INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Acoustic Properties of Formality in Conversational Japanese

Ethan Sherr-Ziarko [1]

[1] University of Oxford, United Kingdom
ethan.sherr-ziarko@stx.ox.ac.uk

Abstract

This paper examines potential acoustic cues for level of formality in Japanese conversational speech using speech data gathered outside the laboratory, with the objective of using any significant cues to develop a model to predict level of formality in spoken Japanese. Based on previous work on the phonetic properties of formality in Japanese [1],[2] and other languages [3], and on a pilot study of informal geminate contractions in Japanese (section 2), the study examined the mean f_0, articulation rate, and f_0 range (the difference between the minimum and maximum f_0 in an utterance) via direct examination of the data and a functional data analysis [4],[5]. Analysis of the speech data shows significant relationships between all three variables and level of formality, and a binary logistic regression indicates that the variables have some potential as predictors of formality independent of lexical cues, although further refinement of any model will be necessary.

Index Terms: Formality, Acoustic Phonetics, Functional Data Analysis, Japanese, Speech Modeling

1. Introduction

This paper describes a study investigating the acoustic properties of the informal register of conversational speech in Japanese, particularly how it compares to the formal register. Japanese was chosen as the language of study because the large number of lexical and grammatical features indexical of speech register [6],[7] and the existence of honorifics make the process of judging the level of formality in speech less subjective than in many other languages.

There are few previous studies of the acoustic properties of different levels of formality in spoken Japanese, and those which have been conducted have generally examined speech elicited in a laboratory setting either via direct prompting [1] or an assigned task [2]. Although these produced some significant results – such as increased f_0 and articulation rate in elicited informal speech – which provided some initial clues for fruitful avenues of investigation to pursue, they are not directly applicable to the current study due to the nature of the speech analyzed.

Instead, the decision was made to create a new small corpus of conversational Japanese for analysis (see Section 3 for further details on collection methodology), with the objective of analyzing more natural speech patterns. However, in order to form an initial hypothesis for the study and to help determine which variables to investigate, a more controlled, lab-based pilot study was conducted first.

2. Pilot study

2.1 Methodology

The objective of the pilot study was to determine possible acoustic properties of informal Japanese speech by analyzing the correlates of one specific target – geminate contractions. This refers to instances in speech where a mora which is a singleton in its canonical lexical form is realized phonetically as a geminate, as in the example in (1).

/wakaɻanai/ *("don't understand")* → [wakan:ai] (1)

This frequently occurs in an informal register of conversational Japanese, and having subjects produce carrier sentences containing these geminate contractions was judged to be a good way to elicit informal speech.

Subjects were brought into the lab and prompted to produce written carrier sentences appearing on a screen which contained singleton/geminate contrasts as in (1). Subjects were five (2 female, 3 male) native speakers of Japanese aged 23-31 raised in the Tokyo region at least until age 18. Subjects also produced distractor sentences not containing geminate contractions.

2.2 Results and analysis

Subjects' speech was both recorded and analyzed using Praat. Based on previous studies on the acoustic correlates of gemination in Japanese [8] and informal speech more generally [1],[2] the target variables examined were mean f_0, speech rate, and amplitude (dB). As the objective was to determine possible general acoustic properties of informal speech, entire utterances were analyzed rather than only the target geminate segments. The initial hypothesis of the study, based on the previous work, was that all three variables would be higher in utterances containing geminate contractions as compared to utterances containing their singleton counterparts.

Analysis of the target variables showed that in utterances containing geminate contractions, mean f_0 was 15.61 Hz higher ($p<.02$) than in utterances containing singletons. Additionally, the mean durations of otherwise identical carrier sentences containing geminate contractions (not including the target segments themselves) were 32ms shorter than sentences containing the singleton counterparts, a significant difference ($p<.05$). Amplitude did not show any significant correlation with the singleton/geminate contrast.

Based on these results f_0 and articulation rate were chosen as target variables for the main study from the pilot. In addition, based on significant results from other studies [3], f_0 range was also included as a target.

3. Data collection and annotation

In order to obtain appropriate conversational Japanese speech data to analyze, the decision was made to create a new small corpus. Although other corpora of spoken Japanese exist, notably the Corpus of Spontaneous Japanese [9] and the Chiba 3-way conversation corpus [10], they were judged to be

http://dx.doi.org/10.21437/Interspeech.2016-132

insufficient for the current study due either to a lack of conversational speech [9] or an insufficient amount of data [10].

3.1 Data collection methodology

The speech data for this study was collected at the NINJAL institute in Tachikawa-shi, Japan via one-on-one interviews between the experimenter and a subject. The interviewer was a non-native speaker of Japanese with a high-level of proficiency, and subjects were 10 native speakers of Japanese aged 31-45 (5 male, 5 female). The age of subjects was kept below 50 in order to minimize any potential influence of the effects of age on f_0 [11], and all subjects were speakers of the Tokyo dialect of Japanese (born and raised in the Tokyo area up to age 18) in order to reduce any possible effects on f_0 from different dialects [12]. Interviews were conducted in a lounge setting rather than a recording booth or lab in order to encourage a more natural, conversational style of speech. Recordings were single-channel mono, made at 48 kHz.

The format of the interviews was similar to the sociolinguistic interview [13] but with less control over the topics discussed, and each subject was recorded for ~30 minutes. All interviews began with self-introductions from both the interviewer and the subject, which were generally quite formal, and then proceeded naturally to other topics as they arose, with the interviewer gradually modulating their speech register to a more informal level to encourage the subject to follow. In general this resulted in a pattern where the first five minutes of the interview consisted mainly of formal speech, minutes 5-10 consisted of a mix of formal and informal speech, with subjects sometimes code-switching within utterances, and the remainder of the interview consisting of mostly informal speech. At the end of each interview, subjects were asked to read a short passage to provide an example of read speech, to use as a control.

In total this resulted in ~5 hours of recorded speech for analysis.

3.2 Data annotation

All of the subjects' utterances were labeled in Praat text grids, with the following exceptions:

- Isolated filler interjections (such as /e:/ or /a:/) were not included.
- Isolated laughter was not included, but was included if it occurred clause-internally.
- Extended pauses (defined as pauses of >1 second) were not included.

Boundary labels were placed either at clause boundaries for full utterances, at the start/end of an extended pause (> 1s) for fragments, or at turn-taking boundaries in the case of back-and-forth conversation containing fragments.

The number of lexical moras within each clause or fragment was manually counted in order to allow the calculation of articulation rate data. Pauses of less than 1s were included and counted as 1 mora per 100ms consistently in order to reduce any effects of more or less frequent pausing in speech on articulation rate.

The speech within each label was judged to be either formal, informal, or read. Although any judgment of the level of formality of a given utterance will be to some degree in the eye of the beholder, because the determination of levels of formality is very important to this study a consistent set of criteria to judge formality was established, as seen in Table 1.

Table 1: *Criteria used in determining utterance formality.*

Criteria	Formal Example	Informal Example
Copular verb	/desɯ/ “to be”	/da/ “to be”
Verb form	/ʃimaʃita/ “did”	/ʃita/ “did”
Sentence-final particles	/-wa/	/-jo/
Question particles	/-ka/	/-kai/
Under/over articulation	/tsɯmaɻanai/ "boring"	/tsɯman:ai/ "boring"
Indexical word forms	/jahaɻi/ “...after all"	/jap:a/ "...after all"
Honorifics	/ikaɻemasɯ/ "to go (HON)"	/ikɯ/ "to go"

The criteria in Table 1 were determined largely by previous examinations of lexical items and phonological forms indexical of different registers of formality in Japanese [5],[6], [14] and of observational evidence of spoken Japanese. Although these criteria were applied consistently, there were a small percentage of utterances (roughly 1/20) which were ambiguous either due to a lack of criteria present in the utterance, or code switching. In such cases, a linguistically naive native speaker of Japanese was consulted, and their opinion followed.

Once all the speech data was labeled, it was then automatically segmented into separate .wav files, and f_0 and articulation rate data was extracted using Praat and bash scripts. In total, this resulted in **2,697** utterances, of which **416** were formal, **2,068** were informal, and **214** were read. In total there were **1,314** utterances by female subjects and **1,383** by male subjects.

4. Data analysis

The variables analyzed in this study were mean f_0, articulation rate, and f_0 range (the difference between the min and max f_0 in a given utterance). The initial hypothesis of the study, based on the pilot study, and on a similar study of the acoustic properties of formality in Korean [3] was that each variable would be significantly higher in informal speech than in formal speech.

4.1 Articulation rate

Table 2 shows articulation rate statistics for each level of formality.

Table 2: *Articulation rate statistics.*

Formality	Mean	Std. Deviation
Informal	7.82 moras/second	1.51 m/s (19% of mean)
Formal	6.64 moras/second	1.63 m/s (24% of mean)
Read	6.93 moras/second	1.16 m/s (17% of mean)

It is immediately apparent from Table 3 that the mean articulation rate of informal speech is quite a bit higher than that of either formal or read. A paired sample t-test shows

the mean difference of 1.18 moras/second between informal and formal speech to be significant at the p<.001 level, as well as a similarly significant difference between informal and read speech (p<.001). There is no significant difference between formal and read speech.

This initial finding agrees with the results of the pilot study, and of previous acoustic studies of Japanese informal speech [1],[2], but does not tell the whole story of the relationship between articulation rate and formality. Figure 1 compares histograms of articulation rate in formal and informal speech.

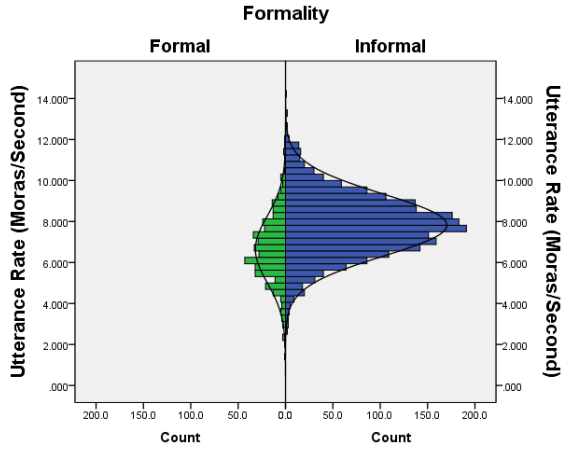

Figure 1: *Histogram of articulation rate in informal and formal speech.*

It is apparent from the overall distributions in Figure 1 that although both informal and formal speech have a similar minimum for articulation rate (~2 moras/second), informal speech appears to have a much higher maximum articulation rate. This indicates that although speakers did not always articulate faster in informal speech (although they did so typically), there was a greater possible **range** of articulation rates in informal speech.

These differences in articulation rate hold regardless of other factors, with a univariate ANOVA showing no significant interaction between articulation rate and speaker ID, age, or gender.

4.2 Mean f_0

Mean f_0 is also different in formal and informal speech. Table 3 shows the f_0 data for each level of formality.

Table 3: *f0 statistics.*

Formality	Mean f_0	Std. Deviation
Informal	178.7 Hz	56.3 Hz (31% of mean)
Formal	158.6 Hz	48.4 Hz (30% of mean)
Read	159.7 Hz	41.6 Hz (26% of mean)

Once again, informal speech shows a significant difference from formal speech, with a paired sample t-test showing the mean difference of 20.1 Hz as significant (p<.02). On the linear Hz scale female speakers have a more pronounced difference in f_0 between informal and formal speech (24 Hz difference for females vs. a 14 Hz difference for males), but when the values are $\log_{10}$ transformed in order to make them more in line with human pitch perception, the differences between the genders disappear, and all subjects appear to follow the general pattern.

Additionally, mean f_0 and articulation rate do not appear to be confounded, with both a univariate ANOVA and Pearson correlations showing no significant relationship, meaning that the mean increase in f_0 does not appear to be caused by the overall mean increase in articulation rate in informal speech.

4.3 f_0 range

The final variable tested was the difference between the max and min f_0 values of each utterance. Although this is a somewhat crude measure, it can still be of use for determining general patterns before more in-depth analysis is conducted. Table 4 shows the f_0 range statistics for each level of formality.

Table 4*: f_0 range statistics.*

Formality	Mean f_0 range	Std. Deviation
Informal	228.72 Hz	130.62 Hz (57% of mean)
Formal	178.44 Hz	137.51 Hz (77% of mean)
Read	172.69 Hz	114.53 Hz (66% of mean)

There is again a readily apparent difference between informal and formal speech, with a paired samples t-test showing the difference in means of 50.28 Hz to be significant (p<.001). However, the very large standard deviations indicate that there are some problems with this method of analysis. This high level of variability likely results from pitch doubling and halving errors in the pitch-tracking, as can be seen in Figure 2.

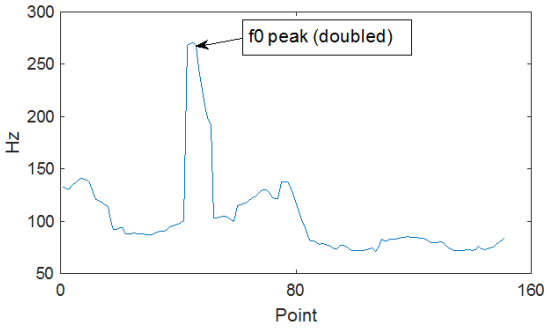

Figure 2: *An f_0 vector showing a pitch-doubling error.*

Such errors present a large problem for this method of analysis, as although it is possible that the errors are spread proportionally among the different levels of formality, it is not possible to determine this without examining every single utterance. It is therefore difficult to know to what degree pitch tracking errors impact the analysis, and because not all utterances contain such errors simply halving pitch peaks does not appear to be a viable solution. To help overcome this problem, and to provide a more in-depth analysis of f_0 range, a functional data analysis method was adopted.

5. Functional data analysis

Functional data analysis refers to a methodology whereby continuous functions (in this case orthogonal polynomials) are fitted to linear f_0 vectors, and the orthogonalized coefficients of the fitted polynomials are related to discrete linguistic variables [4],[5]. This was done by using the *polyfit* function in

Matlab to fit a cubic polynomial to each utterance and then examining its coefficients.

First, full pitch tracking data was taken for each utterance (at intervals of 10ms), and the vectors were normalized using the operation in (2), and then normalized for time using the operation in (3) (where y is the original f_0 vector, yn is the normalized f_0 vector, and x is the normalized time-axis).

$$yn = y/mean(y)-1 \quad (2)$$

$$x = (1:length(yn)-length(yn)/2)/length(yn)/2 \quad (3)$$

In order to further determine the goodness of fit, the sum of the squared differences between the fitted function and the normalized data vector was calculated using the operation in (4) (where yf is the fitted function).

$$d = sum((yf-yn)^2)/length(yn); \quad (4)$$

Examination of the f_0 contour made it apparent that a d of around .02 was necessary for the function to fit accurately. Figure 3 shows an example of such a function.

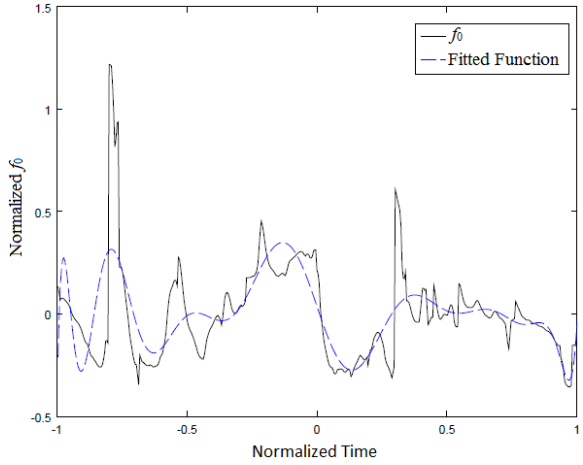

Figure 3: *fitted function with a d of ~.02.*

The function in Figure 3 fits quite closely to the vector, and also appears to achieve the goal of smoothing out a pitch-doubling error early in the vector. Overall, in order to achieve a mean d of .02 for all utterances is was necessary to map a 30 degree polynomial to each utterance. However, with such a large number of coefficients it is difficult to relate each one to a linguistic variable, so an alternative method was adopted.

Each long fitted function was broken down by taking the sections between f_0 troughs – defined as a point where the f_0 contour changes from decreasing to increasing (including from the start and to the end of the vector) – and then fitting a cubic function to each of those sections. The four orthogonalized coefficients of those functions can be interpreted as follows [4]:

1. Coefficient 1 corresponds to the s-shaped 'wiggle' of the function.
2. Coefficient 2 corresponds to the breadth of curvature of the function (how sharply the f0 rises towards and falls from the peak).
3. Coefficient 3 corresponds to the slope of the function, or how the f0 rises or falls overall.
4. Coefficient 4 corresponds to the average height of the function, (i.e. the average f0)

In total this resulted in 28,841 sets of coefficients. In order to avoid skewing the data with poorly fitted functions, any function with a d greater than .05 was excised from the data set, resulting in **27,118** total coefficient sets to be analyzed. A comparison of the average functions (obtained by taking the mean of each of the four coefficients for informal and formal speech) can be seen in Figure 4, and a list of the mean orthogonalized coefficients can be seen in Table 5.

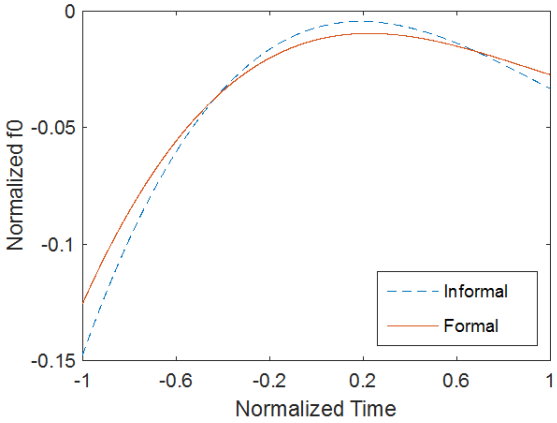

Figure 4*: fitted functions for informal and formal speech.*

Table 5: *List of mean orthogonalized coefficients*

Formality	Coeff. 1	Coeff. 2*	Coeff. 3*	Coeff. 4*
Informal	.0285	-.0834	.0287	-.0072
Formal	.0237	-.0641	.0252	-.0122

* These coefficients are significant in a binary logistic regression ($p<.01$).

Figure 4 reveals a few observations of note. Although the functions are similar due to them both being cubics, there are some visually apparent differences. Although coefficient 1 is very close to 0 for both functions (there is very little s-shaped 'wiggle', as might be expected due to the segments of the function that are being analyzed), there are significant differences in all the others. Informal speech appears to start lower and peak higher (based on coefficient 2), and to also curve up and down more quickly (meaning a faster increase and decrease in f_0, based on coefficient 3). Informal Speech is also slightly higher overall (coefficient 4) indicating a higher mean f_0. A binary logistic regression comparing the coefficients seen in Table 5 further shows coefficients 2-4 to be significant predictors of formality, all at the $p<.01$ level.

6. Conclusion

The findings in this study indicate that there are a number of significant acoustic properties of informal conversational speech in Japanese. All of the tested variables (articulation rate, mean f_0, and f_0 range) appear to be used by speakers in production (whether consciously or not) to indicate level of formality. Informal speech appears to allow a greater level of flexibility in the prosody of the utterance, as is demonstrated by the increased range of possible articulation rates and f_0 values in the informal speech that was examined.

7. References

[1] Ofuka, E., J. McKeown, M. Waterman, P. Roach. (2000). Prosodic cues for rated politeness in Japanese speech. *Speech Communication,* 32, 199-217.

[2] Ito, M. (2002). Japanese politeness and suprasegmentals – a study based on Natural Speech Materials, *Speech Prosody 2002.*

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

[4] Grabe, E., G. Kochanski,, & J. Coleman, (2007). Connecting intonation labels to mathematical descriptions of fundamental frequency. *Language and Speech* 50(3), 281-310.

[5] Ramsay, J. O. (2006). *Functional data analysis*. John Wiley & Sons, Inc.

[6] Cook, H. M. (1998). Situational meanings of Japanese social deixis: The mixed use of the masu and plain forms. *Journal of Linguistic Anthropology* 8(1), 87-110.

[7] Sreetharan, C. (2004). Students, *sarariiman* (pl.), and seniors: Japanese men's use of the 'manly' speech register. *Language in Society* 33, 81-107.

[8] Guion, S. & K. Idemaru. (2008). Acoustic covariants of length contrast in Japanese stops. *Journal of the IPA* 38, 167-286.

[9] Maekawa, K. (2003). Corpus of Spontaneous Japanese: its design and evaluation. *ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, 2003.

[10] Den, Y. (2014). Chiba 3-way conversation corpus. Chiba University.

[11] Harrington, J., S. Palethorpe, & C.J. Watson. (2007). Age-related changes in fundamental frequency and formants: a longitudal study of four speakers. *Interspeech 2007*, 2753-2756.

[12] Kubozono, H. (2012). Varieties of pitch accent systems in Japanese. *Lingua,* 122(13), 1395-1414.

[13] Labov, W. (1972). *Sociolinguistic patterns*. University of Pennsylvania Press.

[14] Okamoto, S. (1999). Situated politeness: manipulating honorific and non-honorific expressions in Japanese conversations. *Pragmatics*, 9(1), 51-74.