Prediction of F0 height of filled pauses in spontaneous Japanese: a preliminary report

Kikuo Maekawa

National Institute for Japanese Language and Linguistics, Japan

Abstract

F0 values of filled pauses (FP) in the Corpus of Spontaneous Japanese were analyzed to examine the mechanism by which the F0 heights of FP were determined. Statistical analyses of the F0 values of FP occurring in between two full-fledged accentual phrases (AP) revealed correspondence between the occurrence timing of FP and the F0 height. Based upon this finding, 5 models of F0 prediction were proposed. Comparison of the mean prediction errors revealed that the best prediction was obtained in a model that linearly interpolate the phrase-final L% tone of the immediately preceding AP and the phrase-initial %L tone of the immediately following AP. This finding suggests that the F0 of FP was specified at the level of phonetic realization rather than phonological prosodic representation.

1. Introduction

Frequent occurrence of filled pauses (FP hereafter) is one of the most salient characteristics of spontaneous speech. There's a wide consensus among the researchers that FP play positive roles in the processing of spontaneous speech. The supposed cognitive roles of FP include prognosis of the perplexity of upcoming word [1], or the complexity of the upcoming clause [2], marking of discourse structure [3], discourse management [4], indication of the degree of factuality of university lectures [5], etc. There are also speech analytic studies on the phonetic characteristics of FP ([6] among others), and, applications-oriented studies including synthesis of dialogue speech [7], recognition of spontaneous speech [8], etc.

Despite its cognitive importance, mechanisms of FP production are left mostly untouched in the study of speech production. In the study of speech prosody, for example, existing theories of prosodic structure do not pay any attention for the intonational or other prosodic characteristics of FP [9]. The lack of scientific knowledge in this field poses, accordingly, serious limitations on the design of prosodic annotation schema for spontaneous speech.

In the X-JToBI annotation scheme, which was proposed for the prosodic annotation of spontaneous speech [10], FP are treated as a special kind of accentual phrase (AP hereafter) whose pitch height is specified tonally either as FH ('filler-high') or FL ('filler-low'). This binary labeling, however, was not proposed on a firm theoretical basis. It is rather a simple extrapolation of established knowledge about the prosody of Japanese that L and H are required for the specification of linguistic contrast and pragmatic information. There is no a priori reason to believe that FP are specified with respect to binary, or whatever, tonal opposition.

In the rest of this paper, corpus-based analyses of FP will be conducted in terms of their location in utterance, timing with respect to adjacent AP, and, F0 height, to know if it is possible to predict the F0 height of FP from their occurrence environment.

2. The data

The 'Core' part of the Corpus of Spontaneous Japanese (CSJ hereafter), which is X-JToBI annotated, was used for analyses [11]. 44 hours of speeches containing about half a million words are included in the CSJ-Core. FP in the CSJ-Core are marked not only in the X-JToBI annotation, they are also marked in the speech transcriptions. Since the criteria of FP recognition are not identical in the prosodic annotation and speech transcription, the total number of FP do not coincide in the prosodic annotation and transcription. The main difference stems from the treatment of a FP (/de/ see Table 1) occurring in the beginning of utterance, which is treated as a FP in prosodic annotation, while it is treated as an ordinary conjunctive in speech transcription. In the present study, FP were recognized according to the criteria of the X-JToBI scheme. The total number of FP analyzed in this study was 35,164.

As for the textual property, 160 different textual shapes were recognized in the speech transcription of the FP in CSJ-Core. Since this classification is too detailed for the present analyses, FP were reclassified into 23 classes based upon the similarity of their segmental shapes. These classes were further reclassified into 8 classes. The results of two-way classifications are shown in Table 1 as Class1 and Class2 respectively. Note that FP whose occurrence frequencies were less than 10 were omitted from the classifications.

Table 1: *Textual classification of FP.*

Class1	Class2	Example
A	A	/a/ /aQ/
AH		/aH/
AN		/ano/ /anoH/ /aHno/ /aHnoH/ etc.
DE	D	/de/ /te/ /Nde/
E	E	/e/
EH		/ee/ /eH/
ET		/eHto/ /eHtoH/ /eHQto/ /eQto/ etc.
KN	K	/kono/
KO		/kou/
M	M	/ma/
MH		/maH/
MO		/moH/
N	N	/N/
NH		/N:/
NT		/N:to/ /Nto/ /N:toH/ /N:Qto/ etc.
UN		/uHN/ /uN/ etc.
SN	S	/sono/ /sonoH/ etc.
U	V	/u/
UH		/uH/
I		/i/
IH		/iH/
O		/o/
OH		/oH/

Not all FP in the CSJ-Core are suitable for the present study, because it is often impossible to measure the F0 value of FP. FP that meet the following 3 criteria were chosen: (a) The F0 value of the FP in question is reliably measurable, (b) The F0 values of the phrase-final L% tone of the AP that immediately precedes the FP in question is reliably measurable, and, (c) The F0 value of the phrase-initial %L tone of the AP that immediately follows the FP in question is reliably measurable. F0 values and their reliability information could be extracted from the XML files containing the X-JToBI annotation data.

In the X-JToBI data of CSJ-Core, the F0 value of a FP was represented by the F0 value measured near the center of the FP duration. This measurement criterion was adopted on the assumption that the F0 pattern of FP in Japanese is nearly flat unlike the ordinary AP in Japanese (see sections 4.1 and 6).

As the result, 4,892 FP were chosen for analysis. Note that the cases where more than two FP occurred consecutively were excluded from the data. The F0 values of the FP and AP were log-transformed and z-transformed for each speaker.

3. Analysis

3.1. Location of occurrence in clause

Occurrence location of FP in a clause is examined in the first place. The frequencies of Class1 FP were computed for each word positions in a clause, beginning from the first up to, whenever possible, 15th position. The results revealed crucial difference between the Class1 DE and all other FP. More than 80 % of DE occurred as the first 'word' of clauses, while the distributions of other FP have their peaks in the clause-medial positions.

3.2. Timing of occurrence

Timing of FP occurrence was analyzed with respect to the adjacent AP. The relative timing of the beginning of a FP is called relPosit and defined as (T3-T1)/(T2-T1), where T1 is the ending time of the AP that immediately precedes the FP in question, T2 is the beginning time of the AP that immediately follows the FP in question, and, T3 is the beginning time of the FP in question (Cf. Figure 4 below). This index distributes in the interval $0.0 \leq \text{relPosit} < 1.0$. When relPosit=0.0, there is pause between the preceding AP and the FP. When relPosit takes a positive value, there is a pause, and the larger the relPosit, the closer the beginnings of FP is to the following AP. Note that relposit can't be equal to 1.0 because a FP has its own duration.

Figure 1 is a boxplot showing the distributions of relPosit computed for each of the Class1 FP having the frequency higher than 10. Here again, DE is unique in that its distribution is concentered near the higher end of ordinate (relPosit). Similar distribution is found in the case of N. On the other hand, the distributions of AH and AN are concentrated near the origin of ordinate. And, the distributions of FP like A, M, MO, O, and U are widely dispersed across the whole ordinate.

3.3. F0 height

3.3.1. Relation between the timing and F0

There is a loose correlation between the mean F0 heights and the mean relPosit values as shown in Figure 2. The abscissa and ordinate of the figure are the relPosit value as divided into 10 classes and the z-transformed value of log-transformed F0 value (F0Logn, hereafter). F0Logn value stays nearly constant within the lower range of abscissa, but increases considerably toward the higher end. Figure 3 shows the result of the same analysis as applied individually to some Class2 FP that have high frequency of occurrence (N>49). All classes other than A and D show the same tendency as in Figure 2, and in the case of class A, F0Logn increases in the area where relPosit > 0.9.

3.3.2. Prediction models of F0 height

Prediction of the F0Logn values was conducted based upon the findings in the previous subsection. Results of 5 different prediction models were compared (See Figure 4). In Model 1, the F0 value of the AP-final L% tone in the preceding AP is copied as the F0 value of the FP. In Model 2, on the contrary, the F0 value of the AP-initial %L tone of the following AP is copied as the F0 of FP. The relPosit values play no role in these models. Model 3 is a hybrid of Models 1 and 2.

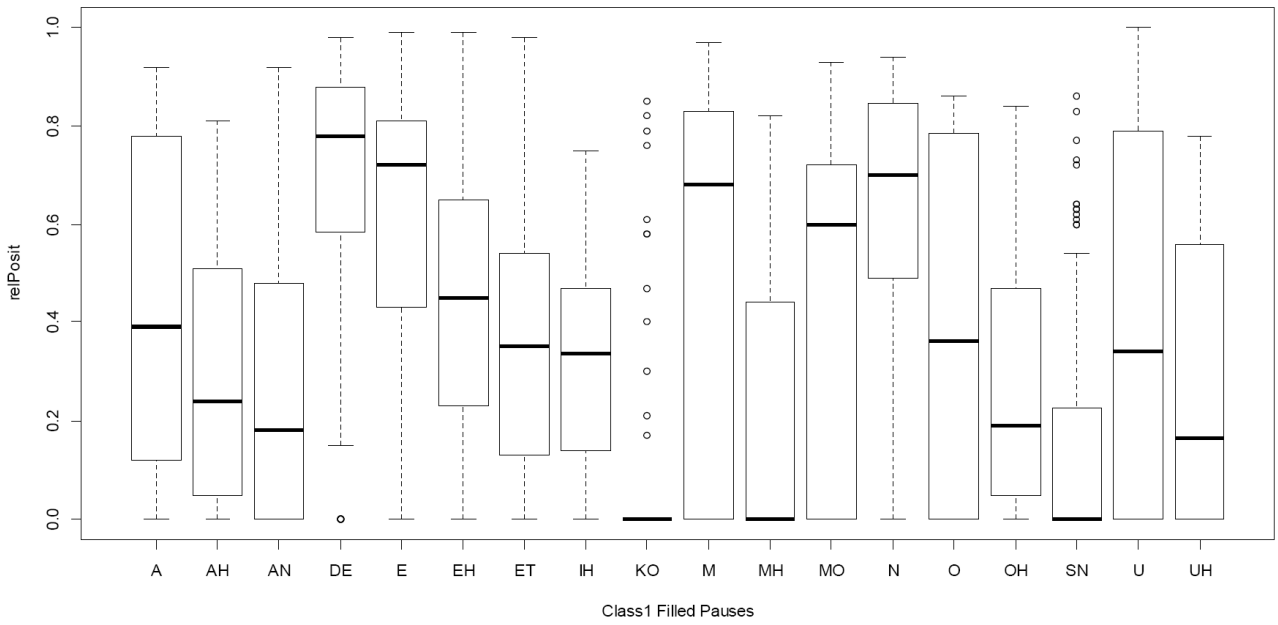

Figure 1: *Distributions of the timing of occurrence (relPosit values) of Class1 FP (N>8).*

In this model, the value preceding L% is copied if the relPost < 0.7; the following %L is copied otherwise. In Model 4, the F0 value of the FP is determined by the interpolation between the values of preceding L% and following %L. Lastly, Model 5 is a hybrid of Models 1 and 4; Model 1 is applied for FP whose relPosit values are smaller than 0.7, and the F0 interpolation of L% and %L is applied for all other FP. Note that in this model, L% is regarded to be copied to the location where relPosit = .7, and the F0 interpolation is conducted between the copied L% and the following %L as shown in Figure 4.

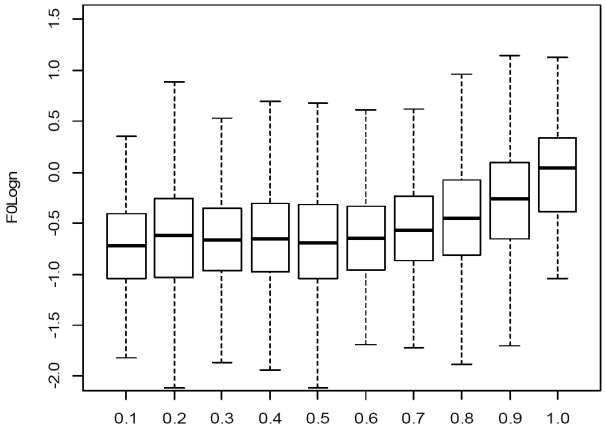

Figure 2: *Relation between the timing and F0 value.*

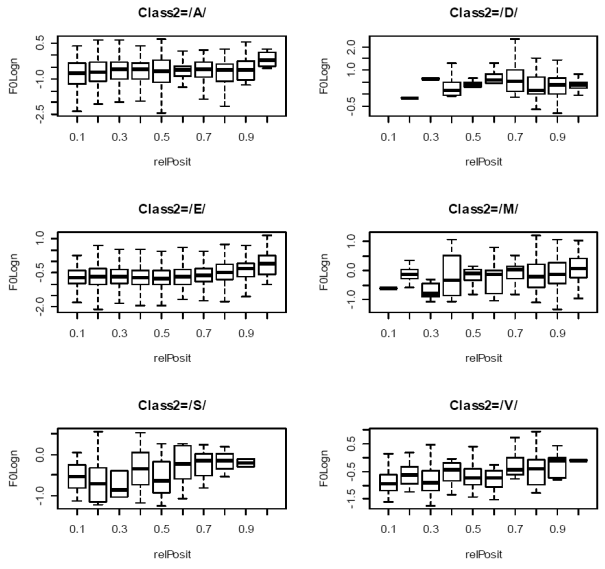

Figure 3: *Timing-F0 relation in individual FP.*

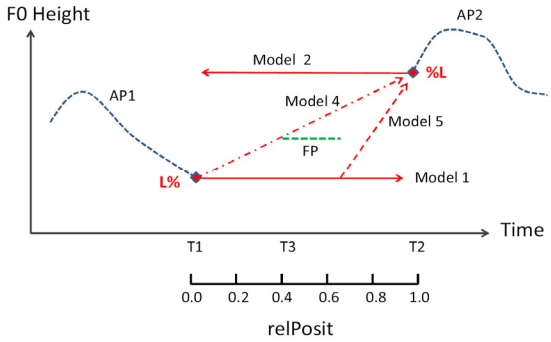

Figure 4: *Schematic representations of 5 prediction models.*

3.3.3. Evaluation of the models

The results of F0 prediction were summarized in Table 2. The RMS errors of prediction were shown in both linear (F0Hz) and log-normalized (F0Logn) values. For each FP, the model showing the least RMS error value is indicated by bold digits, and, the models of two lowest RMS errors are indicated by shaded cells. The FP were classified in terms of Class2. The best prediction was obtained either by Model 4 or 5 in all FP classes. The performances of Models 4 and 5 are very close in all FP except for DE, where Model 4 is superior to Model 5.

4. Discussion

4.1. Evaluation of the precision of prediction

The superiority of Models 4 and 5 over other models are the matter of relative comparison. As shown in the last row of Table 2, the overall RMS prediction errors of the 2 best models (4 or 5) were 0.64 and 0.63 in F0Logn (or, equivalently, 22.2 and 23.0 in Hz). 'Absolute' evaluation of these values is needed. There are reasons to believe that these performances are not bad ones.

First, Figure 5 shows the correlation between the F0 in Hz of FP and the F0 estimated by Model 4. The correlation coefficient is 0.809 and highly significant (t = 96.2074, df = 4890, p-value < 2.2e-16).

Second, Figure 6 compares the density distributions of RMS estimation errors by Model 4, F0 ranges of FP (i.e. the difference between the maximum and minimum F0 in a FP), and the F0 ranges of ordinary AP. F0 are log-normalized. It can be seen from this figure that distribution of the F0 ranges of FP is much narrower than that of ordinary AP. It is also clear that the range of estimation error is even smaller than that of FP. These facts suggest that, from the statistical point of view, it is expected that most of the estimated F0 values are in the fair vicinity of observed F0 values. As a matter of fact, it turns out that 53% of estimated F0 values (in Hz) locate within the ranges of the target FP, and 74% of them locate ± 10 Hz of the ranges.

Table 2: *Comparison of the RMS errors in the F0 estimation.*

Filled Pause	N	Model 1		Model 2		Model 3		Model 4		Model 5	
		F0Hz	F0Logn	F0Hz	F0Logn	F0Hz	F0Logn	F0Hz	F0Logn	F0Hz	F0Logn
AN	411	26.9	0.64	37.0	0.87	28.6	0.70	25.7	0.65	**26.1**	**0.62**
DE	118	76.6	1.83	28.7	0.54	38.9	0.77	**28.5**	**0.52**	46.2	1.02
E	769	26.5	0.76	27.3	0.71	27.0	0.69	20.6	0.58	**18.4**	**0.50**
EH	1790	24.4	0.72	31.4	0.88	25.1	0.72	22.1	0.69	**22.1**	**0.64**
ET	221	30.0	0.79	35.5	0.86	30.5	0.77	**24.4**	**0.65**	28.0	0.73
M	194	31.6	0.91	22.7	0.57	23.0	0.60	**18.0**	**0.49**	21.3	0.59
ALL	3975	27.6	0.77	30.7	0.81	26.2	0.70	22.2	0.64	23.0	0.63

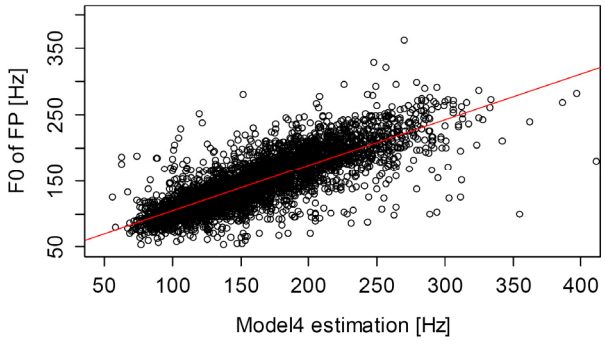

Figure 5: *Correlation between the observed and estimated F0 in Hz. Estimation is by Model 4. Regression line is overlaid.*

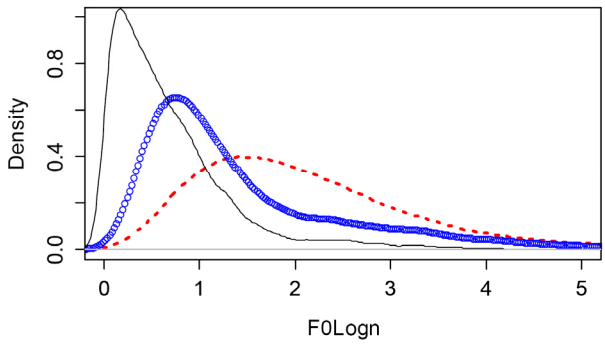

Figure 6: *Comparison of the density distributions of estimation error by Model 4 (real line), F0 range of FP (circle), and F0 range of ordinary AP (dotted line). F0 values are log-normalized.*

4.2. Phonological assessment of models

Five prediction models used in the current study can be classified into two types from a point of view of phonology. Models 1, 2, and 3 are 'phonological' models in that they are based upon the copying (one of the most typical phonological manipulations) of boundary L tones. In these models, a phonological tone is supposed to be associated to FP.

Model 4, on the other hand, is a 'phonetic' model in that it does not include phonological manipulation of tones, and, the FP are not supposed to have any tonal association. The F0 height of the FP is computed on the basis of purely interpolation between the preceding and following boundary L tones. Lastly, Model 5 is a hybrid of 'phonological' and 'phonetic' models.

The results of model comparison revealed that purely 'phonological' models are much inferior to 'phonetic' models in terms of the performance of prediction. From this, it is perhaps safe to conclude that the use of 'FL' and 'FH' labels in the X-JToBI annotation was a non-essential convention. Although there is possibility that these labels reflects native speakers' tendency in the perception of the relative height of FP, the judgments seem to be predictable from the FP's occurrence timing (i.e. relPosit).

To examine the validity of this hypothesis, the relation between the mean relPosit value and the rate of 'FH' labeling was examined for Class1 FP that have frequency higher than 8 and occurring in the environment where %L of the following AP is higher than the L% of the preceding AP.

The Class1 FP that showed the lowest rates of 'FH' label included IH (0.0%, *N*=9), O (0.0%, *N*=20), OH (0.0%, N=33), and UH (0.0%, N=8). And the relPosit values of these FP concentrate in the lower end of ordinate in Figure 1. On the contrary, the FP that showed the highest rates of 'FH' included DE (53.5%, *N*=86), M (28.4%, *N*=130), and MO (21.6%, *N*=28). The relPosit values of these FP tend to concentrate in the higher end in Figure 1.

The overall correlation between the mean relPosit value and the mean rate of FH label is 0.523, and statistically significant (t = 2.605, df = 18, p-value = 0.01791). Note that the correlation is not very high because there are many Class1 FP whose relPosit values scatter widely along the ordinate of Figure 1. Exactly the same tendency is observed when all FP are analyzed (i.e. including the cases where %L is not higher than L%), but the correlation coefficient becomes 0.466 (the correlation, however, is still significant at 0.05).

5. Concluding remarks

The present study revealed the 'phonetic' nature of the F0 height of FP in Japanese. This finding coincides largely with the conclusion of [12] that analyzed the intonation of clause-internal FP in English, but the present finding covers the FP in clause-initial positions as well. The remaining problems include analysis of the cases where more than two FP occur consecutively, finer comparison of Models 4 and 5, inspection of cases where large prediction errors were observed, reexamination of the assumption that F0 is nearly flat within a FP, and so forth. Pilot examination of F0 prediction models by means of synthetic speech is currently underway.

6. Acknowledgements

This study is supported by the Kakenhi grant no.23520483 to the present author. It is also supported by a NINAJL project "Basic Research on Corpus Annotation".

7. References

[1] M.H. Siu and M. Ostendorf "Modeling disfluencies in conversational speech", *Proc. ICSLP '96*, pp. 386–389, 1996.

[2] M. Watanabe, K. Hirose, Y. Den and N. Minematsu. "Filled pauses as cues to the complexity of following phrases", *Proc. INTERSPEECH 2005*, pp. 37–40, 2005.

[3] M. Swerts, "Filled pauses as markers of discourse structure." *Journal of Pragmatics* 30, pp. 485–496, 1998.

[4] T. Sadanobu and Y. Takubo, "Danwa ni okeru shinteki monitaa kikou", Gengo Kenkyu, 108, pp. 74–93, 1995

[5] S. Schachter, N. Christenfeld, B. Ravina and F. Bilous. "Speech disfluency and structure of knowledge", *Journal of Personality and Social Psychology* 60 (3), pp. 362–367, 1991.

[6] E. Shriberg, "Phonetic consequences of speech disfluency." *Proc. ICPhS 1999*, pp. 619–622,1999.

[7] J. Adell, A. Bonafonte, and D. Escudero. "Filled pauses in speech synthesis", *Speech Prosody 2010*, pp. 1–4, 2010.

[8] S. Furui, K. Maekawa, H. Isahara, T. Shinozaki and T. Ohdaira. "Toward the realization of spontaneous speech recognition." *Proc. ICSLP 2000*, pp. 518–521, 2000.

[9] J. Pierrehumbert and M. Beckman. *Japanese Tone Structure*. The MIT Press, 1988.

[10] K. Maekawa, H. Kikuchi, Y. Igarashi and J. Venditti. "X-JToBI: An extended J_ToBI for spontaneous speech", *Proc. ICSLP 2002*, pp. 1545–1548, 2002.

[11] K. Maekawa, "Corpus of Spontaneous Japaese: Its design and evaluation", *Proc. ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, pp. 7–12, 2003.

[12] E.E. Shriberg and R. J. Lickley. "Intonation of clause-internal filled pauses", *Phonetica* 5, pp. 172–179, 1993.

Proceedings of

DiSS 2013

The 6th Workshop on Disfluency in Spontaneous Speech

KTH Royal Institute of Technology
Stockholm, Sweden
21–23 August 2013

TMH-QPSR
Volume 54(1)

Edited by
Robert Eklund

Conference website: http://www.diss2013.org
Proceedings also available at: http://roberteklund.info/conferences/diss2013

Cover design by Robert Eklund
Front cover photo by Jens Edlund and Joakim Gustafson
Back cover photos by Robert Eklund

Proceedings of DiSS 2013, The 6th Workshop of Disfluency in Spontaneous Speech
held at the Royal Institute of Technology (KTH), Stockholm, Sweden, 21–23 August 2013
TMH-QPSR volume 54(1)
Editor: Robert Eklund
Department of Speech, Music and Hearing
Royal Institute of Technology (KTH)
Lindstedtsvägen 24
SE-100 44 Stockholm, Sweden

ISBN 978-91-981276-0-7
eISBN 978-91-981276-1-4
ISSN 1104-5787
ISRN KTH/CSC/TMH--13/01-SE
TRITA TMH 2013:1

Printed by Universitetsservice US-AB, Stockholm, Sweden, 2013