INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

A New Model of Final Lowering in Spontaneous Monologue

Kikuo Maekawa

National Institute for Japanese Language and Linguistics, Japan

kikuo@ninjal.ac.jp

Abstract

F0 downtrend observed in spontaneous monologues in the Corpus of Spontaneous Japanese was analyzed with special attention to the modeling of final lowering. In addition to the previous finding that the domain of final lowering covers all tones in the final accentual phrase, it turned out that the last L tone in the penultimate accentual phrase played important role in the control of final lowering. It is this tone that first reached the bottom of the speaker's pitch range in the time course of utterance; it also turned out that the phonetic realization of this tone is the most stable of all tones in terms of the F0 variability. Regression model of F0 downtrends is generated by generalized linear mixed-effect modeling and evaluated by cross-validation. The mean prediction error of z-normalized F0 values in the best model was 0.25 standard deviation.

Index Terms: final lowering, F0 downtrends, Corpus of Spontaneous Japanese, generalized linear mixed-effect model

1. Introduction

Speech fundamental frequency (F0) lowers gradually from the beginning to the end of an utterance. In addition to this overall declination, it is believed that F0 of declarative utterances lowers locally near the end of utterance so that the F0 approximates the bottom of the speaker's pitch range and indicate finality [1]. This event, called final lowering (FL hereafter), is believed to belong, basically, to the cognitive aspect of speech, and is observed in typologically diverse languages. As for Japanese, however, there is a debate concerning the domain of FL. Poser concludes that the domain of FL is the last mora (or syllable) of intonation phrase [2]. According to his theory, FL is an 'edge effect' of phrase phonology; the occurrence of FL is hence simultaneous to the end of utterance. Pierrehumbert & Beckman, on the other hand, suggested wider domain of FL based upon the comparison of question and declarative utterances, but they did not specify how wide the domain could be [3]. Also, Umeda suggested that FL in Japanese was situation-dependent and observed only in utterances of read-aloud material [4].

Recent studies of spontaneous monologue in the *Corpus of Spontaneous Japanese* (CSJ) provided evidences that FL existed in spontaneous monologue like academic presentation and public speech without prepared script, and its domain covered at least the last accentual phrase (AP hereafter) of utterance [5-7]; unlike Poser's prediction, all phonological tones (see section 2.2 below) in the last AP were located clearly below the line of general declination. In addition, one of these studies suggested the possibility that the tones in the penultimate (last but one) AP were also subject to FL though not as evident as in the final AP [6]. The aim of the present study is to examine this possibility by means of statistical modeling of the same corpus data as in previous studies.

2. Data

2.1. The CSJ-Core

RDB version of the CSJ-Core (ver.2) was used for analysis [8,9]. Data in the CSJ-Core is prosodically annotated by means of the X-JToBI scheme [10]. 177 monologue talks, consisting of 70 academic presentations and 107 extemporaneous public speakings, spoken by 99 males and 78 females were analyzed. The mean length of a talk was 785 sec.

2.2. Constituent tones of AP and F0 normalization

As shown in Fig. 1, AP of Standard Japanese consists of three obligatory phrasal tones: %L tone (in terms of the JToBI notation) that marks the beginning of an AP, which is referred to as ILT (initial low tone) in this paper, H- tone that marks the peak of AP initial pitch rise, which is referred to as IHT (initial high tone), and L% tone that marks the end of an AP, which is referred to as FLT (final low tone) [6,7]. If the AP contains accented word, an extra H* tone that marks the peak of lexical accent appears, which is referred to as ACC. If the accent is on the initial or second syllables of an AP, the IHT is replaced by ACC. FLT is often followed by boundary tones like H% and HL% that make so-called boundary pitch movements (BPM). In this study, however, the behavior of boundary tones other than L% will not be analyzed. It is known that the presence of boundary pitch movements does not affect significantly the height of FLT [6].

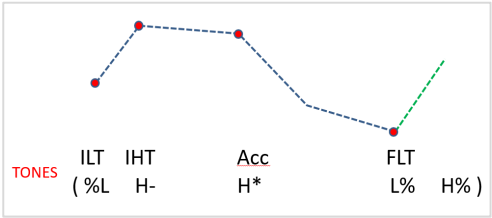

Figure 1: *Tonal representation of an accentual phrase in Japanese*

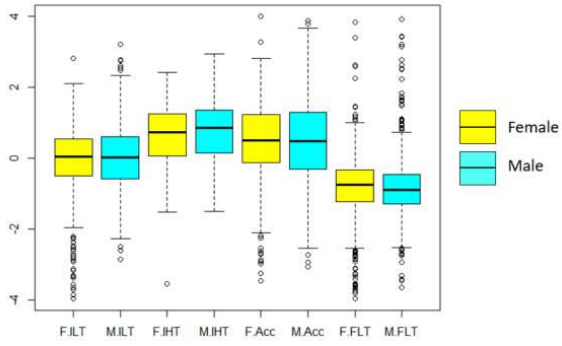

Figure 2: *Distributions of four tones by males and females*

http://dx.doi.org/10.21437/Interspeech.2017-175

CSJ-Core has records of F0 values (in Hz) for all tones except for the cases of vowel devoicing. In this study, the F0 values are normalized for each speaker by computing z-values of the logarithm (log) of F0. Fig. 2 shows distributions of four tones after normalization. Two-way ANOVA of tone-gender revealed significance for the difference among tones (df=3, F=956, p<0.001), but no significance for gender (df=1, F=.0005, p<.942). In this study, normalized F0 values are analyzed so that it is possible to pool male and female samples.

2.3. Extraction of accented AP sequences

In this paper, utterances consisting exclusively of accented APs will be analyzed. Exclusion of unaccented AP simplifies greatly statistical modeling (see section 4 below) with respect to the treatment of downstep [1-3]. Utterances consisting of two to six accented APs were extracted from the CSJ-Core. Filled pauses and word fragments occurring between adjacent APs were ignored. The numbers of extracted utterances were: 145, 122, 79, 51, and 45 respectively for utterances consisting of two, three, four, five, and six APs.

3. Analysis

3.1. Observed F0 behavior

Fig. 3 shows the mean tone values of normalized F0. The unit of the ordinate is standard deviation (SD). Eighty tones are included in the figure, i.e. 4*(2 + 3 + 4 + 5 + 6) = 80. Tones belonging to the same AP are connected by line, and, utterance lengths are distinguished by the types of lines and markers. Three important F0 behaviors that are reported in previous studies [1-3,5-7] can be recognized.

First, there is overall downtrend of F0 from the beginning to the end of utterance. This trend can be interpreted as a composite effect of downstep [1-3] (which is caused by lexical accent in each AP) and F0 declination as a simple function of time [3]. See the discussion in section 4.1 below for the treatment of downstep in this study.

Second, all tones in the final AP of given utterance length locate much lower than the location predicted by above-mentioned overall downtrend, and, they locate in about the same range of F0.

Third, tones in earlier APs tend to become higher as the utterances get longer; this effect, called anticipatory rising, is reported for English [11] and Chinese [12] in the studies of FL.

Because the first and second observations are congruent with the prediction by the theory of FL [1], previous studies concluded that FL existed in Japanese spontaneous monologue, and the domain of FL covered the whole final AP rather than the last mora alone [5-7].

In addition to these previously reported F0 tendencies, there is a novel observation to be noted here: high stability of the FLTs in the penultimate APs. FLTs in the penultimate APs are distributed in a very narrow F0 range around -1.0, regardless of utterance lengths. Fig.4 shows the mean FLT values in the three AP locations in the end of utterance, i.e final, penultimate, and antepenultimate. Note that utterances of 2AP length do not have antepenultimate AP.

Table 1 shows the standard deviations of four tones in the final, penultimate, and antepenultimate APs. It turns out that the variability of FLT in the penultimate AP is by far the smallest of all tones in all locations.

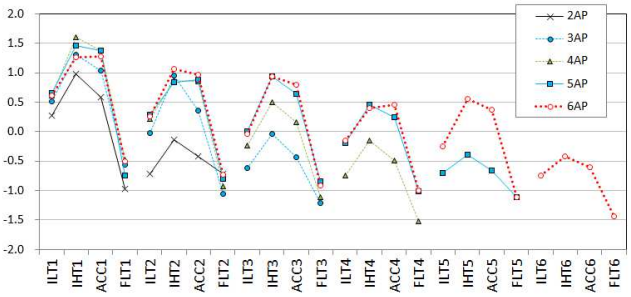

Figure 3: *Mean normalized F0 values of observed 80 tones. Unit in the abscissa is standard deviation*

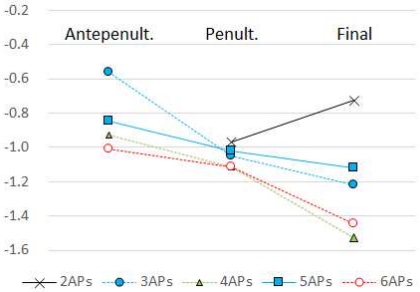

Figure 4: *Comparison of mean FLT values in the three AP locations in the end of utterance*

Table 1 shows the standard deviations of four tones in the final, penultimate, and antepenultimate APs. It turns out that the variability of FLT in the penultimate AP is by far the smallest of all tones in all locations.

Table 1: *Standard deviations of tones in the last three APs*

Tone	Antipenultimate	Penultimate	Final
ILT	0.291	8.571	0.797
IHT	0.744	0.371	2.159
ACC	0.204	0.490	0.338
FLT	0.392	0.086	0.213

3.2. Hypothesis

Tendencies of F0 downtrend observed in Fig. 3 and Table 1 suggest the following hypothesis on the timing of FL: *the temporal target of FL, i.e. the timing when F0 reaches the baseline of a speaker's pitch range is NOT the end of utterance as hypothesized in the past. It is rather the end of penultimate AP*. It is expected under the new hypothesis that FL works on both the penultimate and final APs but in different ways. In the penultimate AP, effect of FL lowers all tones in the AP equally so that the FLT is located on the baseline (hence no interaction between tones and FL). In the final AP, tones are also lowered by the effect of FL, but there should be interaction between the tone types and the effect of FL. Because F0 has already reached the baseline in the end of penultimate AP, the FLT (and ILT perhaps) in the final AP would become too low (much below the baseline) without an interaction that boosts them up so that they locate on the baseline.

Fig. 5 compares schematically the predictions under the traditional and new hypotheses, using as example an utterance consisting of four APs. Real line in blue represents the case without FL. Green broken line represents the FL under the

traditional view. And, the red dotted line represents the prediction under the new hypothesis.

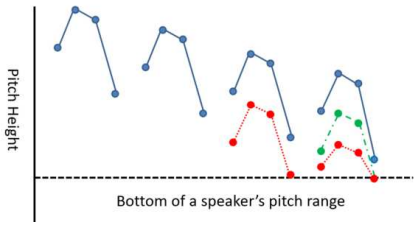

Figure 5: *Schematic comparison of traditional and new hypotheses*

4. Modeling

4.1. Models

F0 downtrend observed in Fig.3 was modeled by use of generalized linear mixed-effect model or GLMM. Table 2 summarizes the variables used in the modeling. Three GLMM models are compared in terms of their performance of predicting tone values which are represented by the 'lnHeight' variable. The models can be written as the following in the notation of the lme4 package of the R language:

Model0: lnHeight ~ tone + loc + final + (1|pBreak) + (1|TalkID)

Model1: lnHeight ~ tone + loc + final + len_loc + penult + (1|pBreak) + (1|TalkID)

Model2: lnHeight ~ tone + loc + final + len_loc + penult + tone:loc + tone:len_loc + tone:final + tone:penult + (1|pBreak) + (1|TalkID)

Variables separated by the '+' symbol is the independent variables. Tems that connect two variables by colon like 'tone:final' or 'tone:penult' stand for statistical interactions. And the tems '(1|pBreak)' and '(1|TalkID)' represent random effects on the intercept.

Table 2: *Variables used in the* GLMM *modeling*

Variables	Gloss
lnHeight	Normalized logarithm of F0
Tone	Type of tones (one of ILT, IHT, ACC, and FLT)
Loc	Location of AP in utterance (1-6)
len_loc	Subtraction of AP location from the utterance length, i.e., the length of utterance after the AP in question (0-5)
Penult	1 if the AP is in penultimate position, 0 otherwise
Final	1 if the AP is in final position, 0 otherwise
pBreak	BI between the current and previous AP (one of 2, 2+p, 2+b, 2+bp, and 3)
TalkID	ID of speakers

Model0 is the simplest model. In this and other models, variable 'tone' stands for inherent values of four tones of AP (see Fig.2). Variable 'loc' is expected to stand for the effect of F0 declination as a function of AP location in an utterance. Because all utterances are sequences of accented APs, the effect of downstep as well as simple declination as a function of time are both concerned with this variable. Note, the so-called 'exponential' nature of downstep [1] is reexpressed here by linear additive function, because the models are concerned with logarithmic variable 'lnHeight'. See also the discussion about random effect below. The variable 'final' is expected to stand for the effect of FL on the final AP. As a whole, Model0 has the closest resemblance to the traditional linguistic models of F0 downtrends [1-3].

Model1 has two extra independent variables. Variable 'len_loc' is expected to express the effect of anticipatory rising (see 3.1 above). Variable 'penult' is expected to stand for the effect of FL on the penultimate AP. Model2 differs from the two previous models in that it has terms of interaction between the variable 'tone' and other four variables 'loc', 'len_loc', 'final' and 'penult'. These terms enable the adjustments of F0 values of each tone at various locations in an utterance.

Lastly, all models have two random effect rems, i.e. '(1|pBreak)' and '(1|talkID).' The former term needs some mention here. In spontaneous speech, unlike read speech, downstep rarely continues over more than three APs. In fact, in the CSJ-Core, 43% of AP boundaries are '3' boundary that resets downstepping. In all three models used in this study, resetting of downstep is treated as a random effect that is beyond the control of experiment design. Accordingly, the effect of variable 'loc' described above is expected to concern mostly with the F0 declination as a simple function of time.

4.2. Results

Three models were fitted to the observed data using the *lmer* function of the *lme4* and *lmerTest* packages [13,14] of the R language (ver. 3.3.1). Results are compared in the upper half of Table 3, where each row of the table corresponds to a variable or interaction term, and in each cell is shown estimate (i.e. the weight) of variables and/or interaction terms with its statistical significance. According to the convention of R, estimates of nominal variable 'tone' are shown as the relative values to one of its levels whose estimate is zero. In this case, the level 'ACC' is set to zero, hence is not shown in the table.

The last two rows of Table 3 show AIC and mean prediction error (unit is standard deviation) of the three models. Model2 shows the best overall performance with respect to both AIC and mean prediction error. Fig. 6 shows the mean tone values predicted by Model2. F0 downtrends observed in Fig. 2 are reconstructed in a successful manner.

Table 3: *Results of model estimation*
Significance codes: 0 '***', 0.001 '**', 0.01 '*'

Variables and interactions	Model0	Model1	Model2
Intercept	0.592 **	0.503 **	0.547* **
ILT	-0.456***	-0.456***	-0.480 **
IHT	0.202***	0.188***	0.298
FLT	-1.447***	-1.450***	-1.616***
Loc	-0.091***	-0.049***	-0.048 *
len_loc	---	0.064***	0.098***
Final	-0.542***	-0.577***	-0.831***
penult	---	-0.214***	-0.327***
ILT:loc	---	---	0.017
IHT:loc	---	---	-0.052
FLT:loc	---	---	0.035
ILT:len_loc	---	---	-0.050
IHT:len_loc	---	---	-0.043
FLT:len_loc	---	---	-0.047
ILT:final	---	---	0.363 *
IHT:final	---	---	0.133
FLT:final	---	---	0.767***
ILT:penult	---	---	0.170
IHT:penult	---	---	0.196
FLT:penult	---	---	0.152
AIC	9548	9404	9278
Mean error (SD)	0.303	0.231	0.176

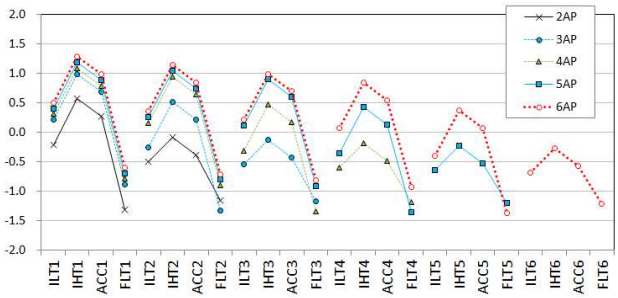

Figure 6: *Prediction made by the Model2 (closed data)*

Estimated random effect values of 'pBreak' variable are -0.224, -0.169, -0.013, 0.023, and 0.384 respectively for levels of '2'(typical AP boundary), '2+p' (AP boundary followed by a pause, which tends to be perceived to be stronger than '2'), '2+bp' (boundary followed by both BPM and pause, stronger than '2'), '2+b' (followed by BPM, stronger than '2') and '3' (intermediate or intonational phrase boundary with resetting of downstep). As expected, '3' has positive value, which implies its F0 resetting function, '2' has negative value, which implies downstep, and, intermediate levels of '2+p', '2+b', and '2+bp' are associated with intermediate values.

4.3. Evaluation of generalizability

The prediction shown in Fig. 5 was based upon closed data set, hence it might have run the risk of overfitting. A sort of cross-validation was conducted to evaluate the generalizability of the model. For this purpose, the following procedure was repeated ten times. The whole data consisting of 4230 tones were divided randomly into two sets; the training set containing 2230 tones, and the test set containing 2000 tones. The training set was used to construct a model (Model2) thereby to predict mean tone values like Fig. 6. On the other hand, the test set was used to compute 80 mean values of 'observed' tone values like Fig. 3. Then, mean prediction error was computed using the predicted and observed values.

Mean and standard deviation of the prediction error across repetitions were 0.25 and 0.02 respectively. The mean error of 0.25 SD corresponds to 9.0 and 12.8 Hz respectively for males and females on linear scale. And the difference of mean prediction error of Model2 between the closed data (Table 3) and open data set (shown above) is less than 4 Hz both for males and females. It can be concluded that the model has high generalizability to new data.

5. Discussions

5.1. Interpretation of the interactions

Analyses in the previous section revealed that the FL and other F0 downtrends in Standard Japanese could be modeled by means of GLMM without losing generalization to new data.

From a point of view of FL, it is important that the new hypothesis described in section 3.2 was supported well by the statistical test of the estimated model parameters. In the "Model2" column of Table 3, both 'final' and 'penult' variables have significant negative estimates; this fact is congruent with the hypothesis that both penultimate and final APs are in the domain of FL. More important is the lack of significant interaction in the case of penultimate AP and the presence of significant interactions in the case of final AP. Again, this is congruent with the hypothesis. It is especially important to note that it is the two L tones of the final AP (i.e. ILT and FLT) that interacted with the 'final' variable, and the estimates of the interaction terms both have positive ---rather than negative--- values. As predicted in 3.2, these interactions are called upon so that the tones in the final AP do not exceed the baseline; it is hence natural that the interaction operates selectively on low tones, rather than high tones. Note also that the interaction terms of ILT*final and FLT*final turned out to be significant in all repetitions of cross-validation, and, the terms IHT*final and ACC*final did not show any significance in any repetitions.

5.2. Implication of the new hypothesis

The newly proposed control mechanism of FL has an important implication for the role of prosody in speech communication. It implies that prosody predicts the end of utterance in advance of the end of text, given that listeners can perceive the timing where F0 reaches the speaker's F0 baseline. In the case of the CSJ, it means that listeners have, on average, temporal margin of 0.66 sec (the mean duration of utterance-final APs) or longer (considering the case there is pause between the final and penultimate APs) before the utterance really reaches its end. This margin can be useful for the speech information processing by listeners. As Ishimoto and Enomoto suggested based on their perception experiment [15], signaling utterance finality at the very end of an utterance is of no practical use for listeners.

5.3. Remaining issue on F0 baseline

The present analysis takes for granted the supposition that the F0 baseline is a fixed value. In the case of Fig. 3, for example, the baseline seems to be located at around -1.0 of ordinate. There are, however, cases where FLT of final AP located lower than the baseline (See the utterances of 4ASP and 6AP). This might well be a mere coincidence, but it is also possible that F0 baseline is not fixed to a single value. Some researchers reported that baseline can be different depending on the position in discourse in English [16,17]. The deviant values in Fig. 3 might be related to the bias caused by discourse positions. This issue should be the theme of a separate study.

6. Conclusion

FL in Standard Japanese is analyzed using the data of spontaneous monologue. The main conclusion is that existing linguistic theory of FL needs to be replaced by a new theory that incorporates the role played by the penultimate accentual phrase, especially its final boundary L tone. The theory brings new insight about the role of FL in speech communication.

Another conclusion that can be drown form this study, which is technical rather than linguistic, is the usefulness of generalized linear mixed-effect model analysis in F0 modeling studies. The generated model is highly congruent with the predictions of the new hypothesis and is generalizable to new data.

7. Acknowledgements

This work is supported by the Kakenhi grants 23520483 and 26284062 to the present author, and the research grant of the Center for Corpus Development of NINJAL. The author thanks Yasuharu Den and Daichi Mochihasi for their advices on statistical analyses.

8. References

[1] M. Liberman & J. Pierrehumbert (1984) "Intonational invariance under changes in pitch range and length." In Aronoff & Oehrle (eds.) *Language Sound Structure*. MIT Press, pp. 157-233.

[2] W. Poser (1984) *The phonetics and phonology of tone and intonation in Japanese*. Ph.D. diss. MIT.

[3] J. Pierrehumbert & M. Beckman (1988) *Japanese Tone Structure*. MIT Press.

[4] N. Umeda (1980) "Fo declination is situation dependent." *Journal of Phonetics*, 10, 279-291.

[5] K. Maekawa (2010) "Final lowering and boundary pitch movements in spontaneous Japanese." *Proc. DiSS-LPSS Joint Workshop 2010*, Tokyo, pp.47-50.

[6] K. Maekawa (2013) "Nihongo Jihatsuonsei niokeru final lowering no seikiryouiki" (Domain of final lowering in spontaneous Japanese). *Proc. 27th Annual Convention of the Phonetic Society of Japan*, pp. 61-65.

[7] K. Maekawa (2014) "Domain of final lowering in spontaneous Japanese." *Journal of Acoustical Society of America* 135 (4), p.2194.

[8] K. Maekawa. "Corpus of Spontaneous Japanese: Its Design and Evaluation." *Proc. ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition (SSPR)*, Tokyo, pp.7-12.

[9] CSJ-RDB: http://pj.ninjal.ac.jp/corpus_center/csj/en/rdb-index.html (as of 5 March 2017)

[10] K. Maekawa et al. (2002) "X-JToBI: An extended J_ToBI for spontaneous speech." *Proc. ICSLP2002*, Denver, pp.1545-1548.

[11] A. Arvaniti (2009) "On the Presence of Final Lowering in British and American English". In Gussenhoven, C. & Riad, T., eds. *Tones and Tunes*, Vol. 2: Experimental Studies in Word and Sentence Prosody. Mouton de Gruyter, Berlin and New York, pp. 317-347.

[12] C. Shih (2000) "A declination model of Mandarin Chinese." In A. Botinis (ed.) *Intonation: Analysis, Modeling and Technology*. Springer Netherlands, pp. 243-268.

[13] lme4: https://github.com/lme4/lme4/ (as of 5 March 2017)

[14] lmerTest: https://cran.r-project.org/web/packages/lmerTest/lmerTest.pdf (as of 5 March 2017)

[15] Y. Ishimoto & M. Enomoto (2016) "Experimental investigation of end-of-utterance perception by final lowering in spontaneous Japanese." *Proc. 2016 Oriental COCOSDA*, Bali, pp. 205-209.

[16] J. Hirschberg & J. Pierrehumbert (1986). "The intonational structure of discourse." *Proc. 24th annual meeting of ACL*, pp. 136-144.

[17] R. Hermann (2000) "Phonetic markers of global discourse structures in English." *Journal of Phonetics*, 28, pp. 466-493.