Prosody, Supporting Real-Time Conversation

Hiroki Oohashi[1], *Tomoko Ohsuga*[2], *Yasuo Horiuchi*[3], *Hideaki Kikuchi*[1], *Akira Ichikawa*[1]

[1]Faculty of Human Science, Waseda University, Japan
[2]National Institute of Informatics, Japan
[3]Graduate School of Advanced Integration Science Chiba University, Japan
hiro084@gmail.com, osuga@nii.ac.jp, hory@faculty.chiba-u.jp,
kikuchi@waseda.jp, a.ichikawa@aoni.waseda.jp

Abstract

We assume that prosody contains information forenoticing segment boundaries, syntactic structures, and turn transitions and enable us to predict these more easily. We examined this assumption using the F_0 model. Concretely speaking, as for forenotices, we examined whether or not the F_0 model parameters can lead to segment boundaries, dependencies of phrases, and turn transitions. On the other hand, as for predictions, we conducted cognitive experiments on turn-taking by presenting stimulations containing only prosody and not phonological information. As a result, the segment boundaries were exactly forenoticed at an accuracy of about 60%, the dependencies of phrases were done at about 80%, the turn transitions were done at about 70%, and the possibility of predictions about turn transitions was indicated.

Index Terms: real-time conversation, word segmentation, turn-taking, syntactic structure, F_0 model

1. Introduction

In spite of the volatile nature of speech sound, speech conversations are smooth. In order to enable a such smooth conversation, the hearer needs to extract words from the mental lexicons consisting of an enormous amount of words, parse syntax structures, comprehend meanings and predict transition relevance places, while listening to continuous speech in real time.

Although previous studies pointed out that prosody contains information about segment boundaries, syntactic structures, and turn transitions [1, 2, 3, 4], examinations from the viewpoint of real-time conversations are insufficient.

We have been examining the assumption that forenoticing information about segment boundaries, syntactic structures, and turn transitions exists in prosody, enables us to predict them and achieves smooth real-time conversations.

So, to establish this assumption, we need to reveal that forenoticing information exists in speech sounds as physical information, and that human beings can percept and utilize them for making predictions in real time.

This paper illustrates the examination cirsumstances for forenoticing information using a quantitative model for generating sentence F_0 contours (called F_0 model) [5]. We explain the method of estimating the F_0 model parameters first and then describe our examinations of the forecasting of segment boundaries, syntactic structures, and turn transitions and on the predictions about turn transitions. In this paper, *forenotice* means that the segment boundaries, syntactic structures, and occurrences of turn transitions are physically determined and indicated in advance. On the other hand, *prediction* means that human beings can percept or judge the segment boundaries, syntactic structures, and occurrences of turn transitions in any way in advance.

However, we don't advocate human beings predicting using the methods from our analysis in this study. These methods are not suitable for the modeling of real-time recognitions.

2. Method of Estimating the F_0 Model Parameters

We adopted the genetic algorithm proposed by Kimura et al. [4], in order to estimate the F_0 model parameters in this study.

The F_0 model is described using (1). (2) is a function of the phrase control mechanism and (3) is a function of the accent control mechanism.

$$lnF_0(t) = F_b + \sum_{i=1}^{I} A_{pi} G_{pi}(t - T_{0i}) + \sum_{j=1}^{J} A_{pj}\{G_{aj}(t - T_{1j}) - G_{aj}(t - T_{2j})\} \quad (1)$$

$$G_{pi}(t) = \begin{cases} \alpha_i^2 t e^{-\alpha_i t} & (t \geq 0) \\ 0 & (t < 0) \end{cases} \quad (2)$$

$$G_{aj}(t) = \begin{cases} min[1 - (1 + \beta_j t)e^{-\beta_j t}, \theta_j] & (t \geq 0) \\ 0 & (t < 0) \end{cases} \quad (3)$$

The symbols in (1)–(3) are F_b : logarithmic value of bias level upon which all the phrases and accent components are superposed to form an F_0 contour, A_{pi} : magnitude of the ith phrase command, T_{0i} : instant of occurrence of the ith phrase command, α_i : natural angular frequency of the phrase control mechanism to the ith phrase command, A_{aj} : amplitude of the jth accent command, β_j : natural angular frequency of the accent control mechanism to the jth accent command, T_{1j}, T_{2j} : onset and end of the jth accent command, and θ_j : ceiling level of the accent component for the jth accent command, in Kimura et al. [4] $\theta_j = 0.9$.

Although α_i and β_j are constant numbers in the original F_0 model, the proposed method by Kimura et al. [4] estimates both parameters as variables .

Here, according to the function formats of the phrase and the accent control mechanism, it is important that the input parameters into each mechanism decide the F_0 patterns at the onset of the F_0 pattern generation.

3. Forenotice of Segment Boundary

In perception experiments conducted by Hatano [1], more than 80% of the subjects discriminated accentual phrase (called AP)

boundaries based on prosody. In this study, we state that the F_0 patterns contain physical information forenoticing the AP boundaries.

3.1. Materials

Hatano [1] recorded 76 speech materials consising of four kinds of AP pairs (3+6 morae, 4+5 morae, 5+4 morae, 6+3 morae), read by a Japanese male.

From these materials, we use 47 of them that were successfully estimated for the F_0 model parameters by using the above-mentioned method and passed the selections mentioned later.

3.2. Procedure

First, we selected materials that consisted of one phrase component, which contains two accent components.

Since then, the F_0 model parameters and AP length were assumed to be explanatory variables and a response variable, respectively, and a regression analysis was done. Concretely speaking, the response variable was a former AP length and the explanatory variables were $F_b, A_{p1}, \alpha_1, \beta_1, A_{a1}, T_{21}, \beta_1 \times T_{21}$ (interaction between β_1 and T_{21}. In the following, the interaction between A and B is described as $A \times B$), and $\alpha_1 \times \beta_1 \times T_{21}$.

3.3. Results

(4) is the regression formula containing all the parameters and Table 1 shows the fifth high ranking for adjusting R^2 and each regression formula. At this time, materials where β_1 were $15 \leq \beta_1 \leq 25$ were used when regarding an extremely small or large β_1 as an estimation error. (4) and Table 1 indicate that $\alpha_1 \times \beta_1 \times T_{21}$ and T_{21} play an important role in forenoticing the AP length.

Subsequently, we examined the regression errors by conducting a leave-one-out cross-validation test. For the examination, we adopted a best model in the adjusting R^2 criterion for use as the regression model. Fig. 1 shows the distribution of the error margins. The regression errors for 63.2% of the materials were less than 1 mora, and the ones for 94.7% were the back and forth 1 mora.

$$\begin{aligned} 0.10 \times F_b + 0.03 \times A_{p1} + 0.28 \times \alpha_1 + 0.11 \times A_{a1} \\ +0.16 \times \beta_1 + 0.33 \times T_{21} + 0.06 \times \beta_1 \times T_{21} \\ -0.33 \times \alpha_1 \times \beta_1 \times T_{21} + 0.65 \end{aligned} \quad (4)$$

Table 1: *The fifth high ranking for adjusting R^2 and each regression formula.*

Formula	Adjusting R^2
$\alpha_1 + A_{a1} + T_{21} + \beta_1 + \alpha_1 \times \beta \times T_{21}$	0.614
$F_b + \alpha_1 + A_{a1} + T_{21} + \beta_1 + \alpha_1 \times \beta_1 \times T_{21}$	0.612
$F_b + A_{p1} + \alpha_1 + A_{a1} + T_{21} + \beta_1 + \alpha_1 \times \beta_1 \times T_{21}$	0.610
$\alpha_1 + T_{21} + \beta_1 + \alpha_1 \times \beta_1 \times T_{21}$	0.599
$F_b + \alpha_1 + T_{21} + \beta_1 + \alpha_1 \times \beta_1 \times T_{21}$	0.585

3.4. Discussion

We examined whether or not the F_0 patterns contained physical information of forenoticing AP boundaries. This examination led to two conclusions. First, T_{21} and the interactions among α_1, β_1 and T_{21} contribute to forenoticing AP boundaries. Second, the F_0 patterns are able to forenotice AP boundaries if the error margins are of back and forth 1 mora, at an accuracy of about 95%.

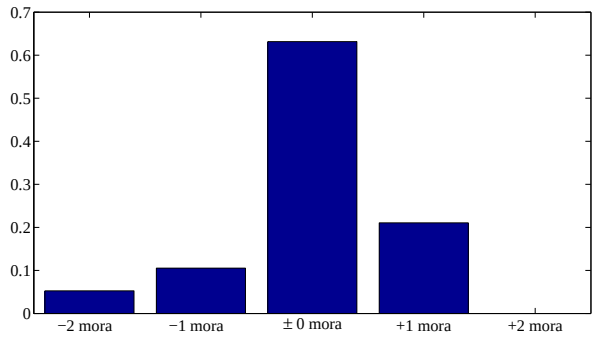

Figure 1: *The distribution of the error margins. The horizontal axis does the range of error per mora and the vertical axis means the rate of materials.*

On the other hand, the regression errors of 63.2% of the materials were less than 1 mora. In summary, only about 60% of the AP boundaries are exactly forenoticed by only the F_0 patterns, and this accuracy is 20% or more less than the results from the perceptual experiments done by Hatano [1]. This is probably because the materials that were used in the previous study contained time information such as the rhythm, but in this study we focused on only the F_0 patterns.

4. Forenotice of Syntactic Structure

Previous studies [6, 7] pointed out down-step, pause and syllable lengthening as solutions for the ambiguities in syntactic structures. In this study, we indicate that F_0 patterns contain physical information forenoticing syntactic structures.

4.1. Materials

We used SET-A of the ATR Phonetically Balanced Sentence [8]. SET-A includes 50 sentences and each sentence was read carefully by 10 Japanese professional speakers (6 male speakers and 4 female speakers). From this data set, we used materials that were successfully estimated for the F_0 model parameters by using the above mentioned method. As a result, we selected 466 APs (298 depend on the following phrase and 168 don't).

4.2. Procedure

We applied a Mahalanobis' generalized distance based on a test for the discriminant analysis, regarding the F_0 model parameters as explanatory variables and whether a preceding AP (called the *PRECEDING*) depends on the following AP (called the *FOLLOWING*) or not as discriminant classes. In order to calculate the correct rates of the discriminant analysis, we conducted a leave-one-out cross-validation test. Then, we selected variables based on the stepwise method with the variable significance level at $P = 5\%$.

We examined the case of using (i) only *PRECEDING* parameters and (ii) *PRECEDING* parameters and a pause between *PRECEDING* and *FOLLOWING*. In particular, in the case of (i), $\widehat{\alpha_1}, A_{a1}, \beta_1, T_{11} - \widehat{T_{01}}, F_{0T11}$ were the proposed variables, and in the case of (ii), the variables in (i) and a pause between *PRECEDING* and *FOLLOWING* were the proposed variables. In the

following, F_{0T1j} is a logarithm of the theoretical value of the F_0 model at T_{1i}. And $\widehat{T_{0i}}$ and $\widehat{\alpha_i}$ are the instance of occurrence of and natural angular frequency of the phrase control mechanism to the phrase command contains the ith accent command, respectively.

4.3. Results

First, the results from (i) using only the *PRECEDING* parameters is described. In this case, the stepwise method selected $T_{11} - \widehat{T_{01}}, A_{a1}$, and β_1, and the correct rate was 0.661 and the Cohen's κ was 0.255.

Next, the results from (ii) where *PRECEDING* parameters and a pause were used is described. In this case, the stepwise method selected $pause, T_{11} - \widehat{T_{01}}, \beta_1, F_{0T11}$, and the correct rate was 0.800 and the Cohen's κ was 0.574. In addition, when using the *PRECEDING* parameters, a pause and the *FOLLOWING* parameters, the correct rate was also about 0.800.

Table 2: *The results from the stepwise method and the correct rate in each condition.*

Condition	Variables	Correct rate
(i)	$T_{11} - \widehat{T_{01}}, A_{a1}, \beta_1$	0.661
(ii)	$pause, T_{11} - \widehat{T_{01}}, \beta_1, F_{0T11}$	0.800

4.4. Discussion

We examined whether or not F_0 patterns contain physical information of forenoticing syntactic structures. As a result, we indicated that only preceding APs were able to forenotice syntactic structures at an accuracy of about 65%, and preceding APs and pauses did it at an accuracy of about 80%.

5. Forenotice and Prediction about Turn Transitions

Previous studies [9, 10] pointed out that phrase final copulas, adjacent pairs, the number of not yet determined dependencies, prosody, and syntax enable us to predict transition relevance places.

In this study, we examined the forenotices and predictions concerning the turn transitions from both physical and cognitive perspectives.

We used a part of the Japanese Map Task Corpus [11]. From this corpus, we ruled out utterances consisting of only proper nouns or back-channel feedbacks, and selected comparatively short and clear ones consisting of 6–14 morae, which had a similar noun (such as "HIDARI"—left, "MIGI"—right, "UE"—above, "SHITA"—below, "SYUPPATSU CHITEN"—departure point, "HAIOKU"—abandoned house, etc.) at the phrase head. Eventually we choose 106 utterances (55 changes, 51 continuous utterances).

5.1. Forenotice of Turn Transitions

From a physical perspective, we examined whether or not the F_0 patterns contained physical information forenoticing turn transitions.

We applied a Mahalanobis' generalized distance based on a test for the discriminant analysis, while regarding the F_0 model parameters as explanatory variables and whether a turn change occurs or a turn continues as discriminant classes. In order to calculate the correct rates of the discriminant analysis, we conducted a leave-one-out cross-validation test. Then, we selected variables based on the stepwise method with the variable significance level at $P = 5\%$.

We used a stepwise method to use A_p and α of the last phrase command, and A_a and β of the last accent command as the proposed variables. As a result, A_p, A_a were selected, and the correct rate was 0.687 and the Cohen's κ was 0.375.

5.2. Prediction about Turn Transitions

From a cognitive perspective, we conducted cognitive experiments in order to examine the predictions about turn transitions. We assumed that the context information I_c, linguistic information I_l, prosody information I_p and information of a body motion—such as the lines of sight, facial expressions, etc.— I_b multimodaly interacted with each other and when the sum of them exceeded threshold T, one can predict the turn transitions. From this assumption, the model in (5) is derived.

$$I_c + I_l + I_p + I_b \geq T \tag{5}$$

We conducted cognitive experiments for the (i) prosody condition—where only prosody information I_p was presented, (ii) text condition—where only linguistic information I_l was presented, (iii) speech condition—where speech sound means $I_l + I_p$ was presented in order to examine the possibilities of predictions in each condition by taking this model into consideration. In particular, subjects judge a change or a continue of a turn based on the information of a preceding utterance without using any information concering that that followed it.

We made materials by reproducing the theoretical values of the F_0 model with triangular waves and smoothed these out by adding power information. These materials contained prosody such as F_0, power, and time structures, but didn't contain phonological information. We presented these materials for the (i) prosody condition. For the (ii) text condition, the character string translated in the kana character was presented in a one by one by one character. For the (iii) speech condition, selected utterances were presented.

The subjects in conditions (i), (ii) and (iii) were as follows: (i) 12 subjects (5 male and 7 female), (ii) 10 subjects (5 male and 5 female), and (iii) 10 subjects (5 male and 5 female). Each subject participated in one experiment. 106 materials were presented, as mentioned above. However, because we are not accustomed to hearing only prosody, in (i), the subjects did prior studies. Each subjects judged the "change" or "continue" of turn and checked the questionnaire forms after veiwing the materials.

Fig. 2 shows the correct rate of each subjects and the Cohen's κ in each condition was follows : (i) 0.072, (ii) 0.329, and (iii) 0.471. Under the prosody condition, the correct rates from 3 of 12 subjects was more than 60% and 1 subject correctly judged at an accuracy of more than 80%. And the possibility expected by chance (the sum of the joint probability of the materials are classified into the one class and the materials join that class when both are independent on each other) was 0.497. The correct rates of 6 subjects exceeded the probability expected by chance.

Table 3, 4 show the detailed the results from (ii) and (iii). According to these results, the subjects were apt to make mistakes when judging the changes of turns based on only the linguistic information. On the other hand, subjects succeeded in judging the changes of turns based on speech sounds. We mean

that the non-linguistic information—prosody enabled subjects to judge changes of turns.

These results indicate the possibility that prosody contains some information that can enable us to predict turn transitions.

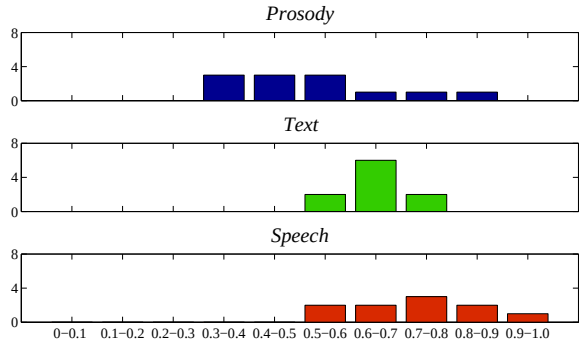

Figure 2: *The results from cognitive experiments on turn-taking. The horizontal axis means the correct rate and the vertical axis does the number of subjects.*

Table 3: *The results from (ii) the text condition.*

Correct class	Discriminant class	
	Change	Continue
Change	259	290
Continue	72	448

Table 4: *The results from (iii) the speech condition.*

Correct class	Discriminant class	
	Change	Continue
Change	367	169
Continue	107	399

5.3. Discussion

From both the physical and cognitive perspectives, we examined whether or not the F_0 patterns forenotice and enable us to predict turn transitions.

Our physical experiments indicated that the F_0 patterns contain information forenoticing turn transitions at an the accuracy of about 70%.

As for cognitive experiments, in the (i) prosody condition presented materials that contained only prosody, but didn't provide phonological information, were not usually familiar to us. That is why prior studies were necessary under this condition. Therefore, there is a possibility that only a few subjects learned and that caused the lowering of the correct rate in (i). However it is important that a few subjects could predict the turn transitions based on only prosody. In addition, a comparison between the results from the (ii) text condition and (iii) speech condition indirectly indicates the contributions of prosody when judging turn transitions.

In these experiments we examined the possibility of predictions being made about turn transitions by not presenting the following utterance. However subjects head preceding utterances from the start to the end and we could not conduct enough examinations to explain the turn transitions with only a short overlap. When taking into account the difficulties in a prior study, other ways may have been better. In particular, one way would be by comparing the results from (iii) with one from experiments in which we presented materials that didn't contain prosody information by using the gate method.

6. Conclusions

We assumed that speech, especially prosody, contains forenoticing and enables us to predict segment boundaries, syntactic structures, and turn transitions, and that these forenotices and predictions help to achieve smoother real-time conversations. In this study, we examined the possibility of these hypotheses by using the F_0 model.

As a result, we indicated that forenoticing information concerning the segment boundaries, syntactic structures, and turn transitions existed in F_0 patterns.

On the other hand, for cognitive predictions, although we showed the possibility of predicting about turn transitions, the predictions concerning segment boundaries and syntactic structure are further task for use to consider.

7. Acknowledgements

For these examinations, we express gratitude to Toshie Hatano for materials of the experiments on segment boundaries and to Minori Chida for assist the experiments on turn-taking.

8. References

[1] T. Hatano, "Prosody Based Speech Segmentation," *Proceedings of the 5th International Conference of the Cognitive Science (ICCS)*, 2006.

[2] T. Ohsuga, M. Nishida, Y. Horiuchi and A. Ichikawa, "Estimating Syntactic Structure from Prosodic Feature in Japanese Speech," *Proceedings of ICSLP2004*, pp.977–980, October. 2004.

[3] T. Ohsuga, M. Nishida, Y. Horiuchi and A. Ichikawa, "Investigation of the Relationship between Turn-taking and Prosodic Features in Spontaneous Dialogue," *Proceedings of Eurospeech*, pp.33–36, September. 2005.

[4] T. Kimura, Y. Horiuchi, M. Nishida and A. Ichikawa, "Analysis of Turn-taking and Prosody in Japanese dialogue using F_0 model [in Japanese]," *IEICE technical report. Speech*, vol.107, no.282, pp.25–30, October. 2007.

[5] H. Fujisaki, and K. Hirose, "Analysis of voice fundamental frequency contours for declarative sentences of Japanese," *Journal of the Acoustical Society of Japan (E)*, vol.5, no.4, pp.232–242, October. 1984.

[6] Y. Hirose, "Speaker's Intention and Hearer's Comprehension : A Latent Function of Lexical Accent in Syntax [in Japanese]," *Cognitive studies*, vol.13, no.3, pp.428–442, September. 2006.

[7] A. Komatsu, E. Oohira, and A. Ichikawa, " Prosodical Sentence Structure Inference for Natural Conversational Speech Understanding, "*Proceedings of Eurospeech*, pp.400-403, 1989.

[8] M. Abe, "Manual of Japanese Speech Database for Research (Continuous Speech Data) [in Japanese]," *Technical Report TR-I-0166*, 1990.

[9] H. Tanaka, Turn-Taking in Japanese Conversation, John Benjamins Pub co, 1999.

[10] K. Takanashi, "An exploration of mechanism of hearers' incremental prediction of on-going sentences [in Japanese]," Sentence and Utterance in Real Time, S. Kushida, T. Sadanobu and Y. Den (eds.), pp.159–202, Hituzi Syobo Pub co, Tokyo, 2007.

[11] Y. Horiuchi, A. Yoshino, M. Naka, S. Tutiya, A. Ichikawa, "The Chiba Map Task Dialogue Corpus Project [in Japanese]," *Research reports of Faculty of Technology, Chiba University*, vol.48, no.2, pp.33–60, March. 1997.