ANALYSIS OF EFFECTS OF LEXICAL ACCENT, SYNTAX, AND GLOBAL SPEECH RATE UPON THE LOCAL SPEECH RATE

Sumio Ohno *Hiroya Fujisaki* *Hideyuki Taguchi*

Science University of Tokyo
2641 Yamazaki, Noda, 278-8510, Japan

ABSTRACT

The speech rate is one of the important prosodic parameters for the naturalness and intelligibility of an utterance. On the basis of the authors' definition of the relative local speech rate, the present paper describes an analysis of the effects of changes in global speech rate, syntactic constituency and lexical accent on the local speech rate, using short utterances in which these factors are systematically controlled. Preliminary results indicate that the span of changes in local speech rate is the syllable rather than *mora*, and also shows the interaction between these factors.

1. INTRODUCTION

In natural speech, the local speech rate is known to vary due to various factors such as lexical accent/stress, syntactic boundary, speaking style (especially the global speech rate), etc., though the magnitude of their effects may differ from one language to another. The synthesis of speech with high degree of naturalness and expressiveness requires an appropriate control of the local speech rate. In order to obtain rules for speech rate control, however, we need to have a clear, quantitative definition of the local speech rate and an objective method for its measurement. While the global average speech rate of an utterance or a set of utterances can be defined by the number of phonetic units such as syllables or *morae* uttered per unit time, the local speech rate has not been clearly defined.

Conventional methods for measuring the local speech rate require determination of specific time instants on the speech waveform or a certain acoustic-phonetic feature such as the short-time frequency spectrum as a function of time. Thus most of the studies on the local speech rate rely on measurements of segmental durations, usually obtained by visual inspection of the speech waveform and/or the frequency spectrum. In many cases, however, segmental boundaries are not well defined nor can be measured objectively. These difficulties cannot be avoided when one tries to measure the absolute local speech rate, but can be circumvented if we try to measure the relative local speech rate, i.e., the local speech rate of a given utterance relative to that of the corresponding portion of a reference utterance with the same linguistic content.

Based on these considerations, we have proposed a new method for quantifying the temporal changes in speech rate of a target utterance relative to another utterance chosen as the reference, and have demonstrated its usefulness in studying the effects of various factors upon the local speech rate. In the present paper, we will first describe the method briefly, and then present some recent results obtained on the effects of lexical accent, syntax and global speech rate on the local speech rate in spoken Japanese.

2. RELATIVE LOCAL SPEECH RATE

2.1. Definition

Provided that we have a way to define a time-axis warping function that maps a given utterance (i.e., the target) onto another utterance (i.e., the reference) of the same linguistic content based on the local similarity of the two utterances, we can define a relative local speech rate without resorting to segmental boundaries. Denoting by $W(t)$ the time-axis warping function where t indicates the time variable of the reference utterance, the relative speech rate of the target relative to the reference can be defined by

$$R(t) = 1 \Big/ \frac{dW(t)}{dt} \ . \tag{1}$$

Since a short-time averaging process is always involved in calculating the local similarity, the above definition should be interpreted as giving the *relative short-time average speech rate at* t, though it can be defined at any given instant t. For the sake of brevity, however, $R(t)$ will be referred to simply as the relative speech rate at t.

2.2. Calculation of Relative Speech Rate

The alignment of the time axis of the target utterance against that of the reference utterance is conducted by a dynamic time-axis warping (DTW) procedure in the 12-dimensional parametric space of FFT cepstrum coefficients. The DTW procedure establishes a one-to-one correspondence between

a sequence of points, represented by t_n $(n = 1 \sim N)$, on the time axis of the reference utterance and the corresponding time points, represented by t'_n $(n = 1 \sim N)$, on the time axis of the target utterance. This correspondence serves as an approximation to the continuous time-axis warping function $W(t)$. By introducing a window function $w(t)$, the relative local speech rate $R(t)$ at any given time instant t can be approximated by the reciprocal of the slope of the weighted regression line as

$$\widetilde{R}(t) = \frac{\sum w_n \cdot \sum w_n {t_n}^2 - (\sum w_n t_n)^2}{\sum w_n \sum w_n t_n t'_n - \sum w_n t_n \sum w_n t'_n}, \quad (2)$$

where $w_n = w(t - t_n)$. In the current analysis, a triangular window of width T is adopted as $w(t)$. The optimum value for the window width T was found to be 270 ms on the basis of perceptual evaluation of naturalness of analysis-resynthesis.

3. EXPERIMENT

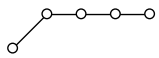
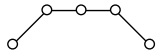
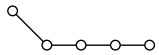
3.1. Speech Material

The text for the speech material consisted of nine short sentences of 10 *morae* each. Each sentence consists of two '*bunsetsu*' phrases. A '*bunsetsu*' is a syntactic unit of Japanese consisting of a content word with or without subsequent function word(s). In the following, we will use the word 'phrase' to mean '*bunsetsu*.' Each test sentence consists of two phrases (henceforth P_1 and P_2). The length of both P_1 and P_2 are 5 *morae*. For both P_1 and P_2, three phrases that are segmentally almost the same but differ in the lexical accent type are selected. The accent types adopted here are:

T_1: Unaccented,
T_2: Accented on the 4th *mora*,
T_3: Accented on the initial *mora*.

Table 1 shows the schematic patterns of subjective pitch of these three accent types, where each circle (○) indicates the relative subjective pitch.

Table 1: Three accent types adopted for the phrases.

Accent Type		Subjective Pitch Pattern
T_1	Unaccented	
T_2	Accented on the 4th *mora*	
T_3	Accented on the initial *mora*	

As for P_2, the following three are selected:

$P_2(T_1)$	*mondaida*	'(is) problem'
$P_2(T_2)$	*monjinda*	'(is) follower'
$P_2(T_3)$	*monbanda*	'(is) janitor'

As for P_1, phrases with maximally similar segmental constituents are selected to constitute meaningful sentences when combined with these P_2s. They are:

$P_1(T_1)$	*nidaimega*
$P_1(T_2)$	*nidaimeno*
$P_1(T_3)$	*nidaimono*

It is to be noted that a P_1 may represent different lexical items when it is combined with different P_2s. Namely, *nidaime* stands for 'the second problem' when followed by *mondai* (problem), but stands for 'the second generation' when followed by *monjin* (follower) or *monban* (janitor). Thus the following set of nine sentences are constructed as shown in Table 2.

Table 2: Nine sentences adopted for the experiment.

No.	Accent Types	Sentence		
S_{11}	$T_1 - T_1$	*Nidaimeno*	*mondaida.*	'It is the second problem.'
S_{21}	$T_2 - T_1$	*Nidaimega*	*mondaida.*	'The problem is the second one.'
S_{31}	$T_3 - T_1$	*Nidaimono*	*mondaida.*	'There are as many as two problems.'
S_{12}	$T_1 - T_2$	*Nidaimeno*	*monjinda.*	'He is the second-generation follower.'
S_{22}	$T_2 - T_2$	*Nidaimega*	*monjinda.*	'The second generation is a follower.'
S_{32}	$T_3 - T_2$	*Nidaimono*	*monjinda.*	'They are followers over two generations.'
S_{13}	$T_1 - T_3$	*Nidaimeno*	*monbanda.*	'He is the second-generation janitor.'
S_{23}	$T_2 - T_3$	*Nidaimega*	*monbanda.*	'The second generation is a janitor.'
S_{33}	$T_3 - T_3$	*Nidaimono*	*monbanda.*	'They are janitors over two generations.'

The speaker was a native male speaker of the common Japanese. The text was read eight times at five global speech rates: very fast (VF), fast (F), normal (N), slow (S), and very slow (VS), corresponding to the average speech rates of 9.5, 8.5, 7.7, 6.8, and 6.1 *morae* per second, respectively.

All the utterances were recorded first on a DAT tape recorder and were subsequently stored in a digital computer at 10 kHz with 16 bit precision. Each utterance was then analyzed at 10 ms intervals using a 25.6 ms Hamming window to obtain 12 FFT cepstrum coefficients for the calculation of the relative local speech rate.

3.2. Results of Analysis

(1) Effects of Global Speech Rate

Figure 1 illustrates the results of analysis of the local speech rate for the sentence S_{12} (*Nidaimemo monjinda.*). The four curves represent the results of VF-, F-, S-, and VS-utterances, each representing the average of 8 utterances, respectively. The results are first obtained by using one of the N-utterances as the reference, then modified in order to have the average of the eight N-utterances as the virtual reference. The abscissa represents the time axis of the initial reference utterance, and the vertical broken lines indicate approximate segmental boundaries obtained by visual inspection of the waveform of the N-utterance used as the initial reference.

The measured local speech rate displays a maximum at each syllable except at phrase-final particles and/or auxiliaries, and a minimum between two adjacent maxima. The location of the first maximum occurs at or very near the word onset, while other maxima occur more or less in the middle of a syllable or a syllable pair. The locations of these maxima and minima are not much affected by the global speech rate and are almost the same for all the four curves. The figure indicates that the acceleration at faster utterances is most prominent at the middle of a syllable (or a pair of syllables when the second syllable is a particle/auxiliary), while deceleration in slower utterance is most prominent toward the end of a syllable (or a pair of syllables). It is to be noted that the span of these acceleration/deceleration is based on syllable rather than *mora*, which is a unit of pitch accent assignment. Thus '*-jin*' in '*monjin*' occupies two *morae* and the accent (*i.e.*, downfall in pitch) occurs at the end of the *mora* '*ji*', but shows only one maximum of local speech rate since '*-jin*' constitutes a single syllable.

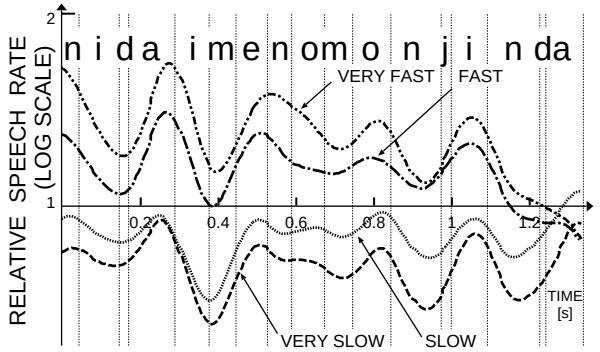


Figure 1: The results of analysis of the local speech rate for the sentence S_{12} (*Nidaimemo monjinda.*).

The distance between a minimum and the following maximum is approximately 100 msec on the time axis of the reference utterance. The figure also indicates that the effect of the global speech rate diminishes at the end of the utterance.

(2) Effects of Accent Type and Syntax

Figure 2 shows the results of analysis for the three sentences: (a) S_{11}, (b) S_{21}, and (c) S_{31} where P_2 is of the unaccented type and P_1 appears in three accent types. Only the results for the F-utterances against the N-utterances as

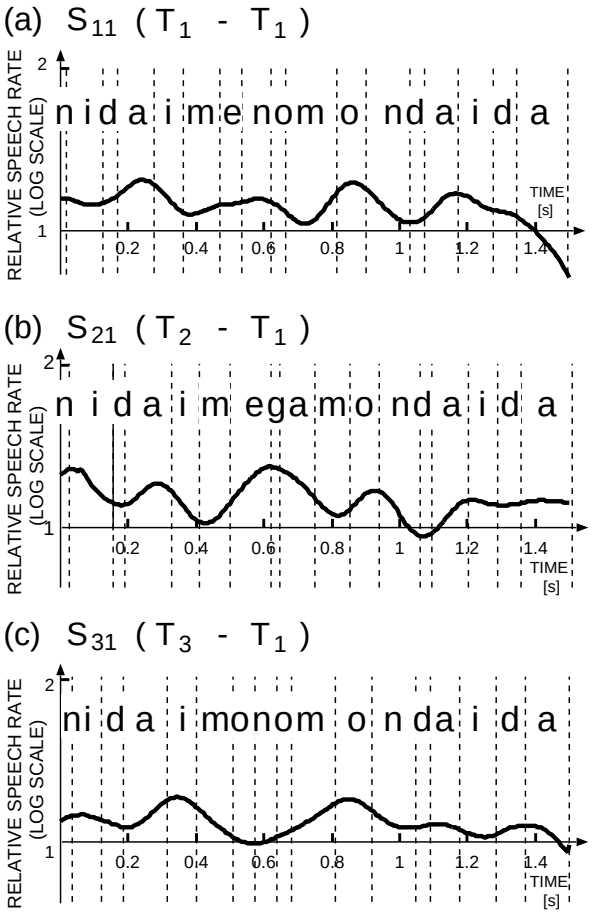


Figure 2: The results of analysis for the three sentences: (a) S_{11}, (b) S_{21}, and (c) S_{31} where P_2 is of the unaccented type and P_1 appears in three accent types.

reference are shown for the sake of simplicity. Panel (b), where the accent occurs at the 4th *mora*, shows the most prominent maximum in local speech rate at the end of the 4th *mora* (*i.e.*, in the middle of the syllable pair '*mega*', where '*ga*' is an unaccented particle), while panel (c), where the accent occurs at the initial *mora*, shows no maximum but a minimum at the corresponding location. In this case, both '*mo*' and '*no*' are particles, while in panels (a) and (b) '*me*' belongs to the content word and '*no*' is the only particle.

Figure 3 shows further results of analysis for the three sentences: (a) S_{31}, (b) S_{32}, and (c) S_{33}, where P_2 is the accented type with accent on the initial *mora* (*i.e.*, the 6th *mora* of the utterance) and P_1 appears in three accent types. In panel (a) where the first phrase is of the unaccented type, the locations of the peaks of the local speech rate are similar to those in Fig. 2. In panels (b) and (c) where P_1 is of the accented type, on the other hand, the effect of the accent position of P_2, *i.e.*, at the initial *mora* of P_2, tends to

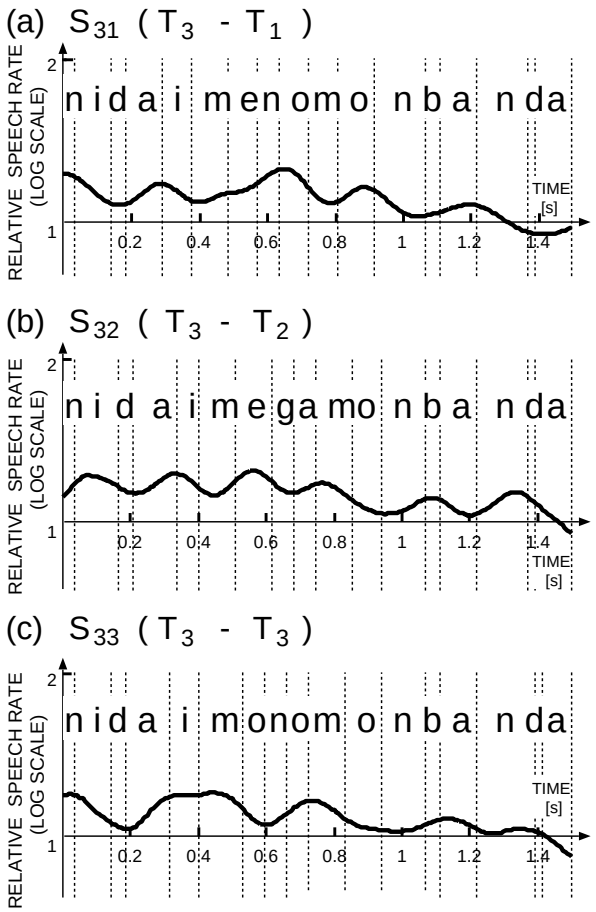


Figure 3: The results of analysis for the three sentences: (a) S_{31}, (b) S_{32}, and (c) S_{33}, where P_2 is the accented type with accent on the initial *mora* (*i.e.*, the 6th *mora* of the utterance) and P_1 appears in three accent types.

shift the location of the maxima of P_2 toward the beginning of each syllable, and gives rise to another maximum near the onset of the final mora '*da*,' which is an auxiliary and does not show a clear maximum in most of the other cases.

4. DISCUSSION AND SUMMARY

Although the results shown in this paper are still limited, they already indicate the existence of interesting interactions between the syntactic constituents and lexical accents in their effects on the local speech rate.

In addition to the factors whose effects are analyzed and briefly described here, other factors, such as discourse focus, speaking style, emotion, etc., are also known to affect the patterns of local speech rate. Work is in progress to analyze and quantify their individual influences on the local speech rate as well as their interactions in actual speech on the one hand, and to investigate their perceptual significance in synthetic speech on the other.

5. REFERENCES

1. Ohno, S. and Fujisaki, H. "A method for quantitative analysis of the local speech rate," *Proceedings of the 4th European Conference on Speech Communication and Technology*, Vol. 1, pp. 421–424, 1995.
2. Ohno, S., Fukumiya, M. and Fujisaki, H. "Quantitative analysis of the local speech rate and its application to speech synthesis," *Proceedings of the 1996 International Conference on Spoken Language Processing*, Vol. 3, pp. 2254–2257, 1996.
3. Ohno, S. and Fujisaki, H. "Analysis of the relative speech rate and its application to speech synthesis," *Proceedings of the First China-Japan Workshop on Spoken Language Processing*, pp. 85–90, 1997.
4. Ohno, S., Fujisaki, H. and Taguchi, H. "A method for analysis of the local speech rate using an inventory of reference units," *Proceedings of the 5th European Conference on Speech Communication and Technology*, Vol. 1, pp. 461–464, 1997.