6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

A Proposal of a Model to Extract Japanese Voluntary Speech Rate Control

Keiichi TAKAMARU, Makoto HIROSHIGE, Kenji ARAKI and Koji TOCHINAI

Graduate School of Engineering, Hokkaido University
N13W8, Kita-ku, Sapporo, Hokkaido, 060-8628, Japan
Tel: +81-11-706-7389 Fax: +81-11-709-6277 E-mail: takamaru@media.eng.hokudai.ac.jp

ABSTRACT

To extract elements of prosodic features which relate to speakers' intentional control is required for speech information processing. Speech rate variation should be a "caution signal" to call listeners' attention strongly. To express and detect such "caution signals", we have proposed a new speech rate model. This model introduces two kinds of force to control the speech rate. One is a driving force which causes a global tendency of speech rate variation, and the other is damming force with speaker's voluntary control. The proposed model have been applied to several spontaneous conversational speech which we have newly recorded. A global tendency of speech rate has been expressed appropriately. In several cases, the intentional rate variations have been expressed appropriately.

1 INTRODUCTION

Recognition and understanding of the paralinguistic information[1] are the important subjects for speech information processing. Paralinguistic information is usually expressed with prosody (fundamental frequency, power, speech rate and so on). Not all prosodic features correspond to paralinguistic information, so extracting elements of prosodic features which relate to speakers' intentional control is required.

Speech rate variations which are occurred with prominence are fewer than other prosodic features (fundamental frequency and power)[2]. Because of this rareness, however, the speech rate variation should be considered as a strong "caution signal" to call listeners' attention, so that such strong "caution signal" should not be missed even in man-machine communication. In our usual conversation, local speech rate variations with speakers' intentional control are actually observed.

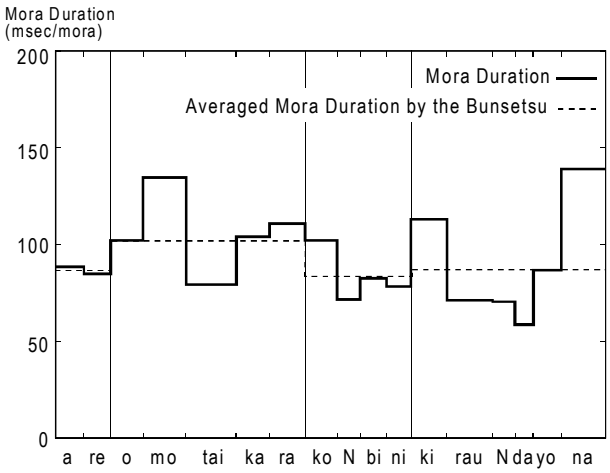

Figure 1: An example of variations of mora duration in Japanese (The meaning of utterance is "It is heavy, so convenience stores hate it.")

From these points of view, we are aiming to detect the portions of intentional local speech rate variations in Japanese spontaneous conversational speech. For this purpose, we propose a new speech rate model to express two forces which control speech rate. In our study, Japanese spontaneous conversational speech which may contains rich prosodic features are recorded. Then mora boundaries in each speech are decided to calculate speech rate. Next, "mora duration adjusting factor (MDAF)" is applied to reduce the influence of extreme shortening morae on the model. Then we apply the proposed speech rate model to spontaneous conversational speech. Our new speech rate model extracts two forces separately. One is a driving force which causes a global tendency of speech rate variation, and the other is a damming force causes non-linear transformation related with speaker's voluntary control.

There are several methods which express a speech rate[3][4]. However they may not be suited for our purpose of detecting intentional rate variation: The method using DTW[4] needs reference speech so they are difficult to use for large amount of spontaneous conversational speech. The methods to decide precise mora duration for speech synthesis may not express speakers' intentional rate control.

In section 2, the expressions and the features of Japanese local speech rate variation are described briefly. Then we will introduce the MDAF. In section 3, the details of the proposed speech rate model are described. In section 4, we apply the proposed speech rate model to several spontaneous conversational speech.

2 DESCRIPTION OF RATE VARIATION OF FREE CONVERSATIONS

2.1 Expression of the Local Speech Rate

To express local speech rate variation, mora duration (sec/mora) is used in our study. To obtain duration of each morae in speech, mora segmentation is needed. In our case of free conversational speech, mora boundaries often disappear because of heavy coarticurations. Thus, the mora segmentations are carried out by human experts.

We employ a *bunsetsu* as a fundamental temporal period(unit) to design our speech rate model. The *bunsetsu* is a semantic unit into which a sentence can be divided naturally in terms of meaning and pronunciation in Japanese[8]. Since the *bunsetsu* is a semantic unit, we can naturally consider that speaker's intentional control can be issued on the *bunsetsu*. An example of variation of mora duration in Japanese spontaneous conversational speech is shown in Figure 1. In figure 1, averaged values of mora duration for each bunsetsu are also shown. These averaged values indicate rough variation of local speech rate in the *bunsetsu*-wise resolution.

2.2 Observed Features of Speech Rate Variation

As a preparation to design a speech rate model, we have investigated the features of local speech rate variation in spontaneous conversational speech. The following features have been observed.

2.2.1 Local Speech Rate Variations by *Bunsetsu*s

We can find that the local speech rate becomes observably slower in several portions of speech. In most such cases, the factor of becoming slower is speakers' thinking. But, in several cases we surely observed slower portions with intentional emphasis.

2.2.2 Variations of Each Mora Duration

Observing the variations of each morae duration, several features such as phrase final lengthening or bimoraic foot are also observed in our data. Especially, shortening of several kinds of mora (moraic nasal, long vowels and double consonants) are observed remarkably in our spontaneous conversational speech. These large variations of mora duration are considered to be involuntary variation based on phonemic nature. It is better to remove the influence of these involuntary phenomena.

2.3 MDAF:Mora Duration Adjusting Factor

The shortening of specific morae is one of involuntary elements of the observed speech rate variation. Before applying our speech rate model, the following "mora duration adjusting factors (MDAF)" are applied to moderate the involuntary turbulence of mora duration. The MDAF is defined as follows:

$$\text{MDAF} = quantize\left[average\left(\frac{MD}{\overline{MD}_{bunsetsu}} \right) \right]$$

where MD is duration of shortening mora and $\overline{MD}_{bunsetsu}$ is averaged mora duration by the *bunsetsu*. The MDAF is applied to the observed mora duration as follows:

$$\text{adjusted mora duration} = \frac{\text{observed mora duration}}{\text{MDAF}}$$

In this study, MDAF are applied to double consonant, moraic nasal, long vowels, diphthong and several devoiced CV morae. Values of MDAF are shown in Table 1. They are decided based on the investigation of duration ratio of specific morae in our study[5], and knowledge on mora duration from several researches[6][7]. Figure 2 shows an example of variation of adjusted mora duration after applying the MDAF. Mora duration adjusted by this process is used in the following new speech rate model proposed in this report.

Table 1: Values of MDAF

Kind of Mora	The Number of Mora	MDAF
Double Consonant	1	**0.75**
Moraic Nasal	1	**0.75**
Long Vowel	2	**1.25**
Diphthong (including /i/)	2	**1.5**
Diphthong (not including /i/)	2	**1.75**
Devoiced /kV/ /fV/	1	**0.75**

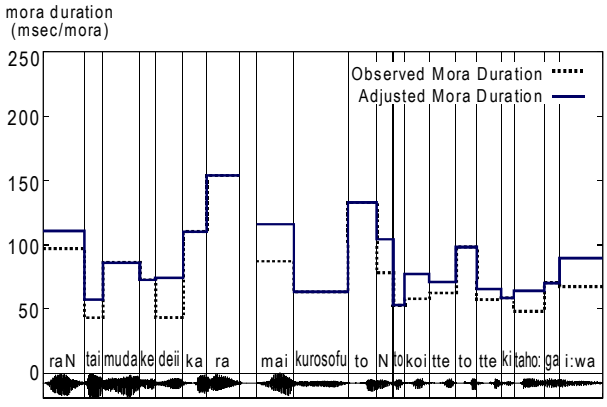

Figure 2: An example of adjusted mora duration (The meaning of utterance is "You need only run-time (library), so you should get it from Microsoft's (site)")

3 MODELING OF THE LOCAL SPEECH RATE VARIATION

3.1 The Concept and Outline of the Proposed Model

Speech rate which we can observe is considered as mixture of variations which are caused by various factors. Those factors can be roughly divided into voluntary elements and involuntary elements. Involuntary rate variations are caused by phonemic nature of the language, whereas voluntary rate variations can be related with speaker's intention. Thus, in our procedure of the modeling, voluntary elements should be preserved and involuntary elements are discarded as noise.

On making a speech rate model, we try to extract two kinds of force (Figure 3) which effect the local speech rate variation separately. They are as follows:

1) a driving force of speech flow
2) a strong force that temporally dams up the flow

The former "driving force" continues throughout a sentence and varies slowly. The slow variation of the force is usually almost involuntary. This driving force causes a global tendency of speech rate variation. As a global tendency of speech rate, averaged speech rate and tendency of increasing or decreasing of its rate should be expressed. So a linear function is applied to variation of mora duration in each semantic unit.

The latter "damming force" is strong, voluntary control, that is main target in our study. We consider that this damming control is applied to a minimum semantic unit (*e.g.*, word), and make local speech rate slower. When the speech rate is made slower voluntarily, morae in the controlled unit may not lengthen uniformly, that is, nonlinear transformation. Mora duration would be lengthened around the most elastic portion in the controlled unit, which seem to depend upon the phoneme in the unit. A simple convex curve should be applied to each controlled unit. We try to use a cosine curve to express this

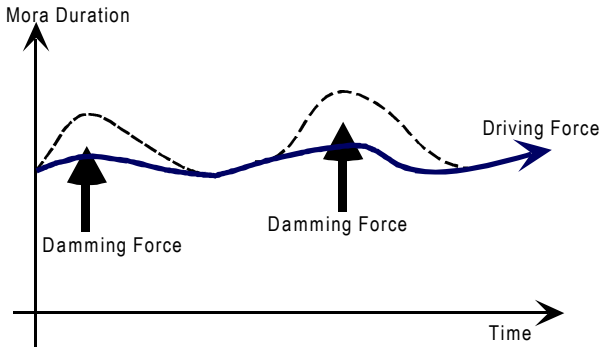

Figure 3: Two forces which affect speech rate variation

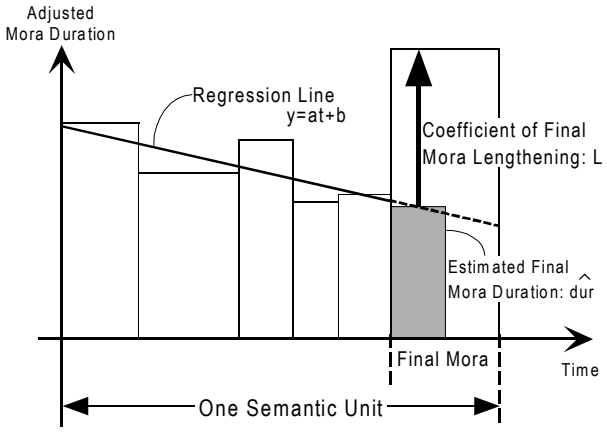

Figure 4: A regression line which expresses a global tendency

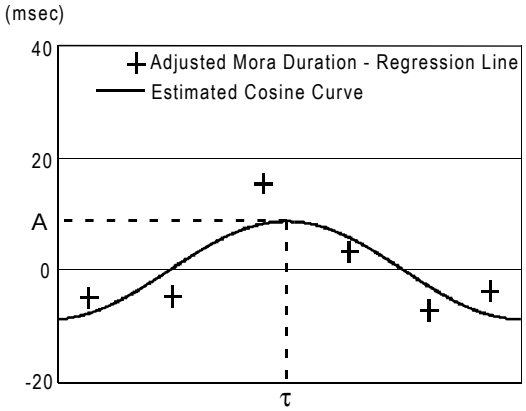

Figure 5: A cosine curve which express "damming control"

damming control.

3.2 Approximation of Global Tendency

To express global tendency of speech rate, straight lines ($y=at+b$) are fitted to the variation curve of observed mora duration (Figure 4). A single line is fitted as a regression line to a single semantic unit "*bunsetsu*".

When the inclination parameter a is negative, the unit has accelerative tendency in speech rate. when a is positive, the unit has decelerative tendency in speech rate.

The final mora in the unit can be lengthened regardless of global tendency of speech rate. This is a well known phenomenon as the "final mora lengthening". At first the final mora of the unit is excluded from the calculation of the regression line. Then duration of final mora is estimated by the regression line. Comparing estimated duration to actual duration, if final mora does not be lengthened, the coefficients of regressive line are calculated again including final mora.

3.3 Approximation of Non-linear Transformation

To express damming force that causes a non-linear transformation of temporal structure in each semantic unit, we introduce a simple convex curve. Amount of the non-linear transformation can be derived by subtraction of the global tendency from the observed mora duration. In this report, a value estimated by regression line is subtracted from each observed mora duration. The value which is subtracted is decided as a value on the regression line at the central time-point of the corresponding mora. To the obtained set of the subtracted value, a simple convex curve is fitted.

In this report, we use a cosine function as a simple convex curve as follows:

$$\hat{y} = -A\cos\left(\frac{2p}{T+|\Delta T|}(t+j)\right) \qquad \text{where} \quad j = \begin{cases} \Delta T & \text{if } \Delta T \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

T is duration of a semantic unit (or a semantic unit without final mora) and A is an amplitude. The amplitude A and the time of the maximum point t = (T-ΔT)/2 are estimated by means of the least mean squares (Figure 5). The A is considered as the strength of damming force. The t indicates the location of the most strongly lengthened portion within the semantic unit.

4 APPLICATION OF PROPOSED MODEL TO SPONTANEOUS CONVERSATIONAL SPEECH

The speech rate variation model proposed in previous section is applied to 48 sentences of spontaneous conversational speech. Speech samples are uttered by two male speakers who are university students and have been newly recorded for our research. Figure 6 and 7 show examples. Lower graph is "adjusted mora duration" and "regression line". Upper graph is "adjusted mora duration minus regression line" and "estimated cosine curve".

The averaged speech rate and tendency of increase and decreasing of speech rate are appropriately expressed by the regression line in most cases. In a lot of cases, a regression line falls. This means the semantic unit has accelerative tendency. And in several cases, cosine curves appropriately express intentional rate variations. The t are located to the middle of the semantic unit in those cases. In other several cases, however, we can find the amplitude A are almost 0 (*i.e.*, no damming force) although there seem to exist nonlinear temporal transformation in the unit. In such cases, more precise investigations are needed. Figure 8 shows the distribution of A and t of all units in samples. The t often takes the value near 0 (beginning of the unit) or 1 (end point of the unit) in a lot of unit.

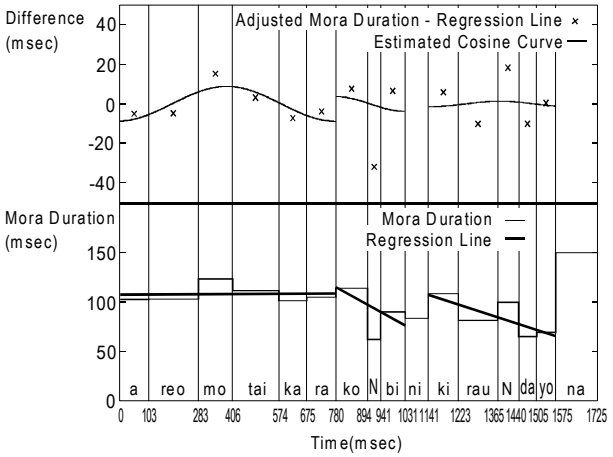

Figure 6: An example of an application of the proposed speech rate model (The meaning of utterance is "It is heavy, so convenience stores hate it.")

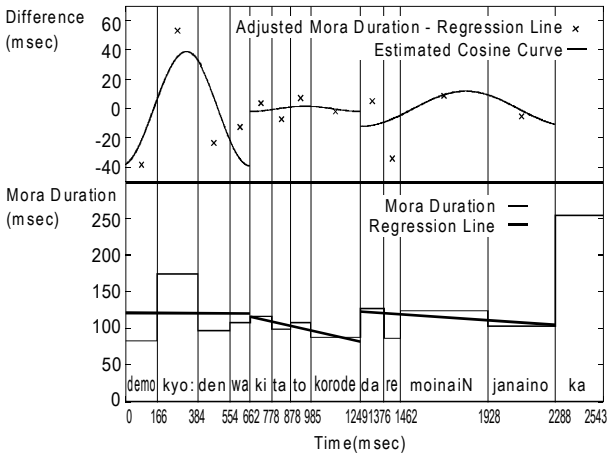

Figure 7: An example of an application of proposed speech rate model (The meaning of utterance is "Phone! But nobody else is here, today?")

When the number of morae is too many in the semantic unit, approximation by the proposed model is apt not to be carried out well. Because increasing the number of morae in the unit means increasing of the noise by phonemic nature. On the other hand, when the number of morae is too small, approximation is also difficult because approximation is apt to be affected by each phonemic nature strongly.

To express voluntary rate control more appropriately, we should consider the improvements of 1) the shape of a curve to use in the approximation 2) the length of the semantic unit to apply the model.

5 CONCLUSIONS

We have proposed a new speech rate model which extracts voluntary rate control from observed speech rate variation. The model have been designed to extract "driving force"(global tendency) and "damming force" (voluntary nonlinear transformation) separately. A regression line is applied to express "driving force" and a cosine curve is applied to extract "damming force" in each semantic unit. We have applied

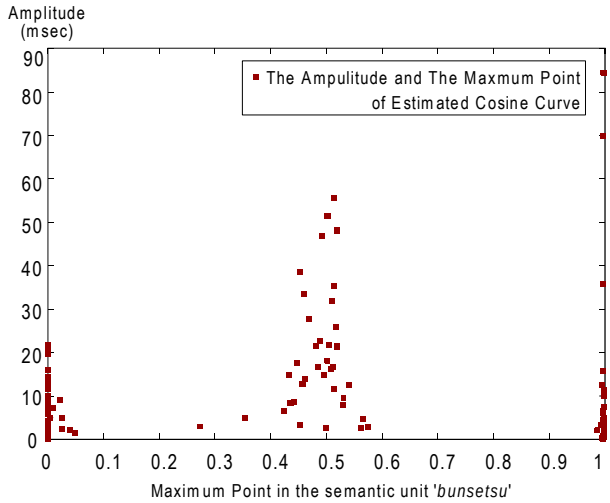

Figure 8: Distribution of A and t of estimated cosine curve

proposed model to several spontaneous conversational speech. Averaged speech rate and tendency of increase and decreasing of speech rate has been expressed by a regression line appropriately. And in several cases, cosine curves appropriately express intentional rate variations.

We need to consider that the validity of using the cosine curve and the validity of applying the model to a semantic unit *bunsetsu*. And we have to investigate relationship between the perception of speech rate variations and the coefficients of proposed model.

REFERENCE

[1] H.Fujisaki: "Prosody, Models, and Spontaneous Speech", In Y.Sagisaka *et al.*(ed.) Computing Prosody, Springer, pp.27-42 (1997)

[2] S.Kobayasi, S.Kitazawa: "Factors concerning paralinguistic feature identification in natural dialogue", Technical Report of IEICE of Japan (1998)

[3] K.Hirose, H.Kawanami: "On the relationship of speech rates with prosodic units in dialogue speech", Proc. ICSLP '98 (1998)

[4] S.Ohno, H.Fujisaki: "Quantitative analysis of the local speech rate and its application to speech synthesis", Proc. ICSLP '96, Vol.3 , pp.2254-2257 (1996)

[5] K.Takamaru, K.Suzuki, M.Hiroshige, K.Tochinai: "A study on several features of local speech rate variations in spontaneous conversational speech -detailed investigations about shortening/lengthening of specific mora duration-" Proc. ITC-CSCC '99, pp.390-393 (1999-07)

[6] Y.Sagisaka, Y.Tohkura: "Phoneme duration control for speech synthesis by rule", Trans. IEICE of Japan, J67-A(7), pp.629-636 (1984)

[7] H.Kawasaki: "Models and data on the temporal regulation of speech: Isochrony in Japanese and English", J. Acoust. Soc. Japan, Vol.39, No.6, pp.389-397 (1983)

[8] M.Nagao: "Nihongo joho-shori", Korona-sha (1984)