音声研究　第13巻第3号
2009(平成21)年12月
85–89頁

Journal of the Phonetic Society
of Japan, Vol.13 No.3
December 2009, pp.85–89

Modulation Spectrum and Rhythmic Units of Japanese

Masahiko KOMATSU* and Takayuki ARAI**

日本語における変調スペクトルとリズムの単位

小松　雅彦*・荒井　隆行**

要旨：日本語発話における変調スペクトルとリズムを構成する単位の関係を調べた。変調スペクトルは，音声の単位の出現するタイミングと関係がある。英語の音節の方が日本語よりも長いにも関わらず，英語発話と日本語発話の変調スペクトルは類似している。「日本語 MULTEXT」から得られた変調スペクトルのピークは，4～5Hz であり，200～250ms の時間長に相当する。モーラの長さは短く，変調スペクトルのピークには対応しない。音節の長さは，ほとんどのものが短かったが，2モーラ以上から構成されている音節は200～250ms に分布していると推定された。連続する2モーラおよび2音節の長さは200～250ms であった。これらのことから，単独のモーラや音節ではなく，2モーラまたは2音節からなる脚が変調スペクトルのピークに対応していることが分かる。変調スペクトルと2モーラまたは2音節からなる脚の関係を示すことができた。2モーラまたは2音節の脚は，かなり一定の間隔で現れ，強さ曲線への影響が大きい。

Key words: Modulation spectrum, mora, syllable, foot, rhythm, Japanese MULTEXT

1. Introduction

The Japanese language is said to have mora-timed rhythm, but researchers do not agree on the definition of such rhythm (see Warner and Arai 2001). Some researchers have related the rhythm to the timing of certain speech units[1]. They assume that the mora is an isochronic unit (Han 1962) or that the length of a higher level structure such as a word is predictable from the number of morae in it (Port, Dalby and O'Dell 1987). The modulation spectrum is thought to be related to the timing of certain speech units, and it is these possible rhythmic units in Japanese speech which we seek to identify in this work.

The modulation spectrum is a 1/3-octave-band, long-term average spectrum obtained from intensity contours (Houtgast and Steeneken 1985), and its importance to speech intelligibility has been repeatedly emphasized. "The modulation spectrum reflects fluctuations in energy associated with articulatory dynamics ... during the production of speech" (Greenberg and Arai 2004). For example, the syllable nucleus generally has greater amplitude than the onset and coda, and hence the intensity rises and falls gradually over time, which is inherently reflected in the modulation spectrum.

The relation of the modulation with speech units has been discussed (Arai and Greenberg 1997, Greenberg and Arai 2004). It was found that the modulation spectra of English and Japanese speech have similar shapes and both have their peaks around 4Hz, although the durations of English syllables are longer than Japanese ones in general. In English, the relation of the modulation spectrum with the distribution of syllable types and durations has been investigated in detail: the central core of the modulation spectrum, namely, 4–5Hz, corresponds to the region where both accented and unaccented syllable durations converge. On the other hand, the relation of the modulation spectrum with speech units in Japanese speech has yet to be explored. Thus far, it has been suspected that units longer than one mora (e.g. bimoraic syllables including a long vowel, and coalescences of a usual mora and its adjacent devoiced mora) contribute to the modulation spectrum.

Research using TEMAX, a spectrographic analysis of amplitude contours, has proposed bimoraic feet as a rhythmic component in Japanese speech, having found two rhythmic components, corresponding to morae and bimoraic feet, respectively (Kitazawa,

* Associate Professor, School of Psychological Science, Health Sciences University of Hokkaido（北海道医療大学心理科学部准教授）
** Professor, Faculty of Science and Technology, Sophia University（上智大学理工学部教授）

Sugiura, Shitaoka and Kobayashi 1996, Ayusawa, Kitazawa and Toki 1998). Because of the technical commonality of the modulation spectrum and TEMAX, the findings observed with TEMAX might also be seen in the modulation spectrum.

2. Analysis

The analysis presented in this paper used all speech data contained in the speech corpus *Japanese MULTEXT* (Kitazawa 2004). This corpus contains speech files of short passages with time-aligned phonemic labels. Three male and three female speakers read 40 passages in two styles: reading and simulated spontaneous speech. In total, the corpus consists of 240 speech and label files (6 speakers × 40 passages) for each style. The total duration of the recording is 107min 52s for reading and 109min 11s for simulated spontaneous speech[2]. The 40 passages include 6,897 morae in total[3].

Fig. 1 shows the average modulation spectra for the read speech and simulated spontaneous speech. It can be seen that they have very similar shapes and both have their peaks around 4–5Hz, as expected from previous studies. The peak in simulated spontaneous speech is slightly higher than in read speech, probably because the simulated spontaneous speech is slightly faster.

Fig. 2 shows the distribution of the durations of morae classified into seven types. CV and CJV morae are, as expected, longer than the other types of morae, namely, morae consisting of a vowel only (V), special morae (N, Q), and morae whose vowel is devoiced (CV devoi, and CJV devoi)[4].

Fig. 3 shows the durations of the mora and larger units. The *mora* in Fig. 3 includes all the mora types mentioned above. A *phonological syllable* refers to a unit that consists of a mora optionally followed by vowel-only morae (V) and/or special morae (N, Q)[5]. A

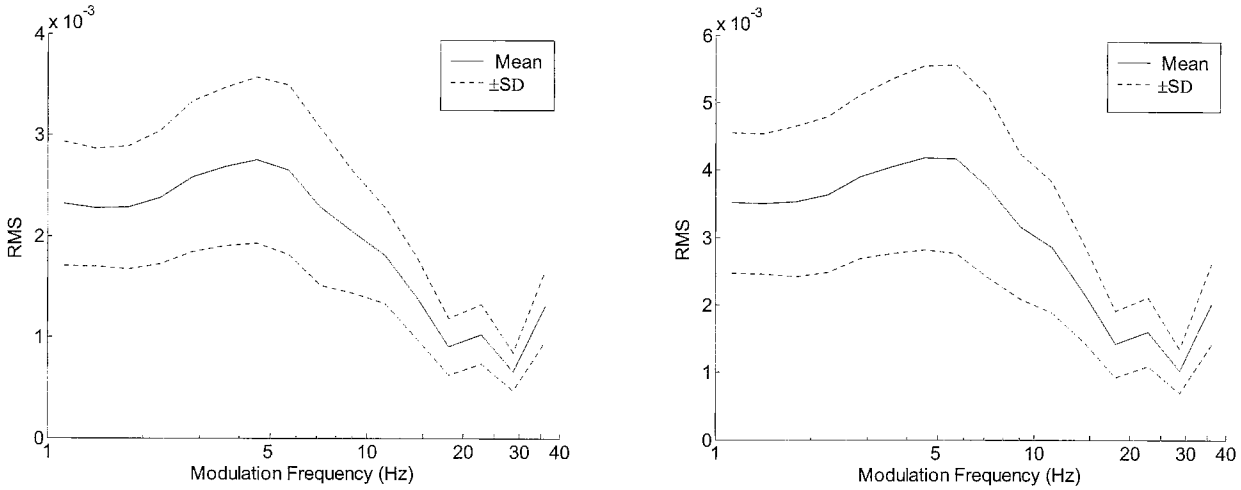

Fig. 1 Modulation spectrum (Left: Read speech; Right: Simulated spontaneous speech)

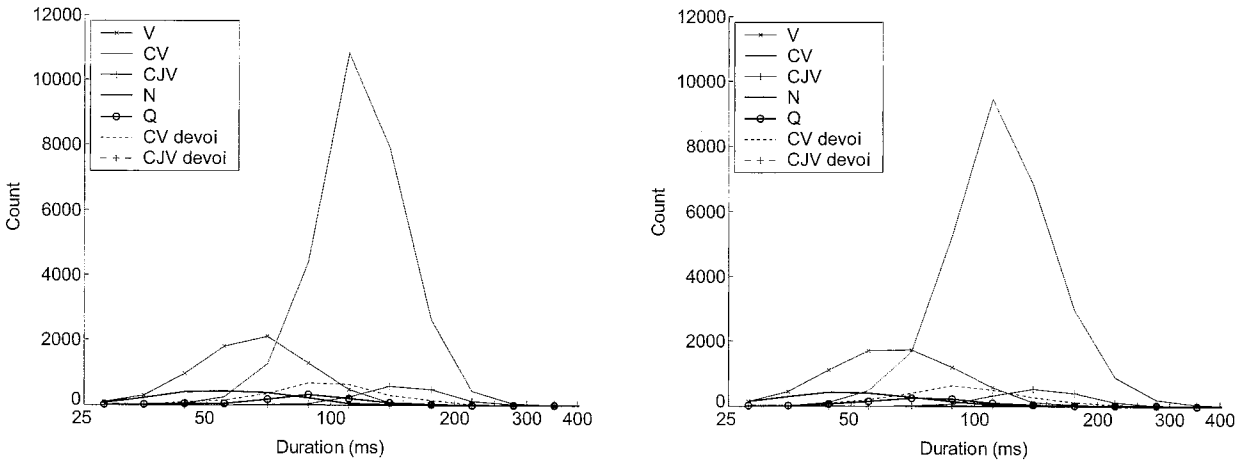

Fig. 2 Duration of each type of morae (Left: Read speech; Right: Simulated spontaneous speech)

phonetic syllable may be a phonological syllable alone, or it can also include devoiced morae (CV devoi, CJV devoi) as prefixes, and/or as suffixes before a pause. In other words, a phonological or phonetic syllable may consist of one mora, or it may include one or more additional non-C(J)V or devoiced morae.

The peaks of the distributions of the mora, phonological syllable, and phonetic syllable are at the same duration. This indicates that most of the phonological and phonetic syllables consist of only one mora.

The "bumps" in the distributions of the phonological and phonetic syllables around 200–250ms in read speech (left graph in Fig. 3) suggest that additional morae maintain their durations, which might contribute to the 4–5Hz peak seen in the modulation spectrum. On the other hand, the slopes are more gradual in simulated spontaneous speech, which suggests that the durations of the additional morae may be shortened.

Fig. 4 shows the duration of two successive morae or phonological/phonetic syllables. All of the distributions have their peaks within or around 200–250ms, which corresponds to 4–5Hz.

3. Discussion

The shape of the modulation spectrum is related to the timing at which local peaks in the intensity contour occur in the speech signal. In particular, the peak in the modulation spectrum is thought to correspond to the units that repeatedly occur at the same interval. In a previous study (Greenberg and Arai 2004), the central core of the English modulation spectrum was demonstrated to be at 4–5Hz, which corresponds to 200–250ms, where both accented and unaccented syllables converge.

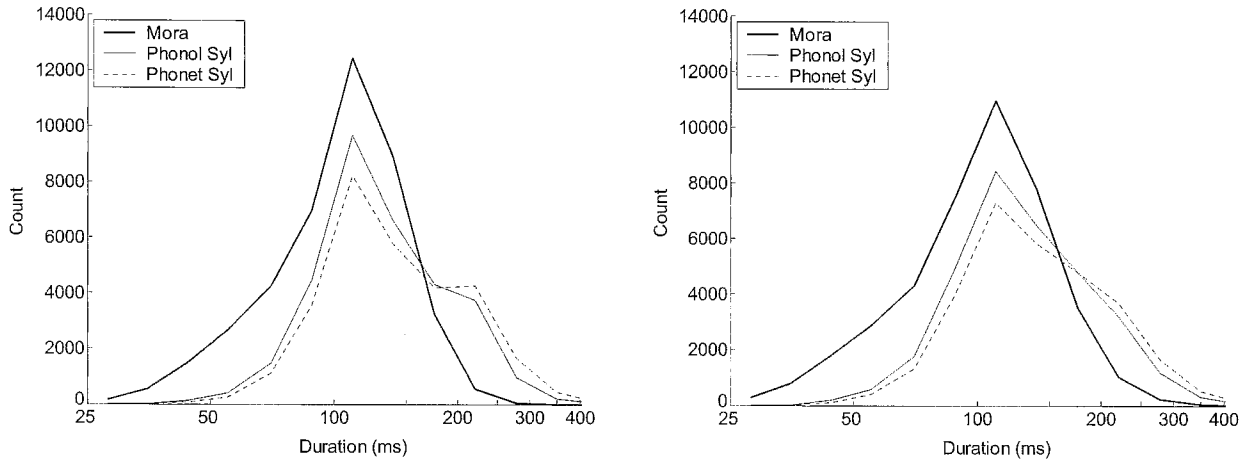

Fig. 3 Duration of morae, phonological syllables, and phonetic syllables (Left: Read speech; Right: Simulated spontaneous speech)

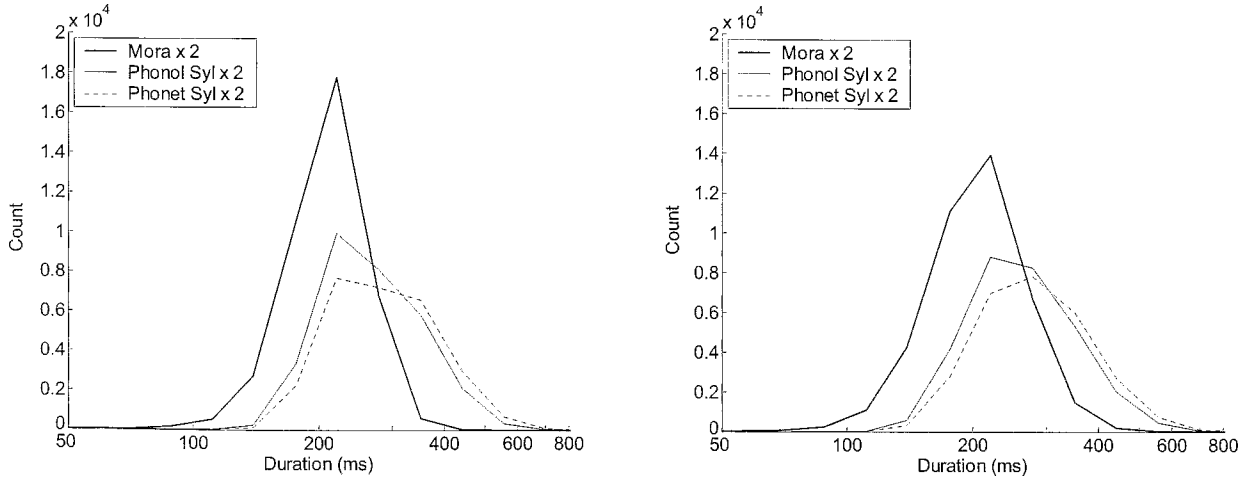

Fig. 4 Duration of two successive morae, phonological syllables, and phonetic syllables (Left: Read speech; Right: Simulated spontaneous speech)

The modulation spectra obtained from the Japanese data here also showed their peaks around 4–5Hz (Fig. 1), which corresponds to the duration of 200–250ms. In the analysis, no types of morae showed 200–250ms peaks (Fig. 2). The peaks of the distributions of phonological and phonetic syllables were not seen at 200–250ms, but those consisting of more than one mora were inferred to be distributed around 200–250ms (Fig. 3). The peaks of two successive morae and phonological/phonetic syllables were found at 200–250ms (Fig. 4). Taken together, these results indicate that bimoraic or bisyllabic feet, rather than single morae or syllables, correspond to the peak in the modulation spectrum.

Put another way, bimoraic or bisyllabic feet have greater effects on intensity contours than single morae or syllables. This corresponds to the implications of previous studies. Arai and Greenberg (1997) emphasized the importance of "syllables" consisting of more than one mora. While they do not use the term *foot* specifically, some of their "syllables" correspond to *feet* as used in this paper. Ayusawa, Kitazawa and Toki (1998) regard feet as rhythmic units, and they also observed that not only diphthongs and long vowels but two-mora coalescence often creates a large bump in the amplitude contour because the dip between the morae is small (p. 40).

This paper describes the rough relationship between the modulation spectrum and bimoraic or bisyllabic feet. These feet occur repeatedly at a rather constant interval and have considerable effects on the intensity contours.

Acknowledgments

This paper is a revised version of Section 3 of Komatsu and Arai (2006). The authors are grateful for the comments from two anonymous reviewers. The research is partially supported by KAKENHI 20242010 from the Japan Society for the Promotion of Science and a Grant-in-Aid for the 2009–2010 Research Project of the Research Institute of Personalized Health Sciences, Health Sciences University of Hokkaido.

Notes

1) The rhythm of Japanese has been approached from various aspects. Others regard the rhythmic difference among languages as a reflection of structural factors, such as syllable structures and the like, rather than timing specifically (Ramus, Nespor and Mehler 1999). Also, some discuss the human competence of coordinating units in speech production (Tajima 1998) and the role of the unit in perception (Cutler and Otake 1997).
2) The duration of the recording is cited from Kitazawa, Kitamura, Mochiduki and Itoh (2001).
3) The number of morae was counted by the author, which is different from the count given in Table 2 in Kitazawa, Kitamura and Itoh (2002).
4) The number of *CJV devoi* is very small, and its distribution is not clearly seen in Fig. 2.
5) The durations of phonological syllables were calculated automatically without considering phonological and morphological constraints. For example, VV sequences which are usually regarded as vowel hiatuses rather than diphthongs in the phonology literature were counted as constituting one syllable.

References

Arai, Takayuki and Steven Greenberg (1997) "The temporal properties of spoken Japanese are similar to those of English," *Proceedings of Eurospeech* '97, 1011–1014.

Ayusawa, Takako, Shigeyoshi Kitazawa and Satoshi Toki (1998) "A study of Japanese rhythm by means of TEMAX and its application in language teaching," *Journal of the Phonetic Society of Japan* 2: 1, 34–40 (In Japanese).

Culter, Anne and Takashi Otake (1997) "Contrastive studies of spoken-language perception," *Journal of the Phonetic Society of Japan* 1: 3, 4–13.

Greenberg, Steven and Takayuki Arai (2004) "What are the essential cues for understanding spoken language?" *IEICE Transactions on Information and Systems* vol. E87-D, no. 5, 1059–1070.

Han, Mieko S. (1962) "The feature of duration in Japanese," *Studies of Sounds* 10, 65–75.

Houtgast, T. and H.J.M. Steeneken (1985) "A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria," *Journal of the Acoustical Society of America* 77, 1069–1077.

Kitazawa, Shigeyoshi (ed.) (2004) *Japanese MULTEXT* [CD-ROM]. Shizuoka University, Hamamatsu, Japan.

Kitazawa, Shigeyoshi, Tatsuya Kitamura and Toshihiko Itoh (2002) "Nihongo MULTEXT ni okeru inritsu jōhō no bunseki to shūroku." In *Proceedings for 2001 2nd Plenary Meeting and Symposium on Prosody and Speech Processing.* (pp.39–50) Tokyo (In Japanese).

Kitazawa, Shigeyoshi, Tatsuya Kitamura, Kazuya Mochiduki and Toshihiko Itoh (2001) "Preliminary study of Japanese MULTEXT: A prosodic corpus." In *Proceedings of International Conference on Speech Processing.* (pp.825–828) Taejon, Korea.

Kitazawa, Shigeyoshi, Hisao Sugiura, Daisuke Shitaoka and Satoshi Kobayashi (1996) "Studies with the TEMAX for

assessment of the rhythmic components in speech," *IEICE Technical Report,* SP 96 (364), 51–58 (In Japanese).

Komatsu, Masahiko and Takayuki Arai (2006) "Tsuyosa kyokusen to henchō supekutoru no nihongo onsetsu to no kankei," *Proceedings of the 20th General Meeting of the Phonetic Society of Japan,* 195–200 (In Japanese).

Port, Robert F., Jonathan Dalby and Michael O'Dell (1987) "Evidence for mora timing in Japanese," *Journal of the Acoustical Society of America* 81, 1574–1585.

Ramus, Franck, Marina Nespor and Jacques Mehler (1999) "Correlates of linguistic rhythm in the speech signal," *Cognition* 73, 265–292.

Tajima, Keiich (1998) *Speech rhythm in English and Japanese: Experiments in speech cycling.* Doctoral dissertation, Indiana University, Bloomington.

Warner, Natasha and Takayuki Arai (2001) "Japanese mora-timing: A review," *Phonetica* 58, 1–25.

(Received Sep. 3, 2009, Accepted Apr. 8, 2010)