ISCA Archive
http://www.isca-speech.org/archive

EUROSPEECH '97
5th European Conference on Speech Communication and Technology
Rhodes, Greece, September 22-25, 1997

EXTRACTION AND REPRESENTATION RHYTHMIC COMPONENTS OF SPONTANEOUS SPEECH

*S. Kitaazawa, H. Ichikawa, S. Kobayashi & Y. Nishinuma**
Department of Computer Science, Faculty of Information, Shizuoka University,
5-1, 3-Chome, Jouhoku, Hamamatsu, 432, JAPAN
Tel. +81 53 478 1471, FAX: +81 53 475 4595, kitazawa@cs.inf.shizuoka.ac.jp
*CNRS, URA 261 "Laboratoire parole et langage", Universite de Provence,
13621 Aix-en-Provence, France

ABSTRACT

Speech speed is measured and displayed with our specific algorithm TEMAX (Temporal Evaluation and Measurement Algorithm by KS). The TEMAX-gram, a sonagraphic output of speech envelope, the DFT using a 1-second window is convenient to set off isosyllabic characteristics. For Japanese traces 2 dark bars, called rhythmic formants: RF1 and RF2: the first one, around 8 Hz, and the second one, at halfway. RF1 corresponds to speech rate, RF2 represents the bimoraic rhythmic foot. As far as English, its isochronic characteristics are observable with a 2-seconds window as RF1. Furthermore, using a 1-second window the periodicity of syllables between stress is displayed as RF2.

1. INTRODUCTION

Temporal aspects of speech may be treated in different ways. Speech rate is a measure of the speed of utterance production. Speech speed changes dynamically according to stylistic factors; it varies during a talk, a lecture and much more in dialogues and conversations.

Speech rate is usually expressed by counting the number of characters or words spoken per minute. But this deferred estimate of speech rate is unfit for the observation of the dynamic characteristics in spontaneous speech. Therefore, we need an instantaneous representation of speech rate such as number of syllables or moras per second. Individual syllables can provide the instantaneous syllable rate as an inverse function of their duration. This speech rate is derived from syllable length changes on each syllable according to its duration which is not a linear function of speech speed.

2. HOW TO MEASURE SPEECH RATE

Speech rate does play a part in speech rhythm. Still, speech rhythm mainly seems to result from repetitions of an element larger than the syllable, such as stress or rhythmic foot in English. If we use duration of individual segments, we need a segment labeling for syllables and stresses in order to count them. Do we need to label manually to measure speech rate? Is it possible to measure speech rate by an automatic processing?

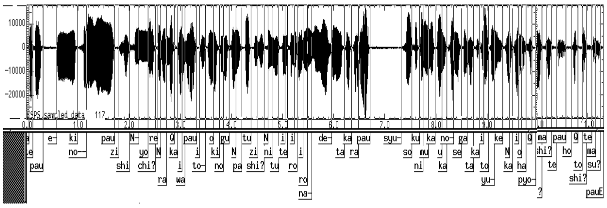


Fig. 1. Speech wave segmented according to mora. [sound A0140S01.WAV]

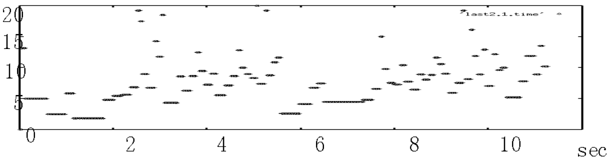


Fig. 2. Instantaneous speech rate estimated as an inverse of mora duration observed in Fig. 1.

2.1 Mora Segmentation

Number of segments observed in a second is a primitive definition of speech rate. This work can be done by manual segmentation of speech wave by hearing. Fig. 1. Shows an example segmentation of Japanese talk on radio. This is accurate, but time consuming and inconvenient to know dynamics during talk.

2.2 Estimation from Duration

An instantaneous estimation of speech rate is an inverse of duration of segment such as mora. Fig. 2. Shows graphically inverse of duration for individual mora as an horizontal line according to its duration. Variation is too large to accept as speech rate.

2.3 Difficulties in Definition

Speech rate may be estimated through a phoneme recognition method, i.e., phonemes are marked along the time axis. The resulting phoneme or syllable lengths, averaged along some intervals, will give a number of syllables per second. Technical implementation is difficult to realize with today's automatic speech recognition technology.

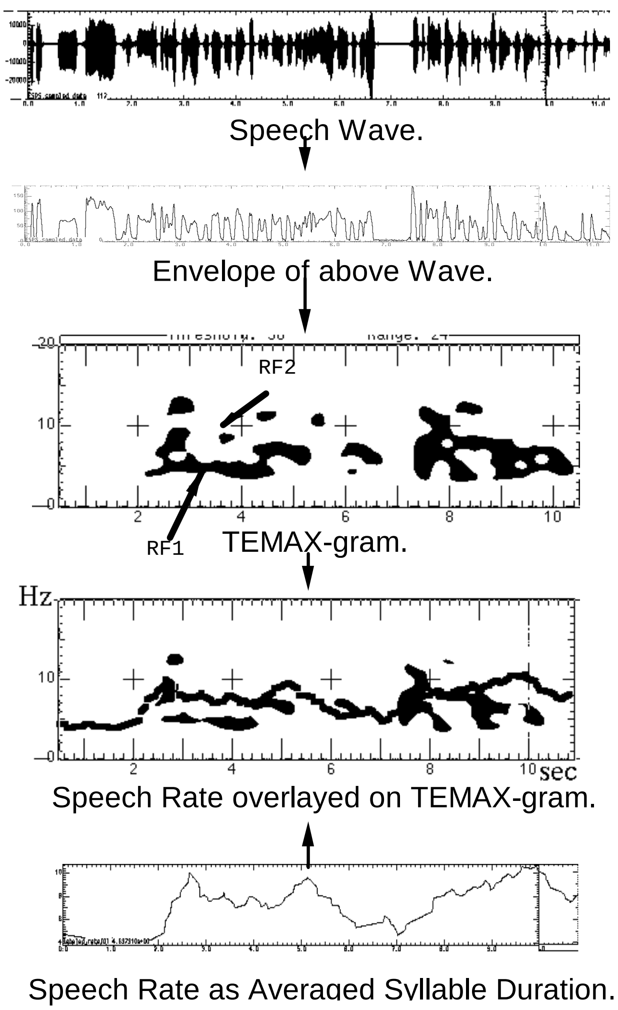


Fig. 3 Procedure for Speech Rate thorough TEMAX.

3. YET ANOTHER TEMPORAL MEASUREMENT

In order to tackle the assessment of temporal aspects of speech, we developed an algorithm called TEMAX (Temporal Evaluation and Measurement Algorithm by ks), which is a signal processing procedure to evaluate speech rate by amplitude envelope of speech wave.[3,4,5] This processing is illustrated in Fig. 3.

3.1 Speech Wave

Speech sample is a spontaneous speech which last about 10 s. This sound [sound A0140S01.WAV] is a male speech on radio broadcast unprepared monologue.

3.2 Speech Envelope

Speech envelope is computed through low-pass filtering of half-wave rectified speech or by the root of the mean squares in a moving window. The cutoff frequency of the low-pass filter is 20 Hz, and window size is 15 ms. Dips in the envelope correspond to consonants or phonemic boundaries, therefore dips within a unit of time are correlated with speech rate.

3.3 TEMAX-gram and Rhythmic Formant

A spectrogram of the above speech envelope. Sampling frequency of envelopes is set to 40 Hz, the frequency of the spectrogram ranges from 0 Hz to 20 Hz. By DFT of the envelope with a 1 second window, the speaking rate appears a dominant spectral peak in a frequency-time plane like a formantic pattern in a spectrogram (we call it a rhythmic formant, RF).

Gray gauged monochrome patterns appear in the 20 Hz frequency region. In this graph, higher energy components produce frequency bands (around 8 Hz) which correspond to a speech rate of about 8 mora/second. This phenomena is due to the isosyllabism or isochronism of the speech rhythm.

One hypothesis in TEMAX-gram is that:

- Within a TEMAX-gram there must be shown continuous variation as dominant dark bars.

In order to satisfy the above requirement, analysis condition is adjusted.

3.4 Averaged Syllable Duration

Syllable duration is smoothed to represent a continuous contour of speech rate. Overlaying this on the TEMAX-gram show consistency between two results.

4. BIMORAIC FOOTS IN JAPANESE[1]

Japanese is called a mora-timed language in a family of syllable-timed languages. Japanese pronunciation unit is a consonant followed by a vowel (a CV syllable) called a mora. Japanese rhythm is kept constant based on mora.

4.1 Mora Merging

Some moras are connect together to compose a larger peaks in envelope. An example is shown in Fig. 4, at the beginning part of speech in Fig. 1.

There is distinction between a mora vowel and two mora vowel (long vowel), the latter has approximately double duration of the former. Apparently these vowels compose

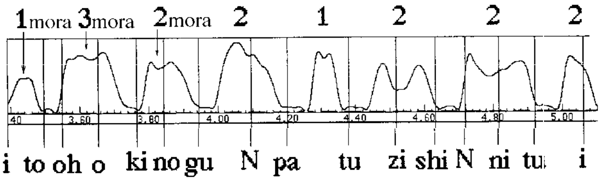


Fig. 4. Peaks Merging Multiple Moras Viewed in a Part of Envelope from Fig.1.

a merged peak on envelope.
Mora before a nasal consonant or syllabic nasals is likely to merge. A syllabic nasal /N/ occupy a unit of duration in Japanese rhythm as a mora.

Vowel concatenation is often occur in Japanese, as a merged peak of envelope. Combinations of vowels are as follows sorted in their frequency:

/ai/, /oo/, /ee/, /aa/, /eN/, /aN/, /ei/

This list includes long vowels and syllabic nasals.

4.2 Statistics of Merged Mora

A statistics, where how many moras are merged in a peak, observed in Japanese samples. Isolated mora peak consists only 42% of total occurrences. Two mora as 40%, 3 mora 12%, 4 mora 4%, 5 mora 1%, 6 mora .3% respectively, and more than 4 mora is very rare.

What we call a peak in envelope is a part where dominant amplitude is separated by two adjacent power dips.

4.3 Multiple Bars in TEMAX-gram

TEMAX signal processing is similar to speech sonagraph. Dominant frequency components like formants are observable in TEMAX gram. In simple case, single major component is significant. Usually multiples of dominant components are evident.

Like difference between wide band sonagraph and narrow band sonagraph, TEMAX-gram differs depending on window size. Shorter window enhances individual phoneme characteristics, while longer window enhance long term trend of smoothed movement.

4.4 Simulated Moraic Envelope

Real speech is too complex to see characteristics. In order to obtain intuition about rhythmic structures, we made TEMAX analysis on artificially generated envelope.

Our assumptions are:

- Syllable-timed (including mora-timed) speech envelope oscillate synchronous with the rate of speech.
- One dominant frequency component will be observed in TEMAX-gram.

In fact, experiment with steady envelope showed dominant component at the fundamental frequency accompanied with number of harmonics'. Experiment with increasing and decreasing demonstrated complex texture in TEMAX-gram.

Fig. 5 upper part is an examined envelope wave simulating 8 mora per second with a half-wave rectified 8 Hz sinusoid where some adjacent 2 peaks or 3 peaks are connected together composing a larger peaks. Ratio of occurrence for uni-mora, 2 mora, 3 mora peaks is proportional to real data (as in 4.2).

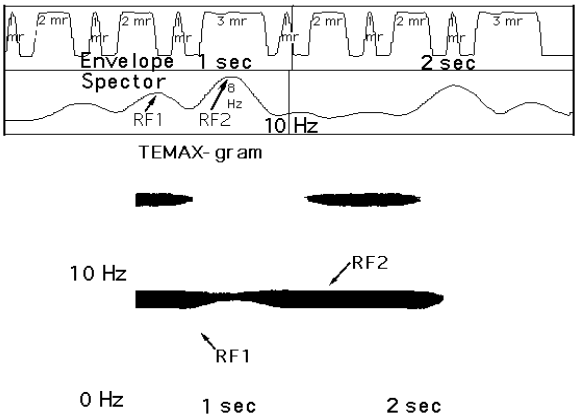


Fig. 5 Simulated Envelope of Japanese Bimoraic Foot and its Spectrum and TMAX-gram.

Middle part of Fig. 5 is spectrum of the test envelope with a 1 s window where dominant 8 Hz peak (RF2) and the secondary peak (RF1) around 6 Hz and so on. A mirror spectrum appears in high frequency region.

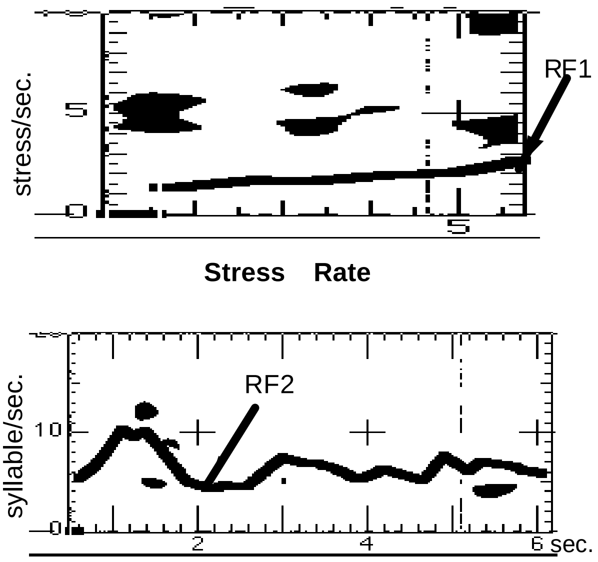
In bottom TEMAX-gram, a continuous display of the envelope spectrum in gray scale, there are dark bars around RF2 and RF1 and the third lower one as well as a higher mirrored picture.

This results suggests how we read TEMAX-gram of real speech that has much complex shape of envelope.

4.5 TEMAX Observation

For Japanese, the TEMAX-gram traces 2 dark bars, called rhythmic formants: RF1 and RF2: one around 8 Hz, and the other at about halfway in Fig. 1. The lower rhythmic formant, RF1, represents the bimoraic rhythmic foot.[1] The upper one, RF2, corresponds to speech rate, which appears almost steady in read speech and monologue, but shows wide variations in spontaneous speech. We found the DFT of the envelope using a 1-second window is convenient to emphasizes isosyllabicity.

As we have seen in above example, we should be careful in observing TEMAX-gram. As an example, in mora-timed language, TEMAX-gram often showed fake bars. So we need to neglect those bars lookin for the real speech rate. Also we need knowledge of rhythmic structure of the language analyzing, such as bimoraic behavior. Syllable-timed languages can be treated in the similar way as Japanese.



Stress Rate

Syllable Rate

"You didn't! You should have rented a car. You see nothing from one of those buses. I told you so remember?"

Fig. 6 TEMAX Analysis of Isochronic English. [sound A0140S02.WAV]

4.6 Japanese by Foreign Students

Japanese learning foreign students have difficulty in mastering rhythm of Japanese. We analyzed of their Japanese and noticed that TEMAX-grams seldom displayed 2 dark bars but one single bar. This means these students do not sufficiently master the Japanese bimoraic foot. Our method may then be a useful tool for Japanese language teaching.

5. ISOCHRONISM IN ENGLISH

World languages are classified into two categories: syllable-timed languages and stress-timed languages.

Although the development of TEMAX is based on the mora-timed Japanese, it turns out to be useful also for stress-timed languages like English.

5.1 Syllable Rate

In syllable-timed language, syllable is speced along time at equidistant point from its adjacent syllables. This can be measured with TEMAX-gram as an dark bar.

Then, English, a stress-timed language, how this syllable-rate performs is not well known yet. A linguist says that syllables are spaced equidistantly between two adjacent stresses.[2] But number of syllables between two adjacent stress is different. Therefore, syllable rate isconstant between stress but it changes discontinuously when passing over the stress point.

5.2 Stress Rate

In stress-timed language, like English, interval between stress is constant. The inverse of this stress interval is the stress rate. Stress is expressed by amplitude and duration of stressed vowel. TEMAX analysis is applicable to demonstrate dark bars as a rhythmic formant for stress with modification of the window to cover stresses (say 2 s, sice stress rate is estimated around 2 or 3 stresses per second).

5.3 TEMAX-gram Reading

An example is shown in Fig. 6, with a speech sample taken from a CD of English dialogue. The figure shows that stress is periodic around 1 Hz and this isochronic characteristic is observable with a 2-second window as RF1. Furthermore, with a 1-second window, RF2 shows the periodicity of syllables between stresses as broken bars around 5 to 10 Hz.

6. CONCLUSION

In this study, we reported an algorithm for measuring speech rate on the basis of syllable periodicity or stress production, i.e. the tempo or the rhythm. We implemented a signal processing algorithm with spectrographic display named TEMAX to evaluate tempo in terms of not only mora/second for mora-timed Japanese, but also stress/second and syllable/second for interstress syllables in English. Thus, two languages with extreme differences as far as their rhythmic structure were successfully analyzed. Our research focuses now on other languages such as Korean, French which are examined to observe their rhythmic structure. These results will be reported in the further conferences

7. REFERENCES

[1] Poser, and William, "Hypocoristic Formation in Japanese," *Proceeings of West Coast Conference on Formal Linguistics*, 3, 218-229, (1984).

[2] Kohno M., “Perceptual Sense Unit and Echoic Memory”, *International Journal of Psycholinguistics*, Vol.8, pp. 13-31, 1993.

[3] Kitazawa S., Kobayashi S., Matsunaga T., and Ichikawa H.: “Tempo Estimation by Wave Envelope for Recognition of Paralinguistic Features in Spontaneous Speech”, *Proc. ICSLP*, Vol. 3, 1691-1694, 1994.

[4] Kitazawa S., Ichikawa H., and Kobayashi S., “TEMAX: Visualization of temporal variation in spontaneous speech”, *J.A.S.A.* 100,4,Pt2, 2851, 1996.

[5] Kitazawa S., Ichikawa H., and Kobayashi S., and Nishinuma Y., “How Can We Extract And Represent Rhythmic Components of Spontaneous Speech?” *Proc. FSA*, 1997.