社団法人 電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report
SP2005-183 (2006-03)

What Is Rhythm? Can We Capture Syllable Shapes From Intensity Contours?

Masahiko KOMATSU † and Takayuki ARAI ‡

† School of Psychological Science, Health Sciences University of Hokkaido
Ainosato 2-5, Kita-ku, Sapporo, 002-8072 Japan

‡ Department of Electrical and Electronics Engineering, Sophia University
7-1 Kioi-cho, Chiyoda-ku, Tokyo, 102-8554 Japan

E-mail: † koma2@hoku-iryo-u.ac.jp, ‡ arai@sophia.ac.jp

Abstract Rhythm can be viewed in two different ways. Acoustic approach views rhythm as the alternating pattern of high and low intensity, which is regarded as a syllable. Phonemic approach attributes rhythm to the phonemic complexity of syllable structure and calculates rhythm based on the durations of consonant and vowel intervals. This paper investigates how well the acoustic approach fits to the phonemic approach. It tests two algorithms adopted in our previous studies, which estimate syllable centers from intensity contours based on the calculation of RMS and correlation with a cosine curve. The results are evaluated against the criteria of the phonemic approach. It concludes that the algorithms are valid, hence syllable shapes can be captured from the intensity contour.

Keyword Rhythm, Syllable, Intensity, RMS, Correlation coefficients

1. Introduction

Rhythm can be viewed in two different ways. On one hand, it is intuitively the alternating pattern of high and low intensity or amplitude (acoustic approach). One cycle of this alternation may be regarded as a syllable. On the other hand, recent research attributes rhythm to the phonemic complexity of syllable structure and calculates rhythm based on the durations of consonant and vowel intervals, or phonemic units (phonemic approach) [1, 2]. Although these two views are underlain by the common idea, their approaches are quite different. It is not clear which of these two views is correct by nature.

The present paper investigates how well the acoustic approach fits to the phonemic approach. Syllable centers are estimated only from the source component features of the source-filter model, especially the intensity contours, and the results are evaluated against the criteria of the phonemic approach.

The present paper tests two algorithms to detect syllables from intensity contours against a labeled Japanese speech corpus. In our previous studies, we tried to capture syllable shapes from intensity contours. Although it seemed that syllables were captured to some extent, the algorithms used were not verified thoroughly. If the algorithms are proved to be valid, it follows that the arguments made in our previous studies were justified, hence the syllable shapes can be captured from intensity contours. This paper fortifies our previous arguments with new experimental results.

2. Previous Studies

2.1. Averaged Syllable Shapes

In [3], averaged syllable shapes (Figure 1) showed cross-linguistic similarities and difference among Chinese, English, Japanese, and Spanish, which linguistically differ in rhythm and lexical accent types.

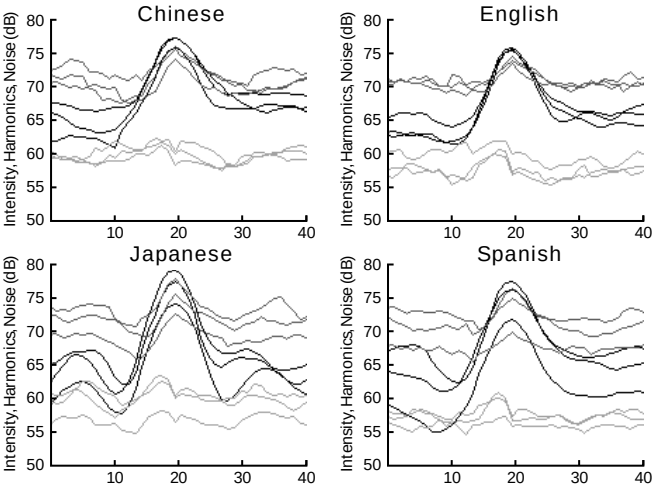


Figure 1. Averaged syllables (reproduced from [3]). *x*-axis: Frame number (1 frame = 10 ms). *y*-axis: Amplitude (dB). The figure shows averaged contours of intensity, harmonics amplitude, and noise amplitude (dark lines, medium lines, pale lines, respectively). Each graph shows averaged contours from three speakers (Hence, there are three lines for each type of contour).

The calculation was conducted in the following way: The automatic algorithm searches for the local peaks of harmonics amplitude. (The algorithm would work almost equivalently using the intensity instead of harmonics

amplitude [4]). Then, it extracts 400 ms intervals of speech signal centering the detected peaks, and averages these intervals.

In sum, the algorithm produces the averaged contours time-aligned at the peaks of harmonics amplitude that can be regarded as syllable centers.

In the present paper, the searching algorithm for the local peaks is tested.

2.2. Syllable Search Method by Correlation Coefficients

In [5], the Syllable Search Method was proposed as a feature extraction method for the neural network for the language discrimination task of English and Japanese. This method was supposed to locates typical syllables in the speech signal, which were then input to the neural network.

First, the procedure calculates correlation coefficients between the power envelope of the speech signal and the reference signal (one cycle of $1 - cos(n)$ at 4 Hz; the same shape as the Hanning window) for each analysis frame. Then, it selects the 10 highest peaks of the contour of the correlation coefficient values during the signal, which were regarded as syllables.

The idea behind this method is that the most frequent syllable duration and the peak of the modulation spectrum of the speech signal is around 4 Hz both in English and Japanese [6]. Therefore, the 4 Hz reference signal is expected to register high correlation with syllable locations.

The present paper tests how well the local peaks of the values of correlation coefficients represent syllables.

3. Experiment

3.1. Purpose and Design

The present experiment verifies how well syllables can be detected by intensity contours. They test two algorithms: the ones using RMS peaks (similar to the one in section 2.1) and correlation coefficient peaks (similar to the one in section 2.2). For both algorithms, two parameters, i.e., the window length of analysis and the threshold value to discard peaks, were varied in the same manner.

The experiment checks how well the detected peaks represent syllables. Following [1, 2], the syllables, or the constituents of rhythm, are considered to be CV(C) sequences, where C stands for consonant(s) and V for vowel(s). Successive V segments (e.g., “aa” and “ioo” in the top panels of Figure 2) are regarded as one V segment; successive C segments, one C segment. The syllable detection algorithm is regarded as successful if each V segment contains only one peak and C segments contain no peak. In this paper, the results of the peak detection are interpreted in terms of how correctly V segments are detected and how correctly the detected peaks represent V segments.

All speech data of Japanese MULTEXT [7] were used in the experiment. It includes, in total, 480 Japanese speech files of short passages and their time-aligned phoneme labeling. The recording includes 3 male and 3 female speakers in two styles: reading and simulated spontaneous speech (henceforth, *reading* and *play* respectively).

3.2. Tested Algorithms

3.2.1. Syllable Detection by RMS Peaks

Because the syllable nucleus (V) usually has larger amplitude than edges (C), local peaks of the amplitude contour are expected to represent syllable nuclei.

In this algorithm, the values in the speech signal were squared and multiplied by the Hanning window, and then RMS was calculated. The analysis window was shifted by 8 ms. The window length was varied: 32, 64, 128, 256, and 512 ms.

In the resultant RMS contour, all local maxima greater than or equal to the threshold were picked up. The threshold value was varied: 0, 0.2, 0.4, 0.6, and 0.8 of the global maximum of the RMS contour.

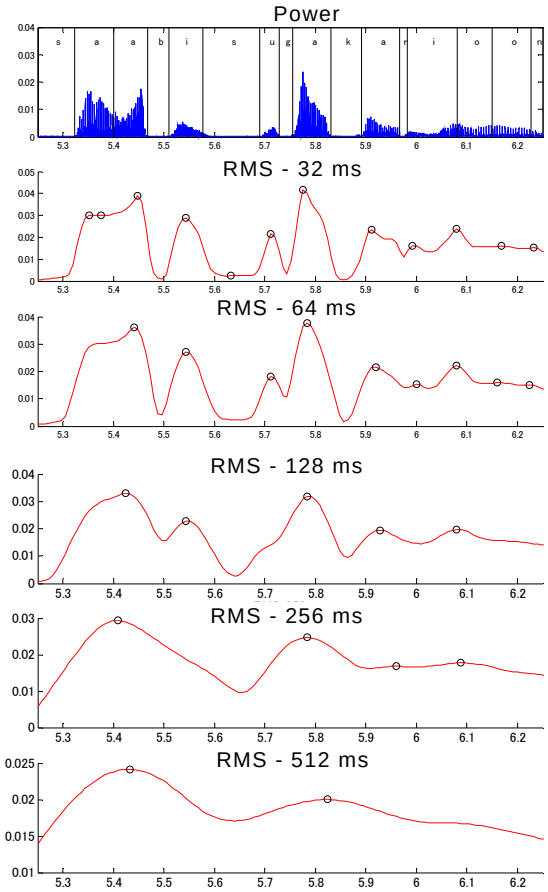
Increasing the window length has the effect of smoothing the contour (see Figure 2a for an example); and increasing the threshold value, the effect of discarding small amplitude segments, hopefully consonants and noise.

3.2.2. Syllable Detection by Correlation Coefficient Peaks

The values in the speech signal were squared, and their correlation coefficients with one cycle of cosine wave (here, the Hanning window was used) were calculated. As with the RMS algorithm, the analysis window was shifted by 8 ms, and the window lengths were: 32, 64, 128, 256, and 512 ms.

Assuming that the typical syllable duration is at 4 Hz, the 256 ms (3.9 Hz) cycle of cosine wave is expected to show high correlation with it.

In the contour of correlation coefficient values, all local maxima greater than or equal to the threshold were picked up. The threshold values were 0, 0.2, 0.4, 0.6, and 0.8 of the global maximum of the contour.



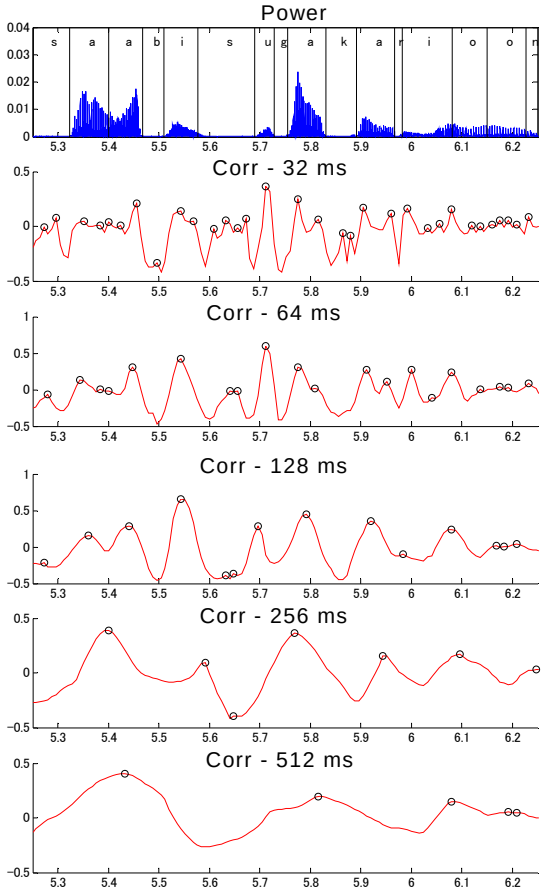


Figure 2. (a) Example of RMS procedure. Threshold = 0. (b) Example of Correlation procedure. Threshold = 0.

Increasing the window length and the threshold value is supposed to have the similar effects as in the RMS algorithm (see Figure 2b for an example).

The calculation procedure of the correlation coefficients is, by definition, is a normalized multiplication. In this respect, the correlation coefficients of the speech signal with the Hanning window in the correlation method is similar to the multiplication of the speech signal by the Hanning window. In order to compare these two similar but different methods, the same window shape, the same window lengths, and the same threshold values were used in this experiment.

3.3. Results

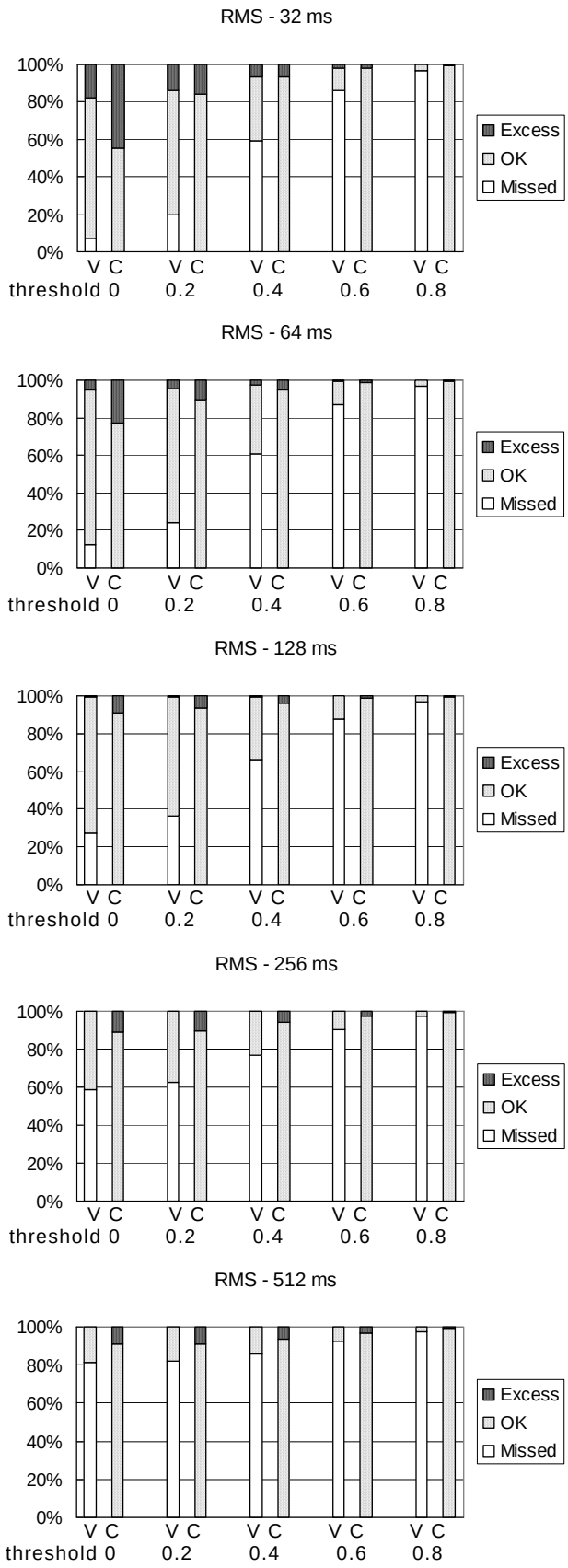
Locations of the detected peaks (local maxima in the above procedure) were checked against the time-aligned phoneme labels in the corpus. The syllable detection algorithm is regarded as successful if each V segment contains only one peak and C segments contain no peak. If V does not contain a peak, it means that the detection missed the segment. If V contains more than one peak or if C contains a peak, it is an excessive detection.

The results indicated almost no difference between *reading* and *play* speeches. Only the results of *reading* are reported due to the space limitation.

3.3.1. RMS Method

Figure 3a (left half of the page) shows how correctly the segments were detected. For V segments, *Missed* indicates the percentage of segments including no peak against the number of all V segments. Likewise, *OK* indicates the percentage of segments including one peak; and *Excess(ive)*, more than one peak. For C segments, *OK* means no peak; *Excess*, one peak or more. Pause segments are excluded from the graphs for simplicity.

In general, as was expected, as the window length increases (from the top panel to the bottom) and as the threshold value increases (from left to right within each panel), missed V segments increases and excessively detected V and C segments decrease. This is because the number of detected peaks as a whole decreases. The detection of syllables is considered to be accurate if both the percentage of correctly detected V and the percentage of correctly undetected C are high (Large *OK* areas of both C and V).



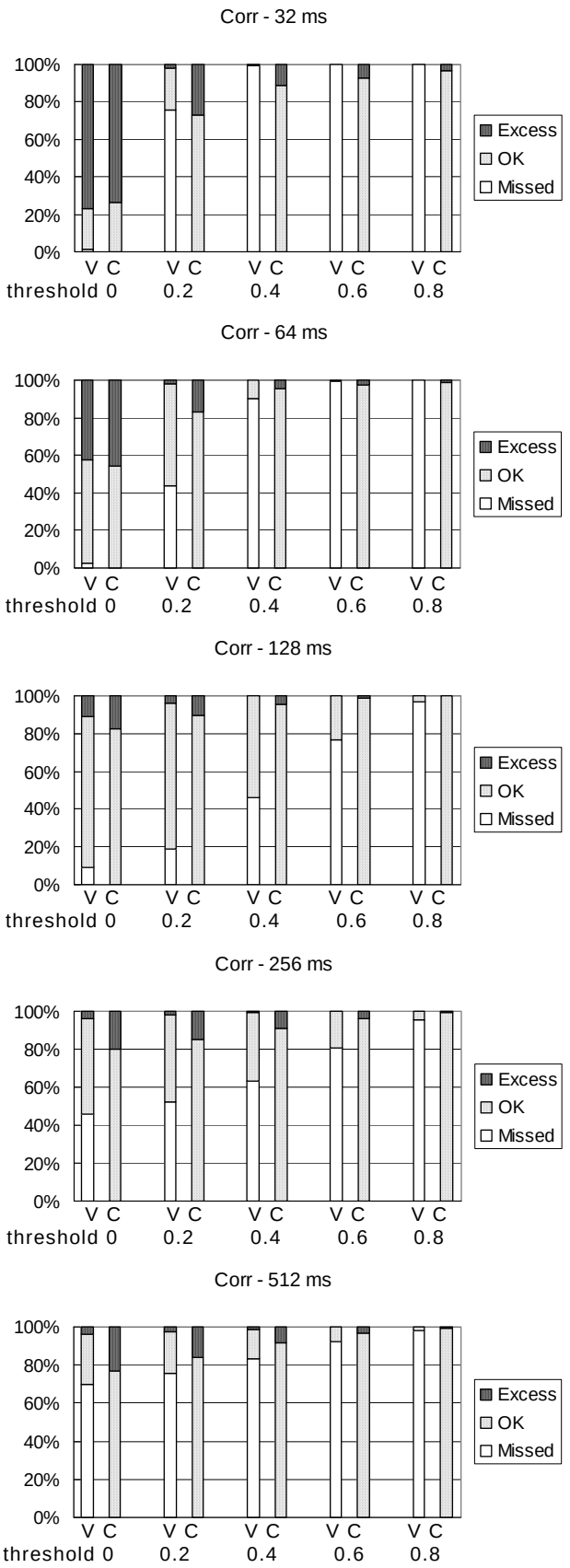
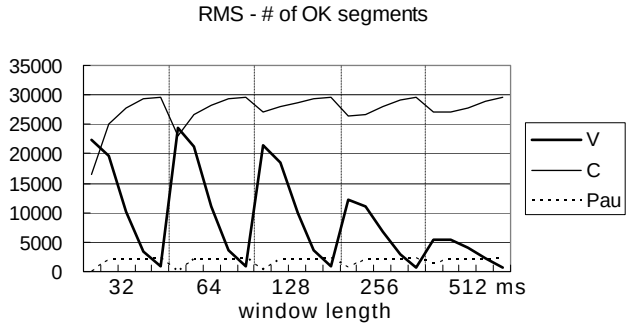


Figure 3. (a) RMS method: % of segments. (b) Correlation method: % of segments.



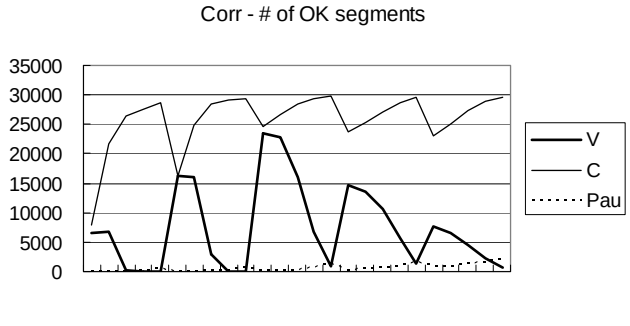
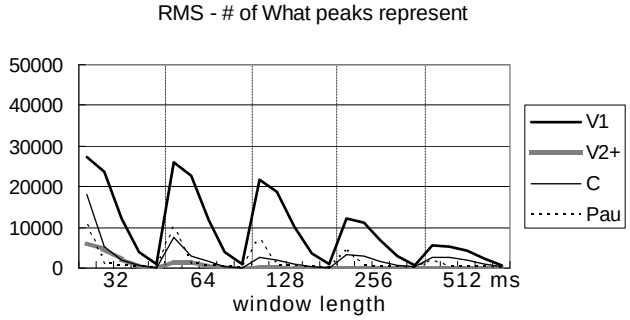


Figure 4. (a) RMS method: OK counts. (b) Correlation method: OK counts.
Threshold = 0, 0.2, 0.4, 0.6, 0.8 (from left to right in each window length)



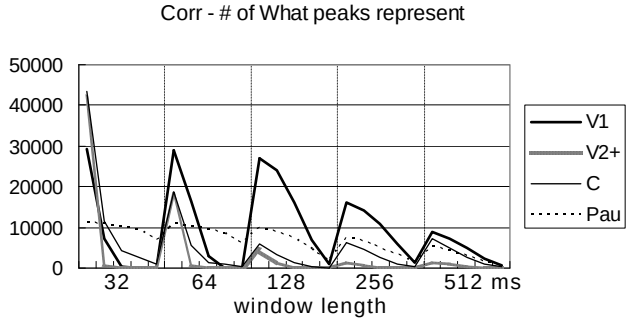


Figure 5. (a) RMS method: What peaks mean. (b) Correlation method: What peaks mean.
Threshold = 0, 0.2, 0.4, 0.6, 0.8 (from left to right in each window length)

Figure 4a shows some of the actual counts of the segments presented in Figure 3a. It indicates the number of V segments correctly detected (*OK* of V in Figure 3a), C segments correctly undetected (*OK* of C in Figure 3a), and pause segments correctly undetected. The numbers of V and C segments shown in the graph indicates that 64 and 128 ms window lengths and 0 and 0.2 thresholds are the best syllable detectors. Taking into account pause segments, a minimum threshold is necessary because the threshold value of 0, i.e., no threshold, produces bad results.

Figure 5a shows the number of detected peaks grouped by what they point to. While Figures 3a and 4a showed how segments were detected (you can read how many segments were correctly detected), Figure 5a shows what detected peaks mean (you can read how many detected peaks are reliable). *V1* is the number of peaks reliable as syllable detectors, and the other categories are the wrong detectors. If there is only one peak in a V segment, it is included in *V1*. If there is more than one peak in a V segment, one of them is classified as *V1* and the others *V2+*. *C* and *Pau* are the peaks in C and pause segments. Note that the numbers in the graph are not exact because the 6th or later peaks in one segment were ignored in the counting procedure.

The figure shows that in the window lengths of 64 and 128 ms, the number of *V1* is quite larger than the other categories. It also shows the change of threshold from 0 to 0.2 reduces *V1* only a little but other categories drastically. These parameter settings were the best ones in Figures 3a and 4a as well.

3.3.2. Correlation Method

In general, the results (Figure 3b) showed the same tendency as in the results of RMS: as the window length increases and as the threshold value increases, missed V segments increases and excessively detected V and C segments decrease.

However, against our expectation, the window length of 128 ms (threshold; 0, 0.2) registered the best results (see Figures 3b and 4b). The superiority of the 128 ms window length (7.8 Hz) is against the expectation that the 256 ms (3.9 Hz) would best capture syllables.

The results produced by the best parameter settings in the correlation method (window length: 128 ms; threshold: 0, 0.2) were approximately the same as the results of the best setting in the RMS method (window length: 64, 128; threshold: 0, 0.2).

4. Discussion

4.1. Previous Studies and Future Improvements

Both the RMS and correlation methods registered high correct rates of syllable detection when the parameters were properly set. It means that the approaches adopted in our previous studies [3, 5] were correct and hence, as was done in those studies, syllable shapes can be captured from intensity contours.

Although the present experiments endorsed the validity of the approaches in general, the details of the analysis procedure are yet to be examined. In [3], the harmonics contour, in which local peaks were detected, was calculated by Praat [8] from the intensity contour (analyzed with the 32 ms analysis window) and the HNR contour (analyzed with the 80 ms analysis window; not explicitly stated in the manual though). This may not have been the best parameter setting and may have detected excessive peaks, but it was devised with other techniques: it discarded the peaks 18 dB below the global maximum of the contour (corresponding to the threshold setting in the present experiments) and the peaks within the distance of 100 ms from their adjacent peaks (removing *V2+* in Figure 5). In [5], the reference signal at 4 Hz was used, which is close to the 256 ms setting in the correlation method tested in this paper, i.e., not the best one. However, only the best 10 candidates were used in [5]; Figure 5b suggests that most of them were the correct syllable identifiers.

Besides the techniques used in [3, 5], other improvements seem available. For example, the thresholds by amplitude in the correlation method seem available. The combination of RMS and correlation methods may produce better results. Incorporation of pitch and HNR may be possible, too.

For the purpose of testing, the present experiment adopted the simplest procedures. In other words, potentially it is possible to capture syllables better than presented in this paper.

4.2. What is syllable?

The results of the correlation method showed that the window of 128 ms (7.8 Hz) was better than the one of 256 ms (3.9 Hz), which may provoke interesting arguments over the definition of syllable. In [6], it was found that in Japanese spontaneous speech syllable durations are around 4 Hz, which corresponds to the peak of the modulation spectrum. Note that in [6] the term *syllables* do not only refer to *phonologically* defined syllables but refer to *phonetic* coalescences of plural phonological syllables. This leads to the question that phonological syllables occur faster than 4 Hz, some of which do not have strong intensity peaks. Weak syllables may be easily merged into adjacent syllables. In phonology, it is known that the *foot* structure often consists of a strong and a weak syllable [9]. Considering that it is phonological feet that correspond to the 4 Hz peak of the modulation spectrum, it is not contradictory that the 7.8 Hz window detected syllables better than 3.9 Hz window in the present experiment.

5. Concluding Remarks

This paper constitutes part of a larger research project [10], which investigates the correspondence of the source component of the source-filter model (acoustic model), including intensity, to prosody (linguistic features), including rhythm. It has been found that rhythm is closely related to syllable structures. Although the syllable detection only from the source component may seem to be a *backend* research topic, what constitutes rhythm and how humans sense rhythm are *frontiers* yet to be explored.

References

[1] F. Ramus, M. Nespor, and J. Mehler, “Correlates of linguistic rhythm in the speech signal,” Cognition, vol. 73, pp. 265-292, 1999.

[2] E. Grabe and E. L. Low, “Durational variability in speech and the Rhythm Class Hypothesis,” in Laboratory phonology 7, eds. C. Gussenhoven and N. Warner, pp. 515-546, Mouton de Gruyter, Berlin, 2002.

[3] M. Komatsu and T. Arai, “Acoustic realization of prosodic types: Constructing average syllables,” LACUS Forum, vol. 29, pp. 259-269, 2003.

[4] M. Komatsu and H. Miyakoda, “Acoustic measurement of rhythm types: A stress language vs. a mora language,” in Linguistik International, Peter Lang, Bern, in press.

[5] T. Aoki, M. Komatsu, T. Arai, and Y. Murahara, “Temporal envelope modulation using syllable search method for robust language identification,” Proc. Forum Acusticum Sevilla 2002 [CD-Rom], Sevilla, Spain, Sept. 2002.

[6] T. Arai and S. Greenberg, “The temporal properties of spoken Japanese are similar to those of English,” Proc. Eurospeech ’97, pp. 1011-1014, Rhodes, Greece, Sept. 1997.

[7] S. Kitazawa, Japanese MULTEXT [CD-ROM], Shizuoka University, Hamamatsu, Japan, 2004.

[8] P. Boersma and D. Weenink, Praat (ver. 4.0.18) [Software], 2002.

[9] M. Kenstowicz, Phonology in generative grammar, Blackwell, Cambridge, MA, 1994.

[10] M. Komatsu, “Essay on acoustic correlates of prosodic typology,” in A new century of phonology and phonological theory, eds. T. Honma, M. Okazaki, T. Tabata, and S. Tanaka, pp. 492-507, Kaitakusha, Tokyo, 2003.