QUANTITATIVE ANALYSIS OF THE LOCAL SPEECH RATE AND ITS APPLICATION TO SPEECH SYNTHESIS

Sumio Ohno, Masamichi Fukumiya and Hiroya Fujisaki

Department of Applied Electronics, Science University of Tokyo
2641 Yamazaki, Noda, 278 Japan

ABSTRACT

On the basis of the short-time relative speech rate defined by the authors, this paper examines the optimum width of the smoothing window by perceptual experiments on the naturalness of re-synthesized speech. With the optimum window of 270 ms, relative speech rates are obtained both for 'fast' and 'slow' utterances of the same sentence, using an utterance produced at a 'normal' speech rate. The averaged results show that the speech rate control function for an utterance can be approximately decomposed into a global component for each sentence and local components for each *bunsetsu* and each major syntactic boundary. Based on these results, a scheme is presented for controlling the local speech rate of a reference utterance to obtain a synthetic utterance of an arbitrary global speech rate.

1. INTRODUCTION

The timing of speech within a discourse is known to vary both globally and locally due to various factors such as lexical or contrastive stress [1], syntactic boundary [2], emotion [3], etc., though the magnitude of their effects may differ from one language to another [4]. Appropriate control of timing is therefore essential for the synthesis of speech with high degrees of naturalness and expressiveness. Since a discourse consists of utterances and pauses, it is necessary to establish measures for expressing their timing in quantitative terms. The timing of an utterance can be expressed by the speech rate, while the timing of a pause can be expressed by its duration. The present paper is concerned only with the speech rate of an utterance.

In order to construct rules for speech rate control in synthesizing utterances, one needs to have a clear definition of the speech rate and an objective method for its measurement. While the global speech rate of a whole utterance 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 [5]. In the present paper, we will describe the result of a perceptual study for finding the optimum width of the time window for calculating the local speech rate, and will also present a method for controlling the local speech rate in speech synthesis.

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 and phonetic content based on the local similarity of the two utterances, we can define a relative local speech rate without resorting to segmental boundaries [6]. 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

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

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.

The advantages of the proposed definition are:

(1) it does not require detection of segmental boundaries,
(2) it can be defined at any given instant along the time axis.

2.2. Calculation of Relative Speech Rate

Although in the current study we have chosen 12 FFT cepstrum coefficients calculated at 10 ms intervals using a 25.6 ms Hamming window, the proposed definition can be applied to any parametric representation of the speech signal.

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 coefficients.

In order to cope with large differences in speech rate to be found in a wide variety of speaking styles, the DTW procedure has the following features:

(1) *Separate processing of silent intervals within each utterance:* Since the presence/absence and duration of silent intervals (i.e., pauses) within an utterance can vary considerably depending on the speaking style, these intervals are detected on the basis of the short-time average power and are treated separately.

(2) *Use of a wider range of slope limitation for time-warping of non-silent intervals:* Instead of the range [1/2, 2] conventionally used in DTW for automatic recognition of spoken words, a wider range [1/3, 3] is adopted to allow for greater differences in speech rate of non-silent intervals. Figure 1 shows the local DP (dynamic programming) matching paths and the weights allowed in the present method.

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 $W(t)$ mentioned in Eq. (1). Note that the time points $t_1, t_2, \cdots, t_N$ are not necessarily equally spaced. 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

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

where $$w_n = w(t - t_n). \tag{3}$$

In the following analysis, a triangular window, given by Eq. (4), is adopted.

$$w(t) = \begin{cases} 1 - |2t/T| & \text{for} - T/2 \le t \le T/2, \\ 0 & \text{elsewhere.} \end{cases} \tag{4}$$

The window width T can be varied according to the required time resolution.

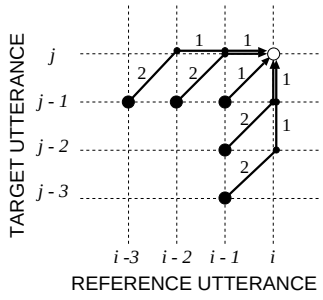

Figure 1: The local DP matching paths and the weights used in the proposed method.

3. PERCEPTUAL OPTIMIZATION OF WINDOW WIDTH

The width T of the window used in obtaining the relative speech rate should be small if we wish to follow the details of changes in the speech rate, but should be large enough to smooth out the granularity of the discrete time-axis warping process. The optimum value of the window width for use in speech synthesis has to be determined by perceptual experiments on the naturalness of synthetic utterances produced by manipulating the local speech rate derived from the analysis of natural utterances.

The speech material for the analysis of the relative speech rate consisted of readings of a sentence in a short story at three global speech rates: 'normal' (7 readings), 'fast' (9 readings), and 'slow' (6 readings). The average speech rate for the three groups of utterances are: 8.0, 9.1, and 6.5 morae per second, respectively. The informant is a male native speaker of the common Japanese. These speech samples are digitized at 10 kHz with 16 bit precision.

The experimental procedure consisted of the following steps:

(1) Selection of the reference utterance.

(2) Selection of the relative speech rate functions for each value of T, and for the 'fast' and 'slow' groups of utterances.

(3) Synthesis of a 'fast' and a 'slow' utterance for each value of T by modifying the local speech rate of the reference utterance.

(4) Perceptual evaluation of naturalness of the synthetic utterances by paired comparison (separate tests for the 'fast' and the 'slow' stimuli).

The time axis of the reference was determined by averaging the time-axis warping functions between all possible pairs of the 7 normal utterances. This was used to produce the reference utterance. The relative speech rate functions were calculated with T values of 40, 80, 160, 320 and 640 ms for each of the 'fast' and 'slow' utterances, and a geometrical mean was calculated for each value of T and for each of the 'fast' and the 'slow' groups. The average compression/expansion factors for the 'fast' and 'slow' groups were 0.81 and 1.16, respectively. These relative speech rate functions were used to generate synthetic 'fast' and 'slow' utterances for each value of T. In addition, two utterances were synthesized by uniformly compressing/expanding the time axis of the reference utterance by the factors shown above.

A paired comparison test was conducted separately for the 'fast' and 'slow' utterances, using all the synthesized utterances as stimuli. All the possible pairs were randomized and presented to the subjects through headphones. The two stimuli in each pair were separated by an interval of 2.5 s, and an interval of 5.0 s was inserted after each pair, during which each subject was asked to record the result of his forced-choice judgment. The subjects were three male speakers of the common Japanese.

The results of the paired comparison tests were processed separately for each subject and for each of the 'fast' and 'slow' groups of stimuli, to obtain the relative positions of each stimulus on a psychological scale of naturalness. Figure 2 shows the averaged results for all the speakers and for both 'fast' and 'slow' groups of stimuli. The data shows a broad peak, and the optimum value of T, obtained by parabolic interpolation, is 270 ms. The effective width of the triangular window ($T/2$) is equal to 135 ms, which is roughly equal to the average duration of one mora in the reference utterance.

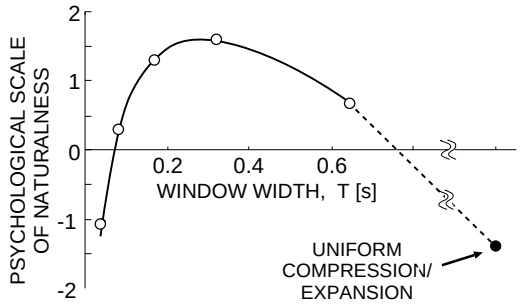

Figure 2: Subjective evaluation of naturalness of time-warped synthetic speech at various values of smoothing window width T.

4. ANALYSIS OF RELATIVE SPEECH RATE

Using the proposed method, relative speech rate was obtained for each of the 'fast' and 'slow' utterances against one of the 'normal' utterances chosen as the reference. Figure 3 illustrates the result of analysis of one of the 'fast' utterance samples. The utterance is a compound sentence consisting of two simple sentences, with a pause in between: "Mukoogishiwa nanimo miezu, marude hirobirotoshita umino yooda." (Nothing can be seen on the other side [of the lake], and it looks just like a vast sea.) The ordinate of the upper panel (a) indicates the time axis of the 'fast' utterance, while the abscissa indicates that of the reference. The speech waveform is displayed along each axis. The time-axis warping function is shown as a piecewise-linear curve in the figure. Since the relative speech rate is defined only for the speech intervals, the panel shows no data for the pause interval between the first and the second sentences of the compound. The lower panel (b) indicates the logarithm of the relative speech rate function smoothed by a triangular window of 270 ms width and plotted against the time axis of the reference.

In order to smooth out sample-to-sample fluctuations and to retain the essential characteristics of speech rate control common to all the samples, the relative speech rate curves were averaged for each of the 'fast' and 'slow' groups, and were corrected for the deviation of the reference sample from the mean speech rate of the 'normal' group. The results are shown in Fig. 4, where the dotted line indicates the average for the 9 'fast' utterance samples and the broken line indicates that of the 6 'slow' samples. The two curves in this figure are almost mirror images of each other. The vertical lines in the figure indicate '*bunsetsu*' boundaries, where a *bunsetsu* is an immediate constituent in the grammar of Japanese, defined as a content word with or without following function words. In general, the absolute values of the deviations from the 'normal' sample show local maxima and minima. Apart from these local maxima and minima, the deviations tend to be larger at the beginning of each of the constituent sentences of the compound sentence.

Assuming exact similarity of the curves of absolute values of logarithm of relative speech rate for both fast and slow utterances, an averaged curve for the absolute value can be derived from these data, and can be used as the prototype of the speech rate control function for the sentence in question. Figure 5 shows such a curve for a global relative relative speech rate of 1.30 (or, equivalently, of 0.77). As a first approximation, the curve can be considered as the sum of two types of components. One type is a slowly-varying component representing the general tendency for each constituent

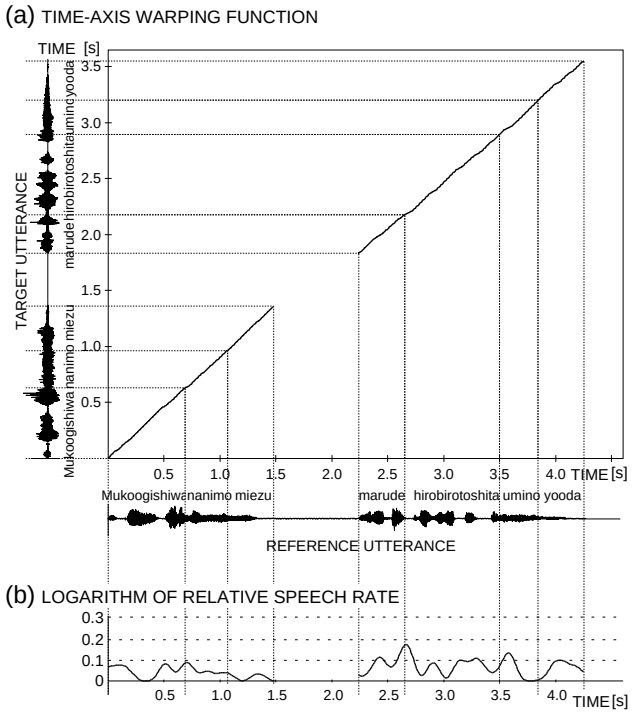

Figure 3: The result of analysis of one of the 'fast' utterance samples.

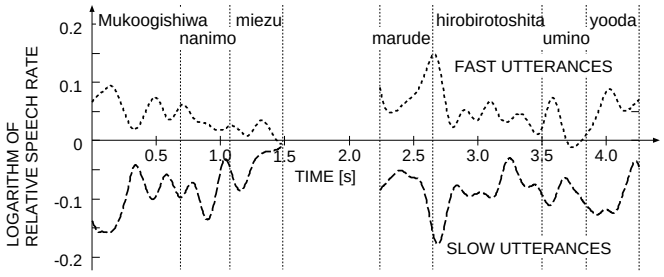

Figure 4: Averaged relative speech rate in logarithmic scale for the 'fast' utterances (dotted lines) and the 'slow' utterances (broken lines).

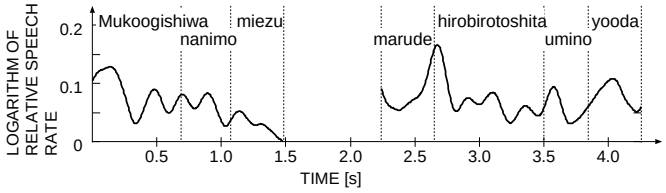

Figure 5: Absolute value of relative speech rate in logarithmic scale for the global relative speech rates of 1.30 and 0.77.

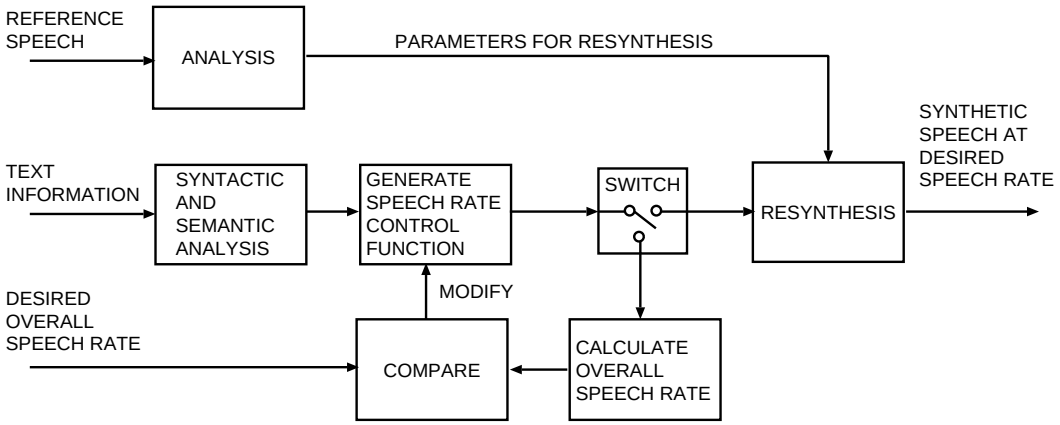

Figure 6: A scheme for synthesizing speech at an arbitrary overall speech rate.

sentence, being larger at the onset of the sentence and gradually and monotonically tending to a positive value. The other type is a fast-varying component representing the local deviation from the general tendency, being always positive and occurring at almost every *bunsetsu* and at major syntactic boundaries.

5. A SCHEME FOR SPEECH RATE CONTROL IN SPEECH SYNTHESIS

Based on the analysis results and the considerations shown in the preceding section, a scheme is proposed for controlling the local speech rate of a reference utterance to obtain a synthetic utterance of an arbitrary global speech rate. The reference utterance can be natural or synthetic.

Figure 6 shows the general outline of the scheme. Here the reference is assumed to be a natural utterance at a normal speech rate. The reference utterance is analyzed and its parameters are retained for re-synthesis at different speech rates. At the same time, syntactic and semantic analyses are made on the text, to derive the prototypal speech rate control function for the modified utterance. The function is constructed from a decaying global component for each sentence, and several local components corresponding to each *bunsetsu* and each major syntactic boundary. In order to obtain the desired overall speech rate, however, it is necessary to adjust the size of the prototypal control function so that the resulting overall speech rate will optimally match the desired value. This is conducted by an iterative procedure until the resulting overall rate will be within the range of tolerance from the desired value. Once this is accomplished, the control function is used to re-synthesize speech at the desired overall speech rate. The procedure can be applied, with very few modifications, to text-to-speech synthesis at a desired overall speech rate.

6. CONCLUSIONS

A method has been proposed to measure the relative speech rate of an utterance against another utterance of the same linguistic/phonetic content chosen as the reference. Since the method calculates the short-time average speech rate, the optimum width of the smoothing window has been determined by perceptual experiments. The method has then been used to derive speech rate control functions for both fast and slow utterances relative to the utterances produced at a normal rate. The averaged results have shown that a speech rate control function for an utterance can be approximately decomposed into a global component for each sentence and local components for each *bunsetsu* and each major syntactic boundary. Based on these results, a scheme has been proposed for controlling the local speech rate of a reference utterance to obtain a synthetic utterance of an arbitrary global speech rate.

Although the results shown here are still preliminary, they are sufficiently consistent to lead to an approximate formulation of the process of speech rate control. Work is in progress to analyze larger amount of speech data for higher reliability and generality of the results, and to formulate the process of speech rate control in more quantitative terms.

Finally, controlling the duration of pauses to match the speech rate of utterances is another problem of importance in speech synthesis. The results of our concurrent study on this problem, however, were not mentioned here due to space limitations but will be presented elsewhere.

7. REFERENCES

1. I. Lehiste. *Suprasegmentals*. The M.I.T. Press, Cambridge, Mass., 1970.
2. I. Lehiste. “The timing of utterances and linguistic boundaries.” *J. Acoust. Soc. Am.*, vol. 51, pp. 2018–2024, 1973.
3. J. Vroomen, R. Collier and S. Mozziconacci. “Duration and intonation in emotional speech.” *Proceedings of EUROSPEECH ’93*, vol. 1, pp. 577–580, 1993.
4. H. Fujisaki, K. Hirose and M. Sugito. “Comparison of acoustic features of word accent in English and Japanese.” *J. Acoust. Soc. Jpn. (E)*, vol. 7, pp. 57–63, 1986.
5. H. Fujisaki and K. Hirose. “Analysis of voice fundamental frequency contour for declarative sentences of Japanese.” *J. Acoust. Soc. Jpn. (E)*, vol. 5, pp. 233–242, 1984.
6. S. Ohno and H. Fujisaki. “A method for quantitative analysis of the local speech rate.” *Proceedings of EUROSPEECH ’95*, vol. 1, pp. 421–424, 1995.