Presented at Eurospeech-97, Rhodes, vol. 2 pp. 1011-1014.

THE TEMPORAL PROPERTIES OF SPOKEN JAPANESE ARE SIMILAR TO THOSE OF ENGLISH

Takayuki Arai and Steven Greenberg

International Computer Science Institute
1947 Center Street, Berkeley, CA 94704 USA
and University of California, Berkeley
{arai,steveng}@icsi.berkeley.edu

ABSTRACT

The languages of the world are generally classified into two types on the basis of their segmental timing. "Syllable-timed" languages, such as Japanese, are considered isochronous, exhibiting a highly regular pattern of syllabic duration. In contrast are the "stress-timed" languages, such as English, whose syllable timing varies greatly, both within and across sentential domains. The present study demonstrates that, even in a language as theoretically isochronous as Japanese, the duration of syllabic segments is as variable as their English counterparts. Moreover, the variability of moraic duration is as high as that observed for syllabic units. Two measures of segmental timing, syllable duration and the low-frequency modulation spectrum, indicate that the coarse temporal characteristics of English and Japanese are remarkably similar. Such common properties may reflect inherent temporal characteristics of physiological mechanisms underlying the production and perception of speech that are shared by all languages of the world.

1. INTRODUCTION

Japanese is widely considered to exemplify a "syllable-timed," isochronous pattern of segmental duration (e.g., [12]). Even among Japanese, the timing of their spoken language is thought to be inherently regular. Underlying this autochthonous notion of isochrony is the *mora,* a segment grounded in the orthography of the Japanese language. In spoken discourse, the mora generally assumes one of three basic forms:□|V| (vowel), |CV| (consonant□+ vowel) and |CjV| (where j designates a glide). In addition, there are three less common forms, |N| (nasal), |Q| (geminate stop) and |V: | (vocalic lengthening) that are associated with the large-scale incorporation of Chinese loan words into the contemporary Japanese lexicon. The mora functions, in some abstract fashion, as a phonological "primitive," acting as a simplified, highly reduced version of the syllable.

A syllable in Japanese generally consists of either one or two morae, although some may contain as many as four morae under certain circumstances. The vocalic portions of many moraic segments are either devoiced or highly reduced. As a consequence, the consonantal components of such mora are often left "standing free" and frequently attach to an adjacent consonantal segment to form units that are essentially syllabic in nature. In addition, such morae as |N|, |Q| and |V: | typically adjoin vocalic segments. This highly intricate pattern of moraic integration makes the partitioning of Japanese speech into syllabic (or moraic) segments a challenging exercise in phonetic transcription and segmentation.

English is considered to exemplify the temporal pattern exhibited by "stress-timed" languages. The theoretical unit of isochrony in such languages is the metrical foot, which in principle, consists of a syllabic pair, one significant longer than the other.

A comparison of the segmental durational properties of Japanese and English would not, in theory, be expected to yield a similar pattern of temporal organization. And yet, as the following analyses indicate, the syllabic timing of these two historically unrelated languages are remarkably similar. The similarity in timing is not confined to syllabic segments, but extends to a more basic measure of timing derived from the modulation spectrum of the acoustic signal.

2. SPOKEN LANGUAGE MATERIALS

The Japanese language materials used in our analyses form a subset of the OGI-TS corpus of spontaneous, informal speech spoken over the telephone by each of 90 native speakers, discussing a topic of their choosing for approximately 50 seconds [11]. Each monologue was carefully transcribed at the phonetic-segment, moraic and syllabic levels by the senior author, a phonetically trained, native speaker of Japanese. Each speech file was carefully edited to eliminate filled pauses, hesitations and other instances of significant interruption in the speech stream. This editing reduced the length of the average monologue to approximately 30 seconds. The segmentation was performed on a Sparc5 workstation, using Entropic software to concurrently display the speech waveform and spectrographic representation of each utterance. In addition, the phonetic- level transcription was provided in order to facilitate syllabic and moraic segmentation of the signal.

The English-language materials were derived from a small subset of the Switchboard corpus. This speech material consists of informal telephone conversations between two individuals discussing a specific, pre-designated topic for several minutes [4]. A portion of this corpus has been phonetically transcribed by a group of highly trained, phonetically experienced native speakers of American English [6,7]. The syllabic durations of this material were computed via a rule-based algorithm derived from a software implementation of Kahn's syllabification rules for English [10].

3. TEMPORAL PROPERTIES OF THE SYLLABLE AND MORA

The durational properties of the English material are discussed in detail elsewhere [6,7], and therefore are only briefly described in the present communication.

The mean syllabic duration for English is 190 ms. Sixty percent of the syllables fall between 106 ms (the 20th percentile) and 260 ms (the 80th percentile). The frequency distribution of the syllabic durations is shown in Figure 2.

The mean syllable duration for the Japanese corpus subset of 30 speakers is 166 ms (standard deviation = 73 ms). Table 1 illustrates the durational data for each of the major syllabic forms in Japanese, and compares these with the frequency of occurrence for certain syllabic forms in English (Switchboard corpus).

The distribution of Japanese syllable durations is illustrated in Figure 1(a). Both its general contour, as well as its mean, mode and median, are similar to that of the English material [6]. One quantitative index of the variability contained within these durational data is the coefficient of variation, which is computed by dividing the mean of the distribution by its standard deviation. The coefficient of variation for the Japanese syllable durations is 0.44, comparable to the coefficient of variation for the English material.

Syllable	n	%	Eng.%	Duration Mean (ms)	Duration S.D. (ms)
CV	3238	60.4	47.2	137.9	58.6
CVV	626	11.7		198.4	59.1
CVC	961	17.9	22.1	229.4	61.7
CVVC	71	1.3		274.9	54.3
VC	64	1.2	4.8	198.2	75.4
VVC	6	0.1		246.2	29.5
V	154	2.9	11.2	87.0	38.5
VV	29	0.5		148.3	37.6
CCV	89	1.7		197.8	63.2
CCVV	72	1.3		212.0	68.6
CCVC	19	0.4		264.5	51.6
CCVVC	8	0.1		255.2	28.8
other	21	0.3			
	5358	100.0	85.3	165.8	72.9

Table 1. Durational statistics for syllable classes (in terms of consonant/vowel composition) contained within the Japanese corpus. The proportion of syllables corresponding to equivalent classes in English (Switchboard corpus) is also indicated.

Mora	n	%	Duration Mean (ms)	Duration S.D. (ms)
V	1148	15.3	72.3	31.2
CV	5589	74.7	130.3	53.2
CjV	182	2.4	141.8	41.6
N	384	5.1	75.4	31.8
Q	183	2.4	123.7	51.7
	7486	100.0	118.7	54.5

Table 2. Durational statistics for moraic classes (in terms of consonant/vowel composition) of the Japanese corpus.

The corresponding durational data for the moraic segments contained within the Japanese corpus are shown in Table 2 and Figure 1(b). The mean moraic duration is 119 ms (standard deviation = 55 ms). The coefficient of variation, 0.46, is similar to that of the syllabic material.

The durational distribution for both moraic and syllabic segments is rather broad, suggesting that spoken Japanese may not be as isochronous as commonly thought. The variability inherent in their segmental duration is typical of the variability observed for English and other stress-timed languages.

The moraic and syllabic durational data are partitioned into coarse phonological classes in Tables 1 and 2. Japanese morae are of six potential types, as shown in the second table. One form, that of vocalic lengthening |V:| is merged with the moraic form |V| in the current analysis. The overwhelming majority of morae (90%) are either of the consonant-vowel |CV| or vocalic |V| variety.

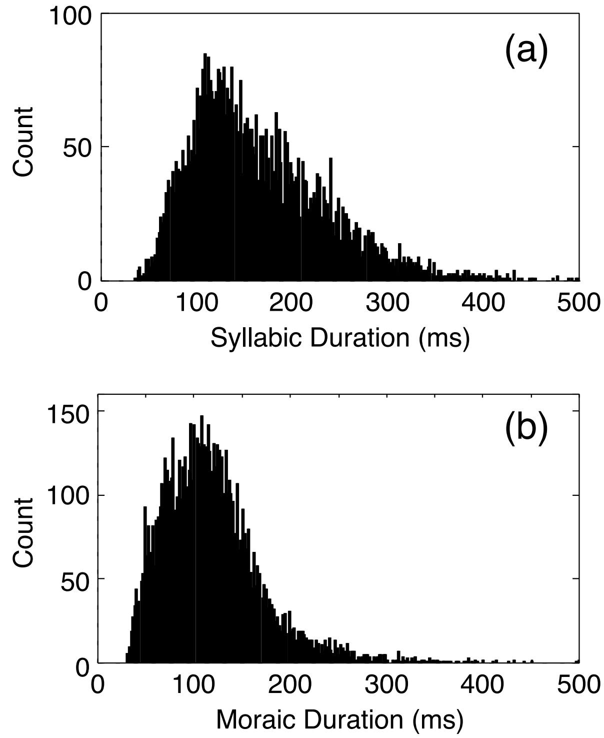

Figure 1. The distribution of syllabic (a) and moraic (b) durations for fifteen minutes of spontaneous Japanese speech spoken by thirty individual native speakers.

Mora	n	%-Devoiced
/shi/	263	74.5
/su/	238	79.0
/ku/	189	43.9
/ki/	128	39.1
/chi/	94	42.6
/tsu/	77	50.6
/hi/	32	75.0
/hu/	26	73.1

Table 3. Frequency of devoicing for a small subset of moraic classes within the Japanese corpus.

The durations of these moraic constituents range between 72 ms for the vocalic type to 142 ms for the CjV form. This variability in moraic durations provides some insight in the manner in which such segments coalesce into generally longer syllabic elements. Certain moraic elements, such as |CV| and |V|, typically merge into longer syllabic entities, such as |CVV|. In this instance, the coalescence appears as a form of vowel lengthening, similar to the process described above. In addition, certain vocalic components of these morae tend to either devoice or reduce, resulting in syllables composed of consonant clusters (e.g., |CV| + |CV| > |CCV|). Through this pattern of consonantal coalescence and vocalic reduction/devoicing the relatively simple, transparent phonological patterns characteristic of the mora are transformed into the phonologically more complex syllabic forms typical of spontaneous Japanese.

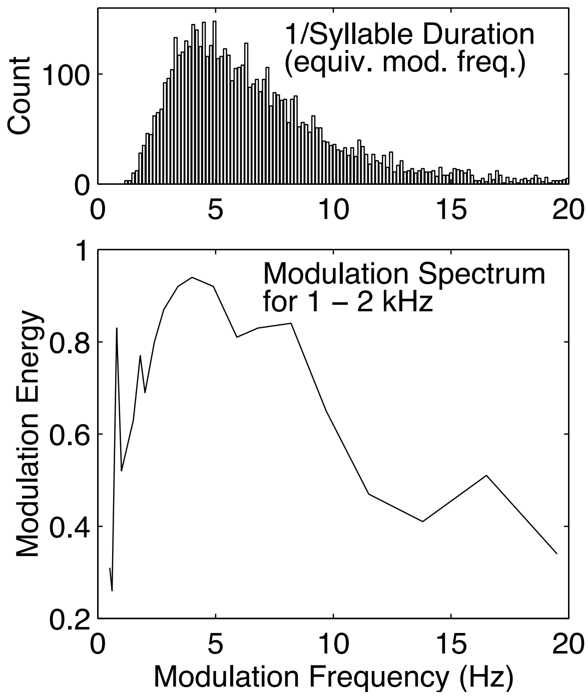

Figure 2. Modulation spectrum and frequency histogram of syllabic durations for spontaneous English discourse (reprinted from [6]). Top panel: histogram pertaining to 2925 syllabic segments from the Switchboard corpus. Bottom panel: modulation spectrum for two minutes of connected, spoken discourse from a single speaker.

This highly complex process is responsible for the non-isomorphic relationship between mora and syllable in Japanese.

This complex relationship between syllable and mora is illustrated for a subset of morae in Table 3. The frequency of vocalic devoicing ranges from 39 to 79 percent. Each instance is associated with a merging of the preceding consonantal segment with the adjacent consonant to form a phonologically more complex syllabic form, resulting in a temporal pattern that is similar to spoken English at this segmental level.

4. MODULATION SPECTRAL ANALYSIS

The low-frequency (2-8 Hz) component of the modulation spectrum plays an important role in the intelligibility of speech [3,9], both for English [5] and Japanese [1]. Its vital role for speech understanding may be a consequence of its isomorphic relationship with the syllabic segmentation of speech [6], an instance of which is illustrated in Figure 2.

The modulation spectrum for a 15-minute portion of the Japanese corpus, shown in Figure 3, is similar to that of English. Moreover, it exhibits the same isomorphic relation to the distribution of syllabic durations (also illustrated in Figure 3, transformed into units of equivalent modulation frequency) as typifies spoken English.

The similarity in patterns of segmental duration and modulation spectra observed in languages as typologically distinct as English and Japanese, suggest that these measures may reflect a universal property of spoken language, that is largely independent of a language's intrinsic phonological structure.

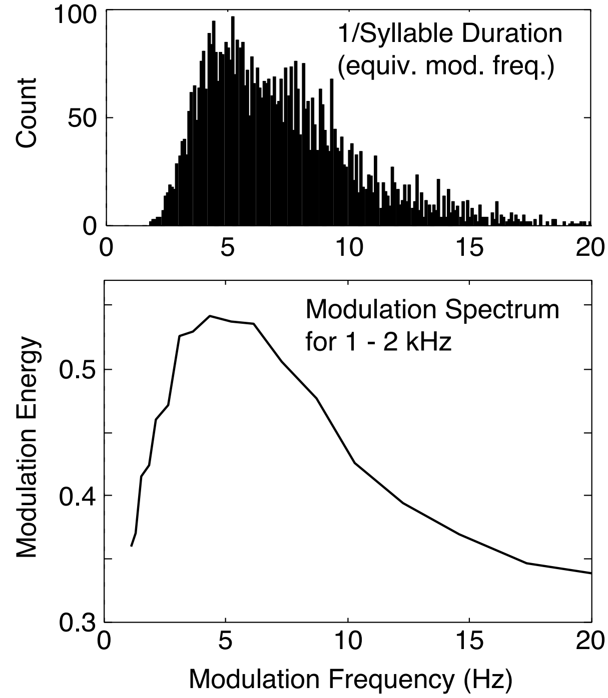

Figure 3. Modulation spectrum and frequency histogram of syllabic durations for the OGI-TS Japanese corpus. Top panel: histogram of durations for 5358 syllables, derived from manual demarcation of syllabic boundaries. Durational data are identical to those in Figure 1(a) but plotted in terms of equivalent modulation frequency. Bottom panel: modulation spectrum of the same speech material as illustrated in the top panel. Details concerning the modulation spectral analysis technique are described in [8].

5. CONDITIONAL DURATIONS FOR SYLLABIC AND MORAIC SEGMENTS

The modulation spectrum and associated patterns of syllabic duration reflect temporal variability over a relatively long span of the Japanese corpus (ca. 15 minutes). In principle, shorter intervals of speech could exhibit a greater degree of isochrony than observed in the corpus as a whole.

The first-order, conditional duration was computed for moraic and syllabic segments in the Japanese corpus. In such an analysis, the duration of the reference ("current") segment is plotted as a function of the previous segment's duration. In a truly isochronous pattern, the conditional durations should fall into a very small region of the data space, since the variance of the segmental durations is theoretically small.

Figure 4 illustrates the conditional durations for moraic and syllabic segments. In each instance, the coordinates of the conditional durations are dispersed over a broad area. Such a pattern of dispersion is not consistent with the theoretical characterization of segmental isochrony, and reinforces the conclusion that the variability of Japanese segmental durations is as high for short spans of speech as it is for longer intervals of time. A similar pattern of conditional durations is observed for syllabic segments in English [7].

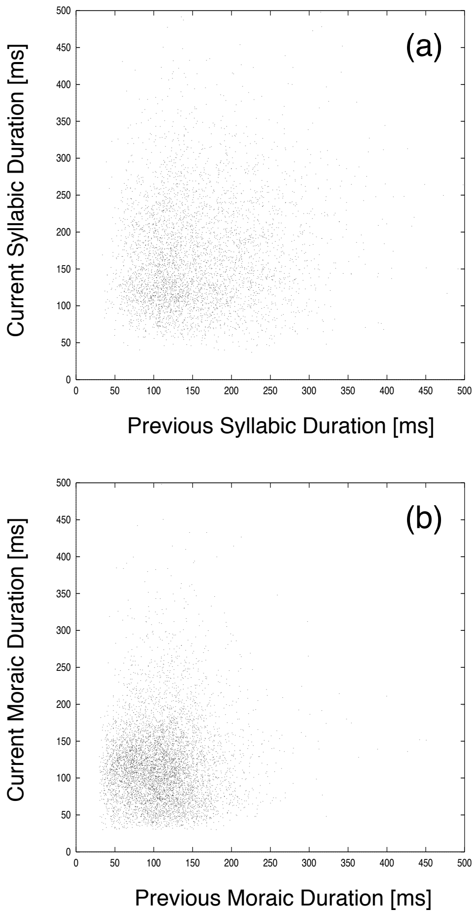

Figure 4. Conditional dependence of segmental duration for contiguous syllables and morae in the Japanese corpus.

6. DISCUSSION AND CONCLUSIONS

Despite broad differences between the phonological patterns of Japanese and English, the temporal properties of their spoken form are remarkably similar. Japanese believe that the timing of their spoken language is inherently regular, based on the *abstract* temporal properties of the mora.

However, this perception of temporal regularity appears to be an illusion, at least in terms of the physical duration of moraic segments in spontaneous speech. Since the mora is derived from an orthographic feature of the language, one would anticipate at least some degree of isochrony in the speech tempo of text-derived spoken Japanese. However, durational analysis of written Japanese, carefully spoken by native speakers indicates that, even under such ideal conditions, the speech tempo of this "mora-timed" language is far from isochronous [2].

Thus, languages as typologically distinct as Japanese and English appear to share many temporal properties in common. The broad distribution of syllabic durations is common to both English and Japanese, and this variability is reflected in the broad bandwidth of the modulation spectrum. Such temporal properties are likely to reflect fundamental constraints imposed by articulatory and perceptual mechanisms common to all languages and bear a potentially important relation to the ability to understand speech spoken under the broad range of acoustic conditions typical of the real world [5,8].

ACKNOWLEDGMENTS

Brian Kingsbury developed the program for computing the modulation spectrum, for which we are grateful. We also express our appreciation to Dan Ellis for his adaptation of Bill Fisher's program for automatic syllabification of phonetic/phonological input strings, as well as to John Ohala and Natasha Warner for their comments on an earlier version of this paper.

REFERENCES

[1] Arai, T., Pavel, M., Hermansky, H., Avendano, C. (1996) Intelligibility of speech with filtered time trajectories of spectral envelopes, *Proceedings of the International Conference on Spoken Language Processing,* pp. 2490-2493.

[2] Beckman, M. (1982) Segment duration and the "mora" in Japanese, *Phonetica*, 39, 113-135.

[3] Drullman, R., Festen, J. M., and Plomp, R (1994) Effect of temporal envelope smearing on speech reception, *J. Acoust. Soc. Amer.*, 95, 1053-1064.

[4] Godfrey, J. J., Holliman, E. C. and McDaniel, J. (1992) SWITCHBOARD: Telephone speech corpus for research and development, *ICASSP-92, IEEE International Conference on Acoustics, Speech and Signal Processing*, 1, pp. 517-520.

[5] Greenberg, S. (1996) Understanding speech understanding: towards a unified theory of speech perception, *Proceedings of the ESCA Tutorial and Advanced Research Workshop on the Auditory Basis of Speech Perception*, W.A. Ainsworth and S. Greenberg (eds.), Keele University, UK, pp. 1-8.

[6] Greenberg, S., Hollenback, J. and Ellis, D. (1996) Insights into spoken language gleaned from phonetic transcription of the switchboard corpus, *Proceedings of the International Conference on Spoken Language Processing*, pp. S24-27.

[7] Greenberg, S. (1997) The Switchboard Transcription Project, Technical Report, 1996 *Johns Hopkins CLSP Workshop on Innovative Techniques for Large Vocabulary Continuous Speech Recognition*, Baltimore, MD.

[8] Greenberg, S. (1997) On the origins of speech intelligibility in the real world, *Proceedings of the ESCA Workshop on Robust Speech Recognition for Unknown Communication Channels*, Pont-a-Mousson, France.

[9] Houtgast, T., Steeneken, H. J. M. (1985) A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria, *J. Acoust. Soc. Am.*, 77, 1069-1077.

[10] Kahn, D. (1976) Syllable-based Generalizations in English Phonology. Bloomington: Indiana University Linguistics Club (Ph.D. Thesis).

[11] Muthusamy, Y. K., Cole, R. A. and Oshika, B. T. (1992) The OGI multi-language telephone speech corpus, *Proceedings of the International Conference on Spoken Language Processing*, pp. 895-898.

[12] Vance, T. J. (1987) *An Introduction to Japanese Phonology*, State University of New York Press: Albany.