The Role of Prosody in the Perception of Formality in Japanese

Ethan Sherr-Ziarko

The University of Oxford
essz.research@gmail.com

ABSTRACT

This study investigates the role of prosody (specifically, mean f_0, f_0 range, and articulation rate) to Japanese listeners' perception of the level of formality of speech. This was accomplished via an experimental task using de-lexicalized speech created using recordings from an existing corpus where the level of formality of each recording had previously been judged[19]. Subjects judged the formality of recordings with both artificially manipulated prosody, and non-manipulated counterparts. Results showed listener responses changed significantly only when all three prosodic variables were manipulated together, indicating a 'super-additive' phenomenon. This was taken as evidence for a theory of category judgment tasks in speech perception based on an 'ideal listener model'.

Keywords: Prosody, formality, speech perception, Japanese, de-lexicalized speech.

1. INTRODUCTION

Recent work on the connection between prosody and the expression of formality (and/or politeness) across a number of languages [2],[8],[19],[21] has shown that there is a significant relationship between the two that is markedly similar in several languages (inc. Japanese, Korean, and Catalan Spanish). However, although some work has been done on the perceptual side of this relationship [2] the question of how various aspects of prosody function together in the expression of formality in speech remains open, as does the question of how salient each individual prosodic variable (e.g., f_0 versus articulation rate) is to the relationship.

This study addresses these questions as they relate to Japanese via an experiment using de-lexicalized speech (see e.g. [6],[13],[16]) which allows for the evaluation of the connection between changes in the prosody of a recording and listeners' perception of its level of formality. Based on previous work showing a connection between prosody and numerous aspects of speech perception [1],[4],[5],[20], and additionally on the production side of this relationship in Japanese [19] the hypothesis to be tested is that manipulating the prosodic variables of mean f_0, f_0 range, and articulation rate upwards will result in listeners judging recordings as more informal, while manipulating them downwards will cause subjects to judge the recordings as more formal.

2. EXPERIMENTAL DESIGN

In order to isolate the effects of prosody from the numerous lexical and grammatical indicators of formality in Japanese, the decision was made to use de-lexicalized speech for the experimental stimuli. De-lexicalized speech has previously been used in studies examining numerous aspects of prosody, and critically has been used successfully to investigate research questions related to f_0 [13] and articulation rate [6]. Because of this is was judged to be an appropriate tool for use in this study.

One previous study which was particularly informative to the experimental design of this project was Laan's [12] study of the connection between prosody and the judgment of the categories of read versus spontaneous speech. [12] examined the effects of prosody by comparing responses to manipulated (both individually and in combination) experimental stimuli to non-manipulated counterparts, and this study takes a similar approach.

2.1. Experimental presentation

The experiment was presented on a computer monitor using GNU Octave [7]. Subjects were presented with a randomly ordered set of de-lexicalized auditory stimuli (see Section 2.2 for specifics of their design and creation), which they were asked to judge to be either formal or informal. The experiment collected their responses on a forced-choice Likert-type scale [18]. The scale itself was presented on screen in the experiment as shown in (1), and input was via the numeric keypad.

(1)

1	**2**	**3**
Informal	Probably Informal	Maybe Informal
4	**5**	**6**
Maybe Formal	Probably Formal	Formal

This scale indirectly asks the subjects to rate their level of confidence in their response (maybe/probably/unqualified), which is done as an attempt to infer the magnitude of the effect of the manipulation of prosodic variables on subjects'

perceptions rather than simply forcing a binary choice between formal/informal.

Subjects were also consistently presented with the following information: they were told (accurately) that the speakers of the stimuli were native speakers of Tokyo-area Japanese, and were shown the age and gender of the speakers of the stimuli. This was done in order to give the subjects a baseline from which to judge changes in prosody, and to set their expectations towards the talkers' speech, as listener expectations have previously been found to effect speech perception in various ways [14].

2.2. Experimental stimuli

2.2.1. Generation of de-lexicalized recordings

The de-lexicalized speech stimuli were created based on speech taken from an existing corpus of spoken Japanese, where recordings had previously been judged as formal or informal based on a fixed set of criteria [19]. A semi-random selection of 100 recordings were used in this study. The stimuli pool was made up of an even split of formal and informal recordings, also evenly split by the gender of the original speaker. The recordings were of full sentences of spoken Japanese from between 2.5 – 7.0 seconds in length. Articulation rate, f_0 (taken at 15ms intervals), and intensity data were measured for each recording, and used to parameterize a Klatt synthesizer [9],[10].

To create de-lexicalized speech $F1$, $F2$, and $F3$ were set to constant values of 500, 1500, and 2500 respectively in order to create the impression of the entire recording being an extended /ə/, without perceptible consonants (see [9]: 986). Intensity was linearly scaled down to fall between 1 and 60 to meet the expectations of the synthesizer, and breaks in voicing (0 f_0 values) were replaced with a linear slope between the breaks in the voicing to improve the naturalness of the synthesized speech. The values for the prosodic variables of interest for the base experimental stimuli are shown in Table 1.

The values in Table 1 largely correspond to previous studies of the prosody of the formal/informal categories in Japanese [19], so the likelihood of the prosody of the stimuli confounding the results is low.

Table 1: Statistics for the prosodic variables of interest in the experimental stimuli. Range is measured as the difference between min and max f_0, and articulation rate is in moras/second.

	Informal		Formal	
Variable	Mean	SD	Mean	SD
Mean f_0	167.6 Hz	44.6 Hz	150.1 Hz	41.3 Hz
Rate	7.84 m/s	1.3 m/s	6.22 m/s	1.0 m/s
f_0 Range	135.1 Hz	50.8 Hz	90.4 Hz	39.1 Hz

2.2.2. Prosodic manipulation

The prosody of the base stimuli was manipulated to allow for the comparison of subjects' evaluations of manipulated versus non-manipulated stimuli. Mean f_0 was manipulated by simply adding or subtracting 20% of the mean from each value, resulting in a complete shift of the f_0 contour. The f_0 range of recordings was manipulated by first z-transforming the f_0 values to center the mean on 0, and then multiplying by either 0.8 to decrease the range, or 1.2 to increase. Finally, the articulation rate was manipulated by re-encoding the recordings (which were originally 16 kHz) at either 12 kHz to decrease the rate, or 20 kHz to increase it, while increasing or decreasing the f_0 values of the recording proportionally.

Each base stimuli was manipulated in several ways; firstly, versions were made where each individual prosodic variable was manipulated in the direction hypothesized to cause listeners to judge recordings as the opposite of what they originally were (i.e. originally formal recordings were manipulated upwards to make them seem more informal). Finally, versions where all three prosodic variables were manipulated together in *both* directions were created, resulting in a total of 6 versions of each stimuli.

3. DATA COLLECTION AND ANALYSIS

3.1. Overview of experimental subjects

In total, 16 linguistically naive native speakers of Japanese – 5 male and 11 female – took part in this experiment. here was no requirement that subjects be from a specific region of Japan, as it was assumed that subjects would have a high degree of familiarity with the 'standard' Tokyo Japanese regardless, but the subject pool was limited to people who were born and raised in Japan until at least age 18.

3.2. Experimental procedure

Subjects were given a brief overview of the experiment which instructed them to expect recordings to be from speakers of Tokyo-area Japanese, and to expect recordings with the actual words obscured. The concept of formality was defined for the subjects for the purposes of this experiment as speech "which appeared to be among friends or colleagues" for informal speech, and "speech towards elders or superiors" for formal speech. Subjects were introduced to the experiment itself via a brief practice section (two stimuli were presented). Subjects were given no specific instruction on the use of the six-point scale (as in 1) other than that their task was to judge each recording as being formal or informal. Finally, subjects were informed that the experiment would involve 300 stimulus objects (half of the total tokens, in an effort to minimize subject fatigue), and that they were allowed to take a break whenever they wished (the experiment was self-paced, and automatically paused after each selection). All instructions were given to the subjects in English, for consistency.

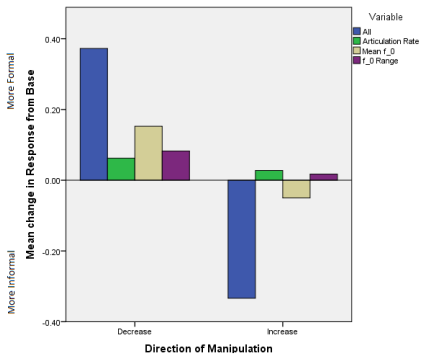
3.3. Data overview

With each subject judging half of the total pool of 600 stimuli, this resulted in a total of 4,800 responses. The overall mean of the responses falls at 3.38 on the 6-point scale, meaning that on average there was very close to a 50/50 split in how stimuli were judged (although it does skew very slightly towards subjects judging stimuli as informal). Analysis using cumulative link mixed models [3] in R [17] (as in 2) showed no significant relationship between the original formality of the base stimuli and subject responses, meaning they were not able to categorize them at a rate better than chance.

(2) **Model Structure**
Response ~ formality + (1+formality|subject)
Model Comparison Results
$X^2(1) = 2.41$, $Pr(>X^2) = 0.12$

Although subjects had difficulty accurately categorizing the base stimuli, a more striking pattern emerges when examining the *changes* in response from the base to the manipulated stimuli. As shown in Figure 1, the manipulation of individual prosodic stimuli does not appear to cause a particularly notable change in subject response, nor are there large differences between manipulating the variables upwards or downwards. However, there is a very obvious difference between the categories when all three variables are manipulated together (~ .7 steps on the Likert scale).

Figure 1: Relationship between the direction of manipulation of each prosodic variable and the change in subjects' responses.



3.4. Modelling analysis

As noted in Section 3.3, the responses (and changes in response) were on an ordinal scale, and therefore statistical analysis was conducted using cumulative link mixed models, a type of model specifically designed for use in analyzing ordinal data. All of the manipulation conditions were tested using the models in (3). In (3), 'Direction' refers to the direction of manipulation of the prosodic variables, while 'subtype' refers to which variables were manipulated. The base formality was included as an interaction term to check if it was confounding the relationship between direction and change in response. P-values were calculated using model comparison of the full model to a 'null' model with the fixed effect of interest removed. The results of all models are shown in Table 2.

(3) *All Data Model*
Model Structure
Response Change ~ Direction * Base Formality + (1+Direction|subject+stimulus+subtype)
Model Comparison Results
Direction: $X^2(2) = 9.46$, $Pr(>X^2) < .01$
Base Formality: $X^2(2) = 5.557$, $Pr(>X^2) = 0.06$

Subtype models
Model Structure
Response Change ~ Direction * Base Formality + (1+Direction|subject+stimulus)
Model Comparison Results
(All) Direction: $X^2(2) = 19.60$, $Pr(>X^2) < .001$
(f_0) Direction: $X^2(2) = 2.20$, $Pr(>X^2) = 0.13$
(Rate) Direction: $X^2(2) = 0.23$, $Pr(>X^2) = 0.63$
(Range) Direction: $X^2(2) = 0.17$, $Pr(>X^2) = 0.67$
Base Formality: $X^2(2) = 4.98$, $Pr(>X^2) = 0.08$

Results show that the overall change in response co-varies significantly with the direction of manipulation, but that the relationship is not significant when manipulating any individual variable. No in-

teraction terms were significant, and base formality was not significant. When all three variables are manipulated together, the differences in change in subject response based on direction of prosodic manipulation is significant.

Table 2: Results of modelling analyses.

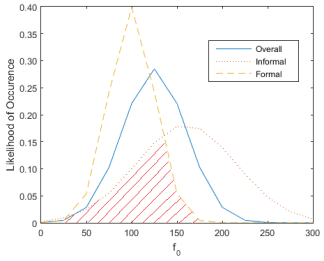
Variable	X^2 (2)	Estimate	p-value
Full Data	9.46	.797 ± .19	<.01*
All	19.60	1.169 ± .21	<.001*
Rate	.23	.101 ± .21	.63
Mean f_0	2.20	.259 ± .17	.13
f_0 range	.17	.072 ± .17	.67

4. DISCUSSION

Overall, the results of this study lend qualified support to the hypothesis described in Section 1. Although there is a relationship between the direction of manipulation of the prosodic variables and change in listener response, the relationship is only significant when all three variables are manipulated together. This result is similar to those seen previously in [12] in that the results appear to be *super-additive*, meaning that the total effect of manipulating multiple variables is greater than it would be if you simply added the effects of manipulating each variable individually.

This discrepancy could be at least partially explained by theories of speech perception espoused in e.g. [11] & [15], where perceptual categories (e.g. vowels, syllables, social categories) are realized as distributions – often Gaussian in shape – of phonetic values and category judgments are made probabilistically (also known as an 'ideal listener model'). To illustrate this concept as it applies to the current study, Figure 2 shows a series of hypothetical normal distributions of f_0 values from an imaginary speaker.

Figure 2: Hypothetical distributions of a speaker's f_0 in overall, formal, and informal categories.



While the distributions in Figure 2 overlap, the formal category is much narrower (lower SD) and therefore it is hypothesized that when a data point falls into a point of overlap in the distribution, it is more likely to be judged as part of the formal category.

If this theory is accurate, then it would make sense that manipulation of a single variable might not be enough to – for example – push a subject to judge a recording to likely be informal if they are considering a series of different distributions of phonetic variables and the other two variables point to the recording being formal. This point was possibly exacerbated by the design of the experiment, wherein individual variables were only manipulated to make recordings appear closer to the opposite of their original category. While this study does not provide enough evidence to claim that this theory is definitely accurate, future similar experiments which more closely control the prosody of the base stimuli and test more directions of manipulation could provide more robust support for the theory.

5. CONCLUSION

While previous work on the connection between prosody and formality in speech has posed the question of which aspects of prosody are most relevant to the relationship [8], this study has shown that the salience of individual variables may be less important than how they work together to inform a listeners' perception of different linguistic categories.

The results of this study indicate that the process of judging the formality of speech based on its prosody is more probabilistic than deterministic, likely involving the consideration of multiple factors simultaneously. However, it seems likely based on the super-additive nature of the results that listeners are not using any one variable in isolation, and therefore any explanation focusing on distributions of a single variable may be somewhat reductive. While the results do lend a certain amount of support to a theory where speakers are accessing cue distributions similar to those seen in Figure 2, a model of this category judgment task should likely be based on a 'combined' distribution of all the relevant phonetic parameters considered together.

In sum, this study shows that prosody is closely related to the perception of formality in Japanese, mirroring results both on studies of other languages [2] and other paralinguistic categories [12]. Methodologically speaking, the results indicate that phonetic research investigating category judgment tasks in speech should take the broader prosodic environment into account, rather than only investigating one or two prosodic variables in isolation.

6. REFERENCES

[1] Abercrombie, D. 1967. *Elements of General Phonetics*. Edinburgh University Press.

[2] Brown, L., Winter, B., Idemaru, K., & Grawunder, S. 2014. Phonetics and politeness: Perceiving Korean honorific and non-honorific speech through phonetic cues. *Journal of Pragmatics*,66, 45-60.

[3] Christensen, R. H. B. 2015. ordinal - Regression Models for Ordinal Data. R package version 2015.6-28. http://www.cran.r-project.org/package=ordinal/.

[4] Collier, R., & 't Hart, J. 1975. The role of intonation in speech perception, in: A. Cohen and S.G. Nooteboom (Eds.) *Structure and Process in Speech Perception*, 107-123. Springer Verlag: Heidelberg.

[5] Darwin, C. J. 1975. On the dynamic use of prosody in speech perception. *Status Report on Speech Research*, 42/43. Haskins Laboratories.

[6] Dellwo, V. 2008. The role of speech rate in perceiving speech rhythm. *Speech Prosody 2008*, 4(8), 375-378.

[7] Eaton, J., Bateman, D., Hauberg, S., & Wehbring, R. 2015. GNU Octave version 4.0.0 manual: a high-level interactive language for numerical computations.

[8] Hübscher, I., Borràs-Comes, J., & Prieto, P. 2017. Prosodic mitigation characterizes Catalan formal speech: The Frequency Code reassessed. *Journal of Phonetics*, 65, 145-159.

[9] Iles, J., & Ing-Simmons, N. 1994. Klatt: A Klatt-style speech synthesizer implemented in C (Version 3.0.4). *CMU Artificial Intelligence Repository.*

[10] Klatt, D. 1980. Software for a cascade/parallel formant synthesizer. *Journal of the Acoustical Society of America*, 67(3), 971-995.

[11] Kleinschmidt, D. F., & Jaeger, T. F. 2015. Robust speech perception: recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological review*, 122(2): 148–203.

[12]Laan, P. 1997. The contribution of intonation, segmental durations, and spectral features to the perception of spontaneous and read speaking style. *Speech Communication*, 27, 43-65.

[13] Morley, E., Klabbers, E., van Santen, J. P., Kain, A., & Mohammadi, S. H. 2012. Synthetic F0 Can Effectively Convey Speaker ID in Delexicalized Speech. *INTERSPEECH 2012*, 434-437.

[14] Niedzielski, N. 1999. The effect of social information on the perception of sociolinguistic variables. *Journal of language and social psychology*, 18(1), 62-85.

[15] Norris, D., & McQueen, J. M. 2008. Shortlist B: A Bayesian model of continuous speech recognition. *Psychological review*, 115(2), 357–395.

[16] Pagel, V., Carbonell, N., & Laprie, Y. 1996. A new method for speech delexicalization, and its application to the perception of French prosody. In *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on Vol. 2*, 821-824.

[17] R Core Team. 2017. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.URL: *https://www.R-project.org/*.

[18] Schütze, C. T., & Sprouse, J. 2014. Judgment data. in Podesva, Robert J., and Devyani Sharma, eds. *Research methods in linguistics*, 27-50. Cambridge University Press.

[19] Sherr-Ziarko, E. 2018. Prosodic properties of formality in conversational Japanese. *Journal of the International Phonetic Association*, 1-22.

[20] Studdert-Kennedy, M. 1979. Speech Perception. Status Report on Speech Research 59/60. Haskins Laboratories.

[21] Winter, B., & Grawunder, S. 2012. The phonetic profile of Korean formal and informal speech registers. *Journal of Phonetics*, 40, 808-815.