the other from Kansai area were asked to rate the degree of formality. To reflect distance of degree in ratings, Magnitude Estimation (ME) was employed. Results show following findings. Firstly, listeners can rate magnitude of formality. Secondly, dialects' influences should be considered carefully. Thirdly, ratings do not reflect addressee's status directly. Finally, speech rate does not seem to be a dominant factor for conveying formality.

1 INTRODUCTION

Even recent speech technology has progressed dramatically in speech synthesis and speech recognition, it is still controversial to regard synthesised speech as "natural". In case of spoken Japanese, native speakers of Japanese are supposed to use appropriate honorific illocutions, which reflects a relationship between speakers. This social requirement also restricts the type of paralinguistics under speakers' control.

According to Brown and Levinson (1978, 1987), the tendency or likeliness of the taken strategies to express "Politeness" varies according to the social background. However, Brown and Levinson explained following facts. Firstly, the relationship between the intentions and the strategies in behind is universal. Secondly, the relationship between a speaker (S) and an addressee (H) is used to specify the strategy for realisation of politeness. And thirdly, the illocution and the usage of paralinguistics seem to share a number of universal characteristics among cultures and language systems. Therefore, surveying on Japanese politeness still has a room to contribute universal features of politeness. Recent studies of spoken Japanese politeness were explored by several researchers (e.g. Kawanami et al. (1997), and Ofuka et al. (2000)) . Since they collected polite/non-polite utterances by acting, naturalness of utterances were not guaranteed. And previous studies used linear scale rating so that distinctiveness between ratings of similar tendencies was not seriously considered.

In present study, to reveal expressions of politeness in natural manner, I focus on "Formality" as a controllable target. I propose following two methods of data collection, for speech production and speech perception. As a speech production collection, I employed Map Task with controlling participants' relationship so as to assure politeness strategies with different degree of formality to be taken in dialogues, and to restrict vocabulary within a certain

extent. As a perception side, I used Magnitude Estimation so that a distance between rated stimulus, which have similar tendency kept maintained.

2 SPEECH DATA COLLECTION

In order to collect data containing politeness and role information in a controlled setting, the Map Task was conducted. The Map Task was originally conceived at the University of Edinburgh, HCRC in 1991, and the Japanese version was conducted by Aono et al., at the Chiba-university in 1994.

The benefits of using the Map Task include the following. First, the politeness relations can be extracted from a controlled dialogue between participants. Since the roles of "Instruction Giver" and "Instruction Follower" can alternate, the effects of role change would give different phrases to be produced and be studied. Second, using utterances from the Map Task enables us to compare fundamental frequencies and some number of characteristics in voice quality. Since one aim of the "Map Task" to make participants concentrate on their task, then their vocabulary and their intentions for every utterance fall within a certain range. By analysis of the accent types and phonetic features of utterances, which occur frequently in the dialogues, it is possible to compare acoustic features. Thus, by controlling the status according to the relationship with the other participants, the utterances from the same speaker, which are lexically similar but not identical, but with different paralinguistics can be compared.

All materials were digitally recorded (16bit, sampling frequency = 48kHz, stereo) on DAT with a close-talking microphone and a DAT channel per each participant, and after down-sampling to 16kHz, they were stored as .sd-format files so as to be used for further analysis using ESPS/XWaves.

One academic staff (speaker ID: ASM) and one postgraduate student (speaker ID: PGT) satisfied two conditions as follows; 1) should talk to higher status and lower status, 2) native speaker of Tokyo dialect, who was brought up in the Tokyo area (where Tokyo dialect is spoken). Target utterances from these two speakers were extracted successfully, and were used for further experiment and analysis.

3 PERCEPTION EXPERIMENT

Speakers may adjust speech rate and fundamental frequency to change the formality level of their speech style, but there may be cases where this adjustment does not occur at all. Therefore, a question arises from the speakers' point of view: are these suprasegmentals main factors which convey formality or not? This experiment is an approach to answer the question above, based on the perception of formality. From the point of view of perception, there are two questions. First, do humans use acoustic parameters for detecting formality? If this is not so then acoustic parameters would convey little information about a degree of formality and humans would use only illocutions, expressions and context for detecting formality. Second, if we suppose that humans use acoustic parameters, are they the suprasegmentals (e.g. pitch, speech rate) or are they other acoustic parameters? To answer these two questions, it is necessary to conduct a pilot experiment concerning the perception of formality. These formality judgement experiments have been carried out based on the speech materials derived from the speech data collection using the Map Task.

Many researchers have pointed out the effect of dialect on speech on the perception side. Otake and Cutler (1999) revealed that there are some differences in perception of prosodic features influenced by dialects, even though Japanese are fairly exposed to Tokyo dialect. It is reasonable to think that this tendency might affect the perception of politeness. Regarding these findings, native speakers consisting of two groups of different dialects were recruited in Tokyo and Kansai for formality perception experiments.

3.1 Method

3.1.1 Magnitude Estimation

To measure formality in perceived speech, the Magnitude Estimation method (ME) was employed. Magnitude Estimation was originally developed by Stevens (1957, 1969, 1975) for psychophysics, to relate physical amount and perception of acoustics. In psycholinguistics, Bard et al. (1996) 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. In the case of using 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 would like to look into distinctiveness, one of the popular methods is the comparative judgement method, presented in pairs. But comparative judgement needs all pair-wise combinations of stimuli to be presented. Another problem comparative judgement has is that the degree of distinctiveness is not measurable with comparative judgement. 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 (1969), this correlation is well explained using Power Function. In this section, estimated magnitude and other acoustical amounts are evaluated on a logarithmic scale, regarding this power function correlation.

There is on the other hand, a problem to be solved, arising from the fact that subjects are used to Linear Scaling methods. Therefore 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 consider the effects of dialects' prosodic features and regional cultural background, two groups of native speakers were recruited to participate in this experiment. The subjects consisted of two native speaker groups, Tokyo dialect and Kansai dialect.

3.1.2.1 Tokyo dialect group

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 have 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.

3.1.2.2 Kansai dialect group

A total of 32 people participated in this experiment. A group of 24 subjects were students without any work experience, and the other 8 subjects had work experience. The majority of them were born and brought up the in Kansai area, and all participants have been residents in the Kansai-area for more than four years.

3.1.3 Materials

In this experiment, the aim to compare listeners' reactions with acoustic features difference, so the utterances, which are lexically similar but not identical, were carefully chosen as materials from previous speech collection, given the following considerations:

- Semantic and contextual influences should be avoided.
- The utterances, which can be produced as independent intonational phrases, should be chosen to avoid considering the phrasal position.

- Intonational phrases, which have accentual rise and fall, are desirable.
- The utterance should not have disfluency.
- The utterances, which were produced, while addressing both to senior and to junior, should be selected.
- The utterances, which are phonetically in similar environment, are suitable to observe voice quality and assimilation.

Thus, lexically similar but not identical utterances, which contain a phrase of /wakarima'shita/ (6 morae 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), are selected from the previous speech collection. Since the meaning of this phrase is "I understand", and participants produce this phrase to confirm of receipt of information, so intentions were focused. These utterances did not have disfluency. Therefore, a total of 18 tokens consisting of nine identical tokens from two speakers, ASM and PGT, were extracted as test stimuli. Nine occurrences of this utterance from each of two speakers are selected, for a total of 18 utterances. Apart from the Map Task, speakers PGT and ASM were recorded reading the script of /wakarima'shita/. These recordings were used as the modulus for the sets of magnitude estimation.

3.2 Procedure

3.2.1 Magnitude estimation

Subjects are tested individually. They hear a set of stimuli from the collection of materials after a modulus, and are required to estimate the magnitude of politeness of each stimulus. For example, subjects are instructed to award a number greater than 1 if a stimulus is politer than the modulus, or to award a number between zero and one if a stimulus sounds more informal than the modulus, so as to convey multiple fraction. For example if the stimuli sounds twice as formal as the modulus, then put 2, while if the stimuli sounds twice as informal as the modulus, then put 0.5 (=1/2). Each input field is displayed on the PC, after a stimulus, and the subjects are asked to give their estimated score. A set of stimuli consists of nine phrases of /wakarima'shita/ (="I understand it") from a speaker. Each session starts with a speaker's modulus followed by a set of nine stimuli. The subjects are presented with sets, which alternate between speaker PGT and speaker ASM. Each set of nine stimuli is presented in two different randomised order, so as to avoid the influence of presenting order. Finally, the target phrases in whole utterances from both speakers are presented once. Thus a total of 60 stimuli are presented. After completing ratings, the subjects are asked to submit the data, collected automatically online.

3.3 Results –tendencies from Magnitude Estimation-

3.3.1 Dialects' effect on data reliability

After collecting data, correlation coefficients between the set of ratings from the first session and the set of ratings from the second session within each subject were computed. Figure 3.3.1 shows a histogram of the distribution of correlation coefficients computed for both dialect groups, between their first session and their second session.

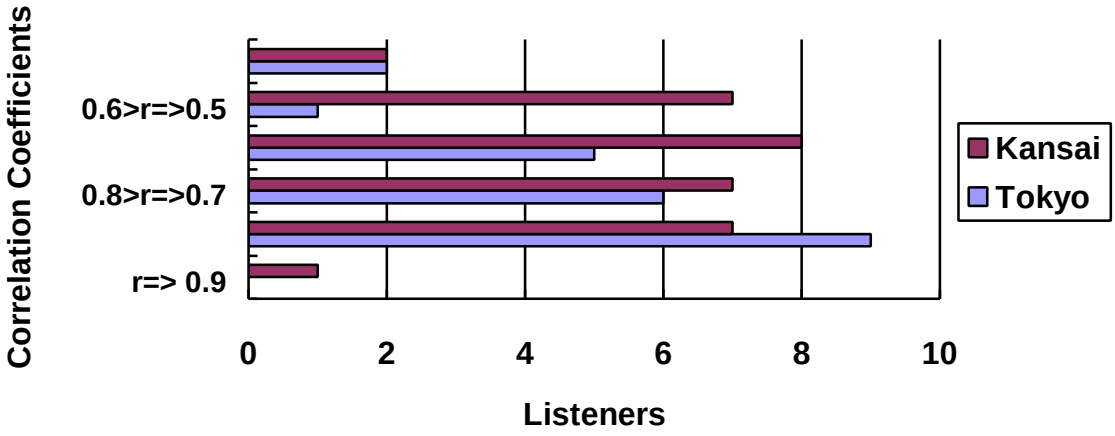

Fig. 3.3.1 Correlation Coefficients between 1st session and 2nd session

Some subjects' data, which does not show a strong enough correlation (r<0.60) between the two sets. These were not regarded as reliable, and so three out of 23 subjects from Tokyo dialect group and 12 out of 32 subjects from Kansai dialect group were excluded from further statistical analysis. As some subjects mentioned that they lacked of confidence in their judgement because they were not familiar with Tokyo dialect, this result shows that Tokyo dialect speakers are likely to show higher consistency in their judgement than Kansai dialect speakers are. This might be reasonably explained as the effect of dialect on the perception of formality. Therefore the data derived from a total of 20 Tokyo dialect listener (16 students and 4 work experienced adults) and a total of 20 Kansai dialect listener (16 students and 4 work experienced adults) were used for analysis.

3.3.2 Analysis on Tokyo dialect Speaker

Figure 3.3.2 (left) shows the Tokyo dialect group's responses to a set of stimuli from speaker PGT, whose suprasegmentals tended to differ according to addressee's status.

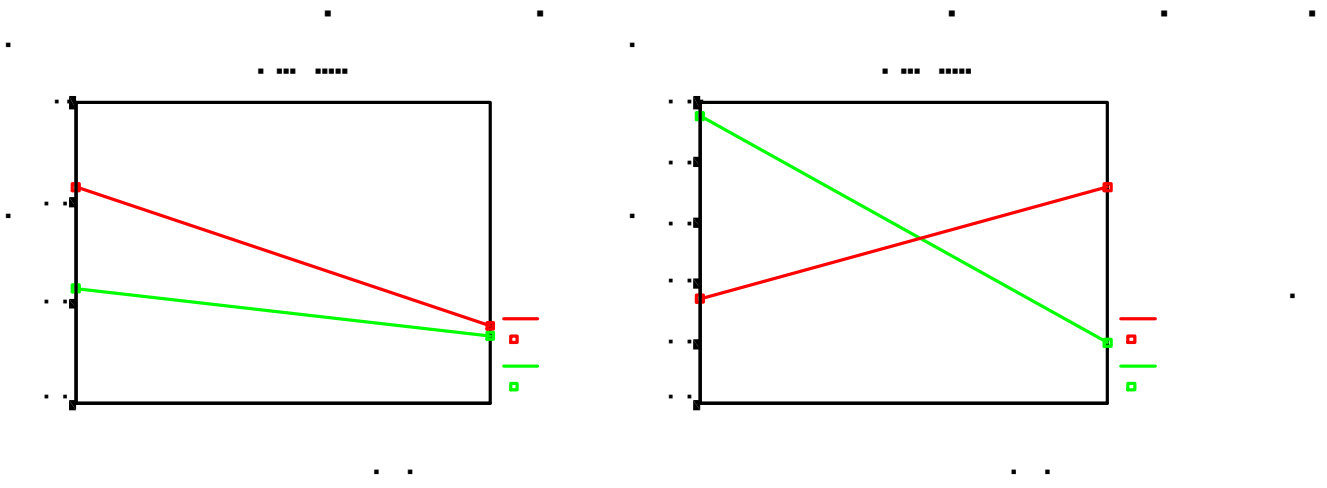

Fig. 3.3.2 Tokyo dialect listeners' ratings : speaker PGT (Left) and ASM (Right).

Figure 3.3.2 (right) shows the Tokyo dialect group's responses to a set of stimuli from speaker ASM, whose suprasegmentals did not correspond with the addressee's status.

These results from the Tokyo dialect group show that they used acoustic features as clues for judging formality, in subject's response depending on manner of presentation, with

contextual/lexical information (when a target phrase is presented in the whole sentence) or without it (when a target phrase is presented solely). Since the aim of this experiment is to survey responses to a target phrase relying on acoustic features, it is important to avoid contextual/lexical information. And their ratings do not always reflect addressees' status.

3.3.3 ME ratings comparison between dialect groups

Correlation coefficients between the responses of the dialect groups were computed. The correlation coefficient for judging all stimuli when they were presented alone is significantly high (r=0.8438), and the coefficient for the PGT stimuli (r=0.9037) is higher than the coefficient for the ASM stimuli (r=0.8156), although the responses for the ASM stimuli still show a high correlation. From this result, we may say that subjects used prosodic features. But regarding the fact that subjects who showed poor correlation coefficients were excluded in this analysis, the remaining Kansai dialect speaker subjects successfully adapted themselves to the Tokyo dialect, and used prosodic features of Tokyo dialects.

These results explain two things. First, the subjects used acoustic features for formality judgement. Second, even when prosodic feature information was not available, in case of speaker ASM, they nevertheless managed to judge formality. There is a need for observation of estimated magnitudes and correlation of acoustical amounts.

3.4 Analysis -correlation of acoustics and formality-

In this section, we will look into the correlation of acoustics and formality. From reliability point, the Tokyo dialect groups responses are employed to see this correlation.

Table 3.4.1 shows the correlation of acoustics and formality. All parameters were computed in a logarithmic scale. F1 and F2 were extracted from /a/ of mora /ka/.

	Speech rate (mora/ms)	*SD of Mora duration.(%)*	*Stop in /ka/ (%)*	*Average F_0 (Hz)*	*F1(Hz)*	*F2(Hz)*
ASM	-0.501	0.316	0.466	-0.805	-0.322	0.490
PGT	0.219	-0.144	-0.379	-0.010	-0.553	-0.849

Table 3.4.1 A correlation of ME ratings and acoustic features (logarithm scale)

In case of speaker ASM, F2, stop and mora duration's SD shows some correlation with formality ratings, but F_0 shows negative correlation. In case of speaker PGT, speech rate shows some correlation with formality ratings, but F2 shows strong negative correlation with formality ratings. From this result, it is questionable to correlate speech rate with a degree of formality.

3.5 Discussion

The results from section 3.3.1 shows that the effect of dialect works on the formality responses for PGT stimuli obviously, though we still see high enough high reliability even from the responses for ASM stimuli. Therefore we may consider the prosodic features, such as duration of phonemes or pitch accent pattern affected from subjects' native dialect, is one of a main factor for judging formality. However, it does not explain subjects' responses in consistency for stimuli ASM, and we need further exploration of acoustic features. The result in section 3.4 suggests that suprasegmentals we observed is hardly to be regarded as main factor to judge formality. Actually, two subjects from the Tokyo dialect group mentioned that they judged formality, not from pitch or speech rate, but from lax/tense voice quality. In further study, I will try to reveal the effect of this voice quality.

Bibliography

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., and Weinert, R. (1991). "The HCRC Map Task Corpus", *Language and Speech*, **34**, 351-366

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)", in *Spoken language understanding and discourse processing* (Research Notes No. SIG-SLUD-9402, pp25-30). Japanese Society for Artificial Intelligence.

Bard, E.G., Robertson D., and Sorace, A., (1996) "Magnitude Estimation of Linguistic Acceptability", *Language*, **72**, 32-68

Brown, P., and Levinson, S., (1987). *Politeness: Some universals in language usage*. Cambridge: Cambridge university Press.

Kawanami, H., and Hirose, K., (1997) "Taido. Kanjo Onsei ni okeru In'ritsuteki-tokucho no ko-satsu", *SIG-HC97-***67**, 73-78

Ofuka, E., McKeown, J.D., Waterman, M.G., and Roach, P.J., (2000) "Prosodic cue for rated politeness in Japanese speech", *Speech Communication*, **32**, 199-217

Otake, T. and Cutler, A., (1999) "Perception of suprasegmental structure in a non-native dialect", *Journal of Phonetics*, **27**, 229-253

Stevens, S.S.,(1957) "On the psychological law", *Psychological Review*, **64**, 153-181

Stevens, S.S.,(1969) "On predicting exponents for cross-modality matches", *Perception and Psychophysics*, **6**, 251-256

Stevens, S.S.,(1975) *Psychophysics: Introduction to its perceptual, neural, and social prospects.* New York: John Wiley.