A METHOD FOR ANALYSIS OF THE LOCAL SPEECH RATE USING AN INVENTORY OF REFERENCE UNITS

Sumio Ohno, Hiroya Fujisaki and Hideyuki Taguchi

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

ABSTRACT

The speech rate is one of the important prosodic parameters essential for the naturalness of an utterance, yet comparatively little is known on the fine structures of speech rate variation in natural utterances. On the basis of the authors' definition of the relative local speech rate, the present paper describes an analysis of the changes in the local rate of speech units, each produced in isolation, when they are embedded in connected speech. The results, together with those already obtained by the authors, will lead to a complete scheme for speech rate control in speech synthesis by concatenation.

1. INTRODUCTION

It is well known that the speech rate varies both globally and locally in natural utterances due to various factors such as lexical or contrastive stress, syntactic boundary, emotion, etc., though the magnitude of their effects may vary from one language to another. Thus an appropriate control of the speech rate is essential for the synthesis of speech with high degrees of naturalness and expressiveness. In order to construct rules for speech rate control, however, one needs to have a clear definition of the speech rate and an objective method for its measurement.

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. Instead of using the segmental duration as a measure of the local speech rate, we have proposed the notion of relative local speech rate, i.e., the local speech rate of a target utterance relative to that of a reference utterance of the same text [1]. Based on this notion we have proposed a method for measuring the local speech rate, and have demonstrated that it can be used to obtain quantitative data concerning the effects of various factors on the local speech rate [2,3].

The present paper describes the method and the results of analyses of the relative speech rate, first in cases where the reference and the target are the utterances of the same text, and then in cases where the references are drawn from an inventory of stored units. Based on these results, 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.

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 Between Two Utterances of the Same Linguistic Content

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

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

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.

Figure 1 illustrates an example of the analysis of relative speech rate of an utterance produced at a fast speech rate against an utterance produced at a normal speech rate. The utterance is "Mizuumiwa yuugureno umino yooda." (The lake looks like a dusky 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. 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.

This method can be applied to find rules for speech rate control necessary to modify the speech rate of a reference utterance into an arbitrary rate.

2.3. Calculation of Relative Speech Rate Between a Reference and a Part of the Target

When the linguistic content of a reference utterance constitute only a part of the target utterance, the time axis alignment procedure has to be modified. Namely, the matching of the reference against the corresponding portion of the target can be accomplished by the technique of word spotting, based on the edge-free DP matching algorithm [4]. When two references correspond to two adjoining parts of the target utterance, their boundary is determined by comparing the distances between corresponding frames for the two references.

Thus the method can be applied to find out how the speech rates of references, say words or phrases uttered in isolation, change in connected speech. The results can then be utilized for constructing rules for speech rate control in speech synthesis by concatenation of stored units. A preliminary experiment along this line will be described in the next section.

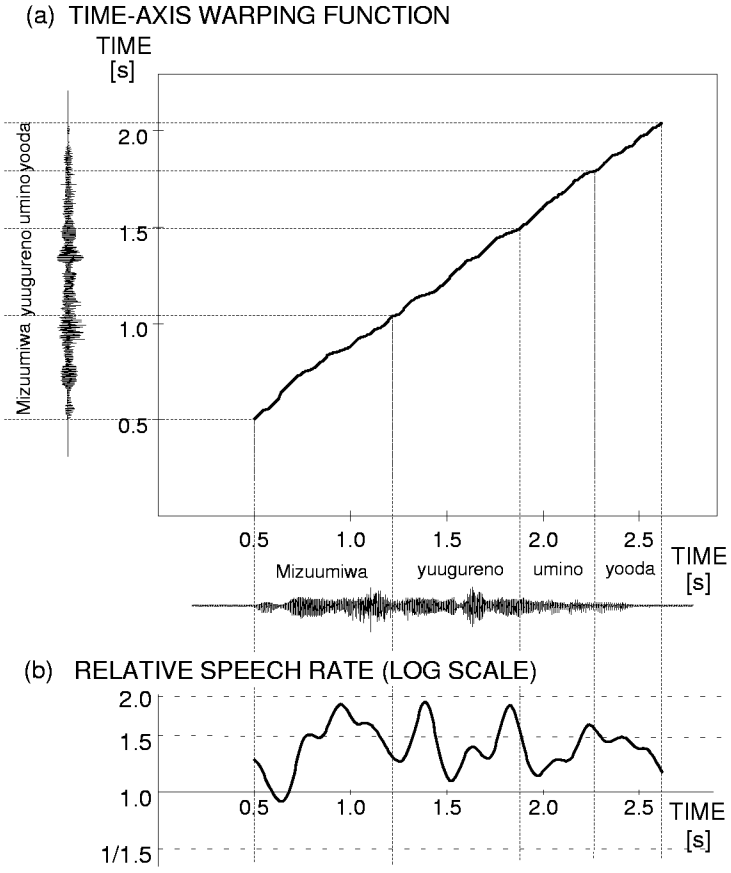

Fig. 1. An example of the analysis of relative speech rate of a 'fast' target utterance against a 'normal' reference utterance of "Mizuumiwa yuugureno umino yooda." (The lake looks like a dusky sea.). (a) DP matching path. (b) Relative speech rate on the time axis of the reference units.

3. EXPERIMENT

3.1. Speech Material

The target utterances for the analysis of changes in the relative speech rate of stored reference consisted of readings of sentences in a short story at three speech rates: normal, fast, and slow. The average speech rate for each group was 7.7 morae/s for the normal, 10.1 morae/s for the fast, and 6.0 morae/s for the slow groups of target utterances. The informant is a male native speaker of the common Japanese.

Each of the sentences was decomposed into constituent *'bunsetsu'* units. (A *'bunsetsu'* is a syntactic unit of Japanese, being a content word with or without following function word(s)). A randomized list of these units were then prepared and read three times by the same speaker at a normal speech rate (6.4 morae per second). These utterances served as the inventory of reference units for the following experiment.

These speech samples were digitized at 10 kHz with 16 bit precision. Parameters used for the DTW procedure were calculated at 5 ms intervals.

3.2. Results

3.2.1. Normal target vs. reference units

Figure 2 shows an example of the analysis of speech rate changes of four reference units in a target utterance of "Mizuumiwa yuugureno umino yooda." of one of the target utterances, where panel (a) shows the DP matching path and panel (b) shows the relative speech rate on the time axis of the reference units.

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 target utterances. The time axis of each of the reference units was slightly adjusted to correct for the deviations of the particular reference

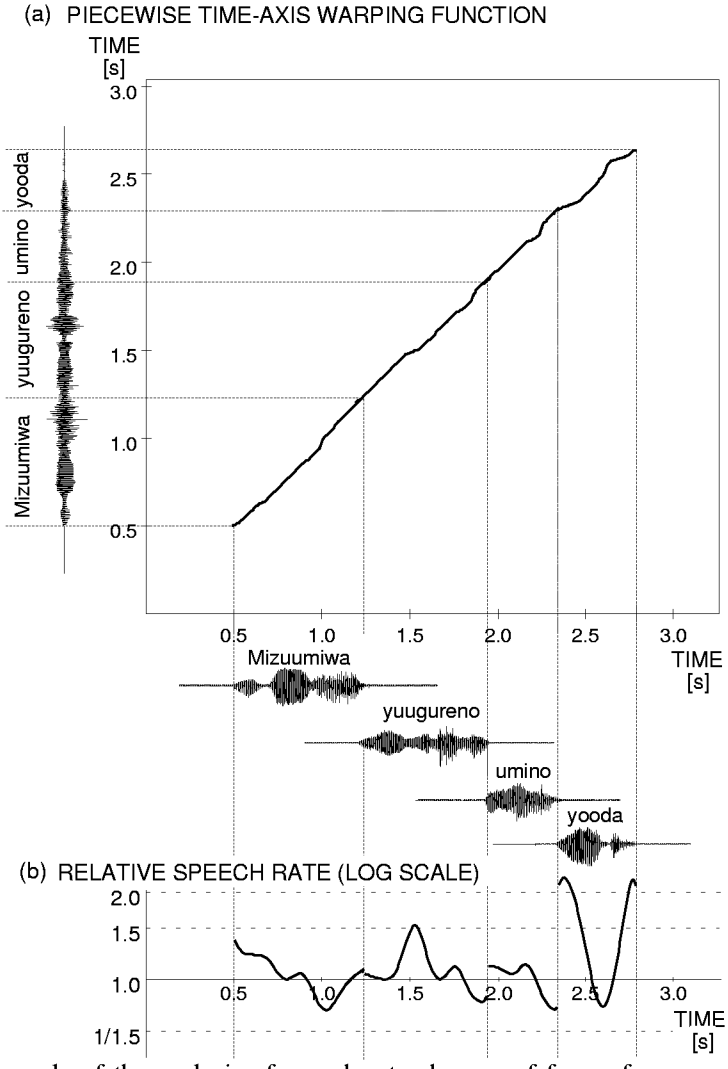

Fig. 2. An example of the analysis of speech rate changes of four reference units in a target utterance of "Mizuumiwa yuugureno umino yooda." (The lake looks like a dusky sea.). (a) DP matching path. (b) Relative speech rate on the time axis of the reference units.

utterance from the mean of the three reference utterances. Figure 3 shows an example of the averaged local speech rate of three target utterances against each of the units. In the target utterance, the mean speech rate is almost the same as that of the reference for the first three units, but the local speech rate shows fine structures of rate control within each *bunsetsu*. The final *bunsetsu* in the target utterance shows a marked tendency of *shortening* rather than lengthening.

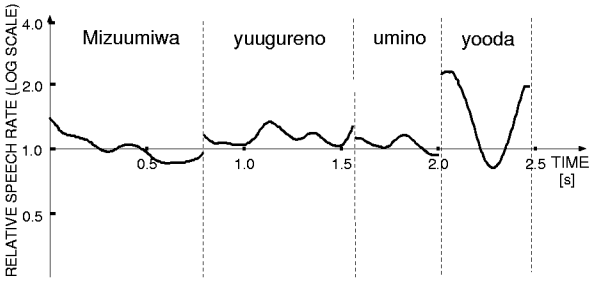

Fig. 3. Averaged relative speech rate in logarithmic scale of the '*bunsetsu*' units in the 'normal' utterances against their corresponding references produced in isolation.

3.2.2. Fast and slow targets vs. normal targets

In order to find out the way how the relative speech rate in connected speech changes from 'normal' to 'fast' or 'slow' utterances, the relative speech rate was obtained for each of the 'fast' or 'slow' utterances against a 'normal' utterance. Figure 4 shows the averaged results for the fast group in a dotted line, and those for the slow group in a broken line.

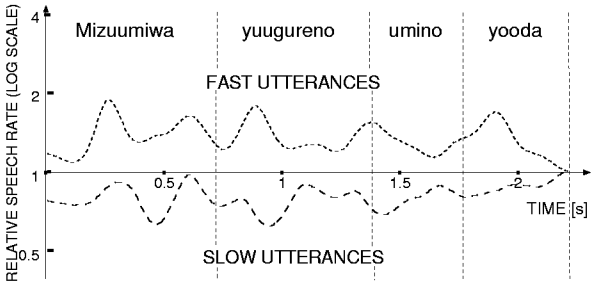

Fig. 4. Averaged relative speech rate in logarithmic scale for the 'fast' utterances (dotted lines) and the 'slow' utterances (broken lines).

4. SPEECH RATE CONTROL IN SPEECH SYNTHESIS BASED ON CONCATENATION OF STORED UNITS

Although the results of the study are still preliminary, they suggest a scheme for speech rate control for speech synthesis by concatenation of pre-stored units. The procedure for obtaining the time-axis warping function for the synthetic speech of an arbitrary global speech rate can be divided into the following two steps:

(1) Derivation of a time-axis warping function for each of the constituent reference units for producing a continuous synthetic utterance of a 'normal' speech rate by concatenation.

(2) Derivation of a time-axis warping function for converting the local speech rate of the synthetic utterance obtained in (1) into that of an arbitrary speech rate.

As a first approximation, the two-stage time-axis warping can be combined into one by multiplying the two warping functions from (1) and (2), or, equivalently, by taking the sum of the logarithms of the two warping functions.

5. SUMMARY AND CONCLUSION

On the basis of the authors' definition of the relative local speech rate, a preliminary study was made to obtain speech rate control data for the units that constitute a connected utterance. The results can be utilized to derive rules for speech rate control for speech synthesis by concatenation from an inventory of pre-stored units. These rules, when combined further with those for modifying the local speech rate of a continuous utterance, will lead to a complete scheme for speech rate control in speech synthesis by concatenation.

REFERENCES

[1] S. Ohno and H. Fujisaki, "A method for quantitative analysis of the local speech rate," *Proceedings of the 4th European Conference on Speech Communication and Technology*, Madrid, vol. 1, pp. 421–424, 1995.

[2] S. Ohno, M. Fukumiya and H. Fujisaki, "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] S. Ohno and H. Fujisaki, "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] H. Sakoe, "Two-level DP-matching — A dynamic programming-based pattern matching algorithm for connected word recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-27, no. 6, pp. 588–595, 1979.