19

Speech rhythm in English and Japanese

KEIICHI TAJIMA AND ROBERT F. PORT

19.1 Introduction

This paper has two main objectives. First, we propose that cross-linguistic variation in speech rhythm is not phonetically manifested simply as acoustic isochrony, but rather as *relative temporal stability* of syllables – i.e. the tendency for certain syllables to occur at particular points in time despite other factors that oppose this tendency. Second, we compare two languages believed to be rhythmically distinct, English and Japanese, and demonstrate that they not only display reliable rhythmic differences, but also striking similarities in the phonetic manifestation of foot-level structure. These goals are accomplished by using speech cycling – an artificial speaking task in which subjects repeat phrases in time with periodic auditory stimuli.

19.1.1 Rhythmic typologies

There has long been an intuition that languages are spoken with different kinds of rhythm. Conventionally, languages have been classified as either 'stress-timed' or 'syllable-timed' (Jones 1918; Pike 1945; Abercrombie 1967), depending on whether it is interstress intervals or intersyllable intervals that are regular. The prediction about timing that is generally seen to follow from 'stress-timed', 'syllable-timed', or even 'mora-timed' (Port, Dalby and O'Dell 1987), is perfect isochrony (equal time intervals) of stresses, syllables or moras. Not surprisingly, however, perfect isochrony has proven to be a difficult test to

satisfy, at least in naturalistic speaking styles (Lehiste 1977; Dauer 1983; Couper-Kuhlen 1993).

In generative phonology, a way of characterizing the rhythm of so-called stress-accent languages has evolved based on the quasi-periodic alternation of strong and weak syllables (Liberman and Prince 1977; Selkirk 1984). Metrical phonology uses metrical grids to formally represent the serial ordering of syllables varying in prominence. However, serially ordered steps by themselves imply nothing about their production in real time. Metrical theory can therefore offer no account of how these representations are to be phonetically interpreted. The simplest hypothesis for predicting specific time intervals would be regular timing of grid symbols (plus noise), but this is unlikely to be empirically supported. In fact, how the grid symbols are interpreted in real time was not of direct concern in many versions of metrical theory. Rather, the appeal of metrical grids stemmed from their being better able than metrical trees to explain certain rhythmic phenomena such as stress retraction (cf. Hogg and McCully 1987).

The role of metrical grids in these later versions of metrical theory is, ironically, fundamentally different from the original conception of metrical grids by Liberman (1975). In Liberman's proposal, grids were directly interpretable in real time; they partitioned time into hierarchically related intervals, much like musical notation. Liberman also made a fundamental distinction between *metrical grids*, which are hierarchically related time intervals, and *metrical patterns*, which are abstract properties of the linguistic text. The abstract patterns of strong and weak syllables in the text were interpreted in real time via alignment of the metrical patterns with the metrical grid. As will be discussed later, the conception of metre and speech presented here does better justice to Liberman's theory of metrical structure than do later versions of metrical phonology.

Rhythmic characteristics of so-called pitch-accent languages have often been investigated using verse-like texts (Lehiste 1990; Sakano 1996). In Japanese, for example, Bekku (1977) has claimed that, although the traditional *haiku* verse form employs an odd number of moras in each line (5, 7, 5), it is orally produced with a musical rest inserted at the end of each line so that the time interval between successive lines contains eight abstract mora-sized units. Homma (1991) in fact found a tendency for a longer pauses to be produced after short (5-mora) lines than after long (7-mora) lines in oral productions of *haiku*. A related claim about Japanese rhythm is that Japanese speakers tend to parse phrases into bimoraic feet, or units of two moras in length (Sakano 1996).

What seems necessary for many of these claims is phonetic evidence. To better examine how rhythmic structure is phonetically interpreted, we suspect that it will prove fruitful to study more constrained styles of speech than spontaneous speech.

19.1.2 Speech cycling

In recent years, we have been developing a method for studying speech rhythm and its variation across languages. In this general class of tasks called 'speech cycling' (Cummins 1997; Cummins and Port 1998; Tajima 1998), subjects produce a short text fragment repeatedly, in time with metronome-like stimuli. In a typical study, subjects might produce a short phrase such as *beg for a dime*, in time with a metronome adjusted to yield speech at a comfortable rate. Speakers show a strong preference for a small number of rhythmic patterns, where *beg* begins each cycle (like the first downbeat of a musical measure), and *dime* occurs on (a) the second downbeat of a two-beat measure, or on (b) the second downbeat or (c) the third downbeat of a three-beat, waltz-like measure. (Readers are encouraged to try these three patterns for themselves.) If these patterns are expressed in terms of the *phase angle* of the timing of *dime* within the cycle from the previous to the next *beg*, then the preferred pattern is to produce *dime* at phase 1/3, 1/2, or 2/3. Speakers are most stable at just these three patterns even though they are given various target phases at which to produce *dime*. These constraints on speech timing are powerful enough that it is quite difficult (without practice) to stably produce other patterns. We interpret these locations to be literally attractors of the onsets of stressed syllables on the phase circle. The phenomena invite a dynamical system interpretation in terms of coupled oscillators (Kelso 1996; Large and Jones 1999). We expect that similar effects will be observable in any language, though we should expect the details to be modulated by properties of each language.

19.1.3 Speech rhythm as temporal stability

Four hypotheses underlie the design of the experiments:

(1) Phases of the repetition cycle such as 1/3 and 2/3, called *simple harmonic phases*, are attractors of prominent syllables. That is, prominent syllables will 'want' to line up at these phases and exhibit temporal stability.
(2) An abrupt onset of acoustic energy at the beginning of vowels is a likely time point to exhibit temporal stability. It is this *beat* near vowel onsets that tends to be stably located in time (see Section 19.2.1.2 below).
(3) Changing the structure of the text, by changing segments or syllables, should cause perturbations of timing in performance of speech cycling.
(4) Languages differ in their criteria for syllable prominence. Therefore, even with comparable texts, speakers of different languages are expected to differ in the patterns they find stable in speech-cycling tasks.

19.1.4 Predictions about English and Japanese rhythm

This study uses the speech-cycling method to induce overtly rhythmic forms of speaking. By imposing a particular metrical organisation onto speech, the method allows direct evaluation of how metre is phonetically interpreted, and how it is influenced by the linguistic structure of the text.

English and Japanese were compared by using roughly comparable text materials and the identical speech-cycling task. In English, prominence is tied closely to stress. Thus, we expect stressed syllables to be regularly timed and have their onsets (their beats) located at simple harmonic fractions (e.g. 1/3 and 2/3) of the repetition cycle. In Japanese, which has no stress, predictions are more problematic. It is likely that word onsets are prominent, especially if combined with word-initial pitch accents. It is also possible that the initial moras of bimoraic feet are prominent. These moras may therefore be attracted to simple harmonic phases.

To test temporal stability, the timing of syllables was given phonological perturbations by constructing phrases that were identical except for one difference in segmental or syllabic content. The timing of corresponding syllables was then compared between these minimally contrasting phrases.

19.2 Experiment 1

19.2.1 Methods

19.2.1.1 Design of text materials

Text materials in Experiment 1 systematically varied in the duration of certain syllables, through inversion of adjacent long and short vowels. Table 19.1 gives phonemic transcriptions of some of the test phrases. The phrases contained nonexistent but phonotactically plausible words. Each phrase contained five open syllables. There were 16 phrases in each language, consisting of four sets of four phrases that followed specific patterns of durational contrasts, illustrated in the upper part of Figure 19.1. Patterns A and B constitute a 'minimal pair'; they are identical except for the underlined portion, which contains an inversion of vowels in the third and fourth syllables.[1] Since the diphthong /aj/ is inherently longer than the vowel /ə/, the third and fourth syllables are inherently long–short in A but short–long in B. Phrases C and D are analogous, except that the vowel inversion occurs in the second and third syllables. Japanese phrases were constructed in the same manner, using low versus high vowels for the long versus short contrast.

Table 19.1 *Phonemic transcription of sample phrases in Experiment 1. Two sets out of the four are shown for each language*

Set	Pattern	English	Japanese
I	A	/ˈgow fɔr ˈbaj.gə ˈdej/	/ku.ɾo ba.ki da/
	B	/ˈgow fɔr ˈbʌ.gaj ˈdej/	/ku.ɾo bi.ka da/
	C	/ˈgow baj ˈgʌ.nəˑ ˈdej/	/ku.ba ki.ɾi da/
	D	/ˈgow.bə ˈgaj.nəˑ ˈdej/	/ku.bi ka.ɾi da/
II	A	/ˈbi ðə ˈkow.bə ˈgaj/	/ta.te ka.bi da/
	B	/ˈbi ðə ˈkʌ.bow ˈgaj/	/ta.te ki.ba da/
	C	/ˈbi.kow ˈbʌ.ɾəˑ ˈgaj/	/ta.ka bi.ʃa da/
	D	/ˈbi.kə ˈbow.ɾəˑ ˈgaj/	/ta.ki ba.ʃa da/

Even though the vowel inversion is the same for the A-B pair and the C-D pair, the inversion was designed to cross a foot boundary in the latter pair but not in the former. The expected prosodic foot structure of the phrases is depicted in the middle part of Figure 19.1. If prominent syllables are 'attracted' toward the downbeats of the waltz rhythm, then they should show greater resistance to temporal perturbation caused by segmental manipulations. That is, the effect of vowel inversion should be smaller in the C-D pair (in which the third syllable is prominent) than in the A-B pair (in which the fourth syllable is not prominent).

In English, the prominent syllables (σ́) in Figure 19.1 were stressed syllables, and the brackets corresponded to stress-feet. The brackets also corresponded to major word boundaries, so that the prominent syllables were also word-initial (e.g. [go for] [Byga] [Day]). In Japanese, the prominent syllables (σ́) were morpheme-initial syllables, and the brackets corresponded to morphemes. Each Japanese phrase consisted of two unaccented bimoraic morphemes followed by the copula *da* (e.g. [naka] [niwa] [da]).[2] The prominent syllables were also the initial syllables of bimoraic feet. That is, if Japanese speakers were to show a preference toward parsing each phrase left-to-right into groups of two moras, then the test phrases would be parsed as shown by the brackets in Figure 19.1.

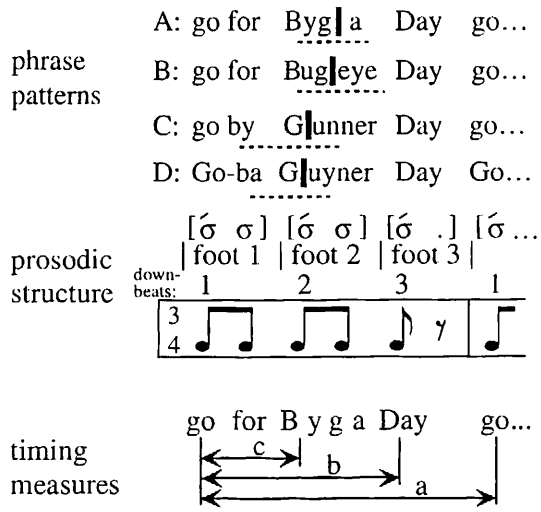

Figure 19.1 Top panel: phrase patterns A-D. Vowel inversion in the underlined portion should affect the fourth syllable beat for the A-B pair, and the third syllable beat for the C-D pair (as shown by the thick vertical lines). Middle panel: expected prosodic structure of the phrases and musical notation of the waltz rhythm. Since the third syllable is foot-initial, temporal displacement should be smaller in the C-D pair than in the A-B pair. Bottom panel: two relative measures of timing: (1) phase of a syllable relative to the cycle of successive repetition onsets (e.g. *c/a* and *b/a*), and (2) timing of a syllable relative to the first and last beats of a single repetition (e.g. *c/b*)

19.2.1.2 Procedure and measurement

Fourteen native speakers of American English and thirteen native speakers of Tokyo Japanese participated in a speech-cycling task in which they were instructed to produce the odd-numbered syllables at metrically strong positions (downbeats) of the repetition cycle. Subjects listened to an isochronous series of 50 ms, 600 Hz pure tones, presented through headphones. The metronome period was 1200 ms for English, and either 1100 ms or 1000 ms for Japanese. The following instructions were given:

> On each trial, listen to the first four beeps. Start repeating the phrase on the fifth beep, aligning the beginning of the phrase with each successive beep. Repeat the phrase rhythmically, using a waltz-like, three-beat rhythm. That is, repeat the phrase while keeping in mind the rhythm '1.. 2.. 3.. 1.. 2.. 3..', with the 1s falling on the beeps. Stop after eight repetitions. Do not insert breaths between repetitions.

Test phrases were presented using the native orthographic system of the language. Japanese phrases were written in *kanji* and *kana* in such a way that the

written forms suggested the expected morphological and accentual patterns. After familiarisation with the text materials, subjects were given as much feedback and practice as necessary to perform the task before starting the test trials. The 16 phrases were presented in a random order. A total of 432 trials were analysed [16 phrases X 27 subjects].

Measurement of the recorded utterances began by first converting the speech signal into a series of *beats*, by placing a beat near the vowel onset of each syllable. This was done semi-automatically, using a 'beat extractor' that created a rectified smoothed energy profile of the formant frequency region of the speech signal, and placed a marker (beat) halfway up every local rise in this contour (see Cummins and Port 1998 for details). Visual inspection was used to correct the output of the algorithm. The series of beats over time was used to capture the timing of vowel onsets in each trial.

Measurement based on beats contrasts with conventional measures of syllable and segment durations. For example, in a CVCVCV... sequence, instead of measuring syllable durations starting from the acoustic onset of a consonant up to the offset of a vowel (i.e. [CV][CV][CV]), these beats delimit syllables from vowel onset to vowel onset (i.e. C[VC][VC]V). This measure was adopted because abrupt rises in acoustic energy (as in a vowel onset) are correlated with more activity in the auditory nerve than are less abrupt rises or acoustic offsets (as in a consonant onset) (Delgutte 1982). Also, for syllables with voiced stops and nasals, the vowel onset is roughly equal to the so-called 'perceptual moment of occurrence' or P-centre, of the syllable (e.g. Scott 1993).

Of the eight repetitions in each trial, the first two and the last were discarded to reduce transient effects. Beats from the remaining five repetitions were converted into two relative measures of timing, illustrated in the lower part of Figure 19.1. *Phase* measures the time of occurrence of a syllable relative to the cycle starting at the initial beat of the current repetition and ending at the initial beat of the following repetitions (e.g. *b/a* or *c/a* in Figure 19.1). Additionally, we measured the time of occurrence of a syllable relative to the interval between the first and last syllable beats of the current repetition (e.g. *c/b*).

19.2.2 Results

19.2.2.1 Production of waltz rhythm

To examine how well the subjects as a group maintained the three-beat waltz rhythm, Figure 19.2 plots histograms of the timing of the third and fifth syllables of each phrase. Data are collapsed across patterns A-D. Given that phase is defined in the scale {0, 100}, a perfect waltz rhythm would imply that the third and fifth syllables would begin at phases 33.3 and 66.7, as shown by the vertical lines.

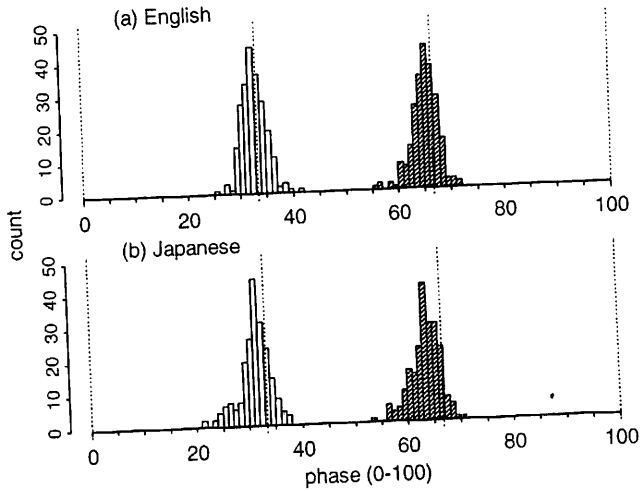

Figure 19.2 Histogram of phase of the third (unfilled bars) and fifth (filled bars) syllable beats from all trials in English (top) and Japanese (bottom)

Each cluster is based on trial means across all phrases and speakers. Trial means were calculated by averaging the observed phases of each syllable beat across the five repetitions measured in each trial. The figure shows that each cluster has a single distributional mode, with very little variability in the data. Performance was therefore highly consistent across subjects and phrases in both languages. Impressionistic judgement also suggested that speakers accurately produced the waltz rhythm.

19.2.2.2 Temporal stability of foot-initial syllables

To assess whether foot-initial and foot-internal syllables differ in temporal stability, Figure 19.3 shows the timing pattern of one matched set of English phrases, averaged across speakers. The x-axis here shows the percentage of the interval from the initial to final syllables of the phrase. This measure was adopted because the relative timing of the final syllable tended to vary across phrases. The vertical line at 50% corresponds to the second downbeat of the waltz, and is the expected phase of the third syllable. For each phrase, the vertical lines indicate the timing of the syllable beats, averaged across speakers.

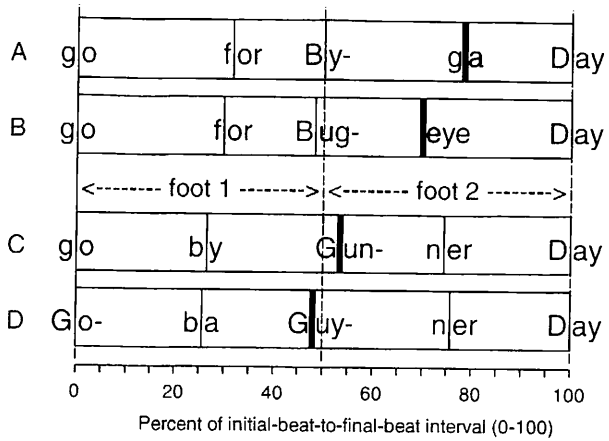

Figure 19.3 Timing of syllable beats in one matched set of English phrases. Each beat is the mean of trial means from all 14 American speakers. The x-axis shows the percentage of the interval from phrase-initial to phrase-final beats. The vertical lines at 0, 50 and 100% correspond to the three downbeats of the waltz. The thick lines are the target beats in question. Timing difference of these beats between A and B and between C and D are shown in the following figure

The thick lines in Figure 19.3 are the target beats; duration of the vowels before and after the target beats was manipulated. In the A-B pair, the fourth syllable beat is temporally displaced as a result of the inversion of the third and fourth vowels. In the C-D pair, a similar displacement is found for the third syllable, but the displacement is smaller in magnitude in this pair than in the A-B pair, as predicted.

Figure 19.4 directly compares the magnitude of displacement between the A-B pair and C-D pair, for all sets of phrases. Grey bars indicate displacement of the fourth syllable in the A-B pair, and black bars indicate displacement of the third syllable in the C-D pair (these syllables correspond to the thick lines in Figure 19.3). The displacement is expressed as a percentage of the interval between phrase-initial and final beats. Beat displacement in the C-D pair is smaller in magnitude than that in the A-B pair for all cases except set IV in Japanese. Overall, it appears that there are consistent differences in temporal displacement between foot-initial and foot-internal syllables. Foot-initial syllables are more resistant to temporal perturbations than are foot-internal syllables. This effect was found in both English and Japanese, contrary to previous claims that foot-level organisation is much less phonetically salient in Japanese than in English (Beckman 1994).

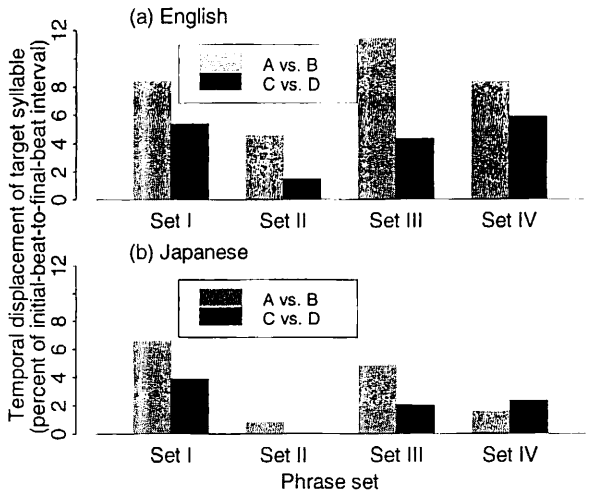

Figure 19.4 Temporal displacement of target syllables in pairs of minimally contrasting phrases. Grey bars indicate timing difference of the fourth syllable beat in the A-B pair, and black bars indicate timing difference of the third beat in the C-D pair (these syllables correspond to the thick vertical lines in the previous figure). Each bar graph shows the mean displacement across all speakers

A repeated-measures ANOVA was conducted with Syllable (foot-internal versus foot-initial) as a within-subjects factor and Language as a between-subjects factor. Results showed a highly significant main effect of Syllable [$F(1,25) = 32.78$; $p < .001$], indicating that the overall difference in temporal displacement between foot-initial and foot-internal beats is statistically reliable. Also significant were the main effect of Language [$F(1,25) = 24.12$; $p < .001$], and its interaction with Syllable [$F(1,25) = 7.15$; $p = .013$], showing that displacement of the beats was significantly larger overall in English than in Japanese.

19.2.3 Discussion

These data demonstrate similarities in rhythmic organisation between English and Japanese. The degree of temporal displacement of a syllable beat is sensitive to the prosodic foot structure of the phrases. In particular, the beats of foot-initial syllables are more resistant to temporal perturbations – i.e. more temporally stable – than are non-initial syllables.

To express the temporal displacement in absolute terms, a simple transformation was used to convert the y-axis of Figure 19.4 into milliseconds, noting that the interval between phrase-initial and final beats is approximately two-thirds of the metronome period. On average, syllable beats in English were

displaced by 66 ms, measured from the beginning of the phrase, if the target syllable was foot-internal, but only by 35 ms if the target was foot-initial. Similarly, displacement in Japanese was 24 ms for foot-internal syllables, and 14 ms for foot-initial syllables. Thus, in both languages, foot-initial beats moved roughly half as much as did foot-internal beats under similar segmental perturbations to the syllable's duration.

19.3 Experiment 2

Experiment 2 is similar to Experiment 1, but timing was perturbed through insertion of an extra syllable to a foot- or word-level unit. Text materials were made prosodically similar in English and Japanese by operationally equating English stressed syllables with Japanese accented moras. There is some limited evidence for this decision. Another speech-cycling study (Tajima 1998) compared Japanese phrases which were of the same length, but had different patterns of pitch accents and word boundaries. There was a slight tendency for word-initial accented syllables to occur at simple harmonic fractions of the phase cycle, such as 1/3 and 1/2, yielding distinct rhythmic patterns for these phrases. Thus, it is possible that both English stress and Japanese accent are attracted toward harmonic phases. However, conventional labels such as stress-timed versus mora-timed suggest that the insertion of a syllable in a phrase affects the timing of Japanese more severely than English.

19.3.1 Methods

Table 19.2 lists sample test phrases. There were 12 phrases in each language, consisting of four sets of three phrases each that followed a specified prosodic pattern, as illustrated below. The expected foot structure and downbeats of the waltz rhythm are also shown, much like the middle panel of Figure 19.1.

Pattern E:	[σ́ σ]	[σ́ σ σ]	[σ́]
Pattern F:	[σ́ σ]	[σ́ σ]	[σ́]
Pattern G:	[σ́ σ σ]	[σ́ σ]	[σ́]
Feet:	| foot 1	| foot 2	| foot 3 |
Downbeats:	1	2	3

Table 19.2 *Sample test phrases in Experiment 2*

Set	Pattern	English	Japanese
V	E	Búy the 'Dáily' for Gúy	dóno bíruma-da
	F	Búy the 'Dáy' for Gúy	dóno bíru-da
	G	Búy the 'Todáy' for Gúy	dókono bíru-da
VI	E	Gó to Báker for nów	mí-ta bíjo-mo-da
	F	Gó to Báker nów	mí-ta bíjo-da
	G	Gó to the báker nów	mí-se-ta bíjo-da

Looking at pattern F, this is a five-syllable pattern similar to the phrases in Experiment 1. Patterns E and G are six-syllable phrases that minimally contrasted with pattern F by the insertion of a syllable either to the second word-sized unit or to the first, respectively.

In English, the prominent syllables (σ́) were lexically stressed syllables, and the brackets corresponded to stress-feet. The prominent syllables were often also word-initial syllables. In Japanese, the prominent syllables were lexically pitch-accented syllables except for the phrase-final syllable which was always the unaccented copula *da*. The brackets roughly corresponded to words. Some of the words in patterns E and G were trimoraic, so the brackets shown in (1) did not correspond to bimoraic units parsed left-to-right. Thus, unlike Experiment 1, foot structure and word structure did not always coincide in the phrases used in Experiment 2.

Prosodic structure of the phrases were made reasonably comparable between English and Japanese by equating English stress-feet with short Japanese words with initial accent. In some conditions, these two units have the same intonation contour, with a high tone on the initial syllable, and relatively lower pitch elsewhere.

Despite the prosodic similarity, different predictions are made for English and Japanese. If English is stress-timed, then the stressed syllables should show relatively small temporal displacement across patterns E-G, despite the addition of an extra weak syllable in E and G. By contrast, if Japanese is mora-timed, then the addition of a mora to a word should yield a proportionate increase in word duration. This should make it difficult to keep word onsets near the downbeats of the repetition cycle.

Subjects in Experiment 2 were the same as those in Experiment 1. Two of the Americans did not participate in Experiment 2 because they had difficulty performing the task and took too much time to complete Experiment 1. Data

were therefore obtained from 12 Americans and 13 Japanese. Subjects were told that some phrases were longer than others, and were instructed to maintain the waltz rhythm while repeating the phrases in a comfortable fashion. All other aspects of the experiment were the same as Experiment 1. A total of 300 trials were analysed [12 phrases X 25 speakers].

19.3.2 Results

19.3.2.1 Production of waltz rhythm

Figure 19.5 shows the same type of histograms as Figure 19.2, showing how well the subjects as a group maintained the waltz rhythm. Again, data are collapsed across patterns E-G. The unfilled and filled bar graphs are from the phrase-medial and phrase-final prominent syllables, respectively (i.e. English stressed syllables or Japanese word-initial accented moras).

The histograms in Figure 19.5 show greater variability than those in Figure 19.2. This is especially noticeable in Japanese, where the distributions for both the third and fifth syllable beats are decidedly non-Gaussian. Timing of the prominent syllables is therefore less consistent across phrases and speakers in Experiment 2 than in Experiment 1, particularly in Japanese.

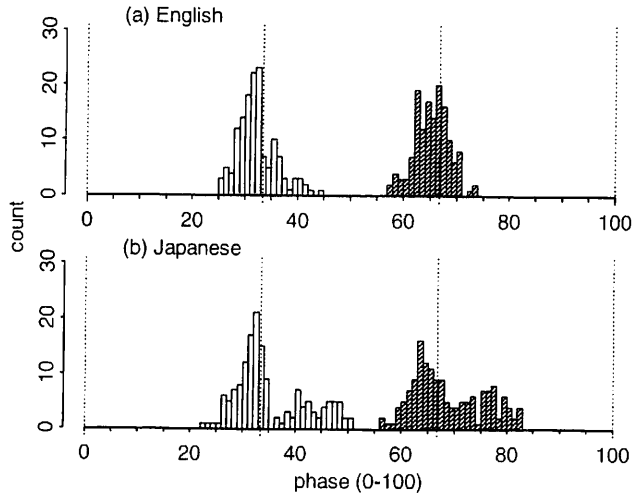

Figure 19.5 Histogram of phase of the phrase-medial (unfilled bars) and phrase-final (filled bars) prominent syllables from all trials in English (top) and Japanese (bottom). Each cluster is based on trial means from all phrases and speakers in each language

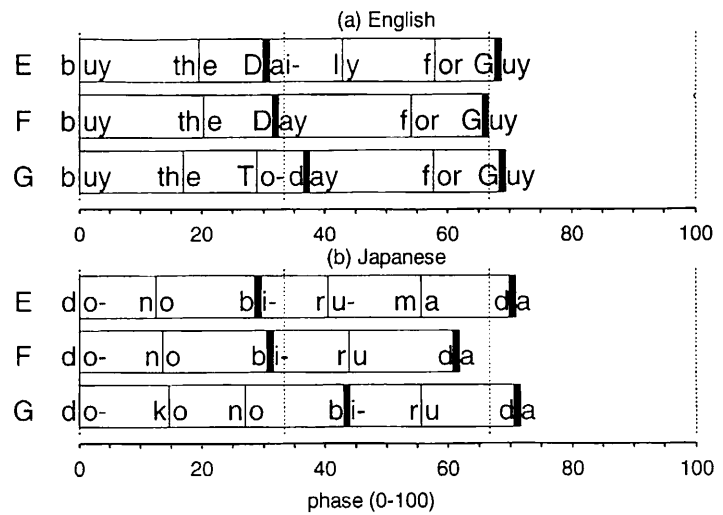

Figure 19.6 Relative timing of syllable beats in one matched set of phrases in English and Japanese. Each beat is the mean of trial means from all speakers in the language. The x-axis is phase relative to the phrase repetition cycle

19.3.2.2 Cross-linguistic differences in temporal stability

To examine how the insertion of an extra syllable affects the timing of phrases, Figure 19.6 shows the phase of syllable beats in one matched set of phrases in English and Japanese, averaged across speakers. The thick lines correspond to the prominent syllables in both languages.

Figure 19.6 shows that the prominent syllables undergo smaller temporal displacement across the phrases in English than in Japanese. English stressed syllables are produced reasonably close to the downbeats of the waltz metre. By contrast, Japanese shows a larger displacement of the word-initial accented moras. They exhibit much less tendency to occur close to the metrically strong positions. This property of Japanese is consistent with the traditional observation that moras tend to be isochronous.

Repeated-measures ANOVAs were carried out for the entire data set, with Pattern (E versus F versus G) and Language as factors. Separate tests were conducted for the phrase-medial and phrase-final prominent syllables. All main effects were significant at the .05 level, for both medial and final syllables. Importantly, Pattern-by-Language interaction was highly significant for the medial syllable [$F(2,46) = 21.69$; $p < .001$] and the final syllable [$F(2,46) = 46.91$; $p < .001$]. This indicates that the magnitude of displacement of the medial and final syllables across the three patterns was significantly different between English and Japanese.

In addition, Japanese showed more temporal variation across individual speakers than English did. It appears that all English speakers adopted a single

strategy for repeating the phrases, in which the stressed syllables were produced near the downbeats of the waltz. The Japanese speakers, however, showed greater inter-speaker variation. Figure 19.7 shows the mean phase of syllables from two sample speakers who apparently produced qualitatively distinct rhythmic patterns. For example, speaker 2 aligned the fifth syllable of the phrases near the third downbeat, at phase 66.6. For speaker 1, however, the sixth syllable of patterns E and G is closer to that downbeat than is the fifth syllable, perhaps suggesting that the speaker deviated from a three-beat rhythm. While the speakers differed from each other in the rhythmic patterns produced, each speaker was reasonably consistent across trials.

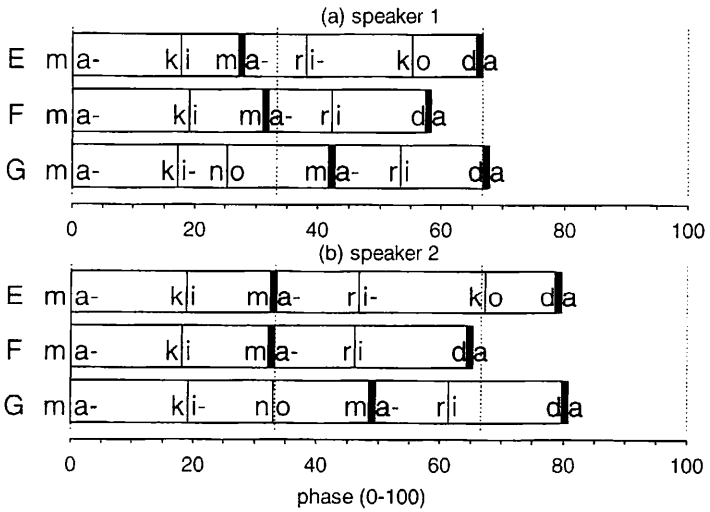

Figure 19.7 Illustration of inter-speaker variability in Japanese. Phase of syllable beats is shown for two sample speakers for one matched set of phrases

19.3.3 Discussion

Experiment 2 revealed measurable rhythmic differences between English and Japanese. English stressed syllables showed greater temporal stability across changes in foot size than Japanese word-initial accented syllables did across changes in word size. Also, compared to Experiment 1, phrases in Experiment 2 were not produced as consistently in the prescribed rhythm. This was especially true for Japanese, where individual speakers found alternative ways of aligning the text with the waltz rhythm.

Thus, the phonetic interpretation of metrical structure is highly sensitive to linguistic properties of the text. The instruction to speak in a waltz rhythm may

be followed more or less accurately, and in distinct ways, depending on what language one is a native speaker of.

19.4 General discussion and conclusions

19.4.1 Toward a theory of metre in speech

As an account of what speakers do in our speech-cycling task, we propose that two distinct cognitive structures are involved. The first is a *metre system* defining a pattern of attractors at downbeats in real continuous time. It is hypothesised that this system of attractors on the phase circle draws certain events (e.g. vowel onsets in prominent syllables) toward them. The attractors may result from a system of oscillators coupled at simple ratios such as 1:2 or 1:3 (Gasser, Eck and Port 1999; Large and Jones 1999). An oscillator is a cognitive process that has an instantaneous phase cycling around from 0 to 1 (or 360 degrees). Phase zero of each oscillator seems to serve as a downbeat, or an attractor of prominent events. In a 1:3 waltz rhythm, for example, the slower cognitive oscillator entrains to the stimulus metronome period. The second, faster oscillator is phase-coupled with the slower oscillator and makes three complete cycles for each cycle of the slow one, thus generating three isochronously spaced attractors for beats in each cycle. Thus, in asking our subjects to produce phrases 'using a waltz-like pattern', it seems that we effectively invited them to activate a 1:3 metre system. A similar metre system presumably underlies the perception and production of musical waltzes as well.

The second cognitive structure is a *linguistic text* containing a sequence of syllables having varying degrees of prominence. The relative prominence is determined by language-specific factors, including: (1) lexical stress (in English, at least), (2) lexical pitch accent, (3) word-initial syllables, (4) initial syllables in bimoraic feet (for Japanese, at least), and possibly other factors.

On this view, the speaker's problem when doing speech cycling is to align the intrinsic prominence pattern – the strong and weak syllables in the text – to the periodic structure of the metre system in real time. In this process, speakers of both languages exhibit the constraint that inherently prominent syllables are attracted toward phase zeros, i.e. the downbeats of the metre system. This tendency for attraction of prominent syllables to simple harmonic phases is a possible candidate to be a universal of human speech. The universality of this attraction seems reasonable since probably all languages are occasionally spoken in rhythmical ways, whether in verse, song or chant. On the other hand, what differs between English and Japanese seems to lie in the criteria for syllable prominence.

19.4.2 Metre system versus metrical grids

The above conception of how metre relates to speech is similar to Liberman's (1975) proposal of two distinct formal structures, a *metrical pattern*, to represent the abstract prominence relations among linguistic elements, and a *metrical grid*, to represent just a pattern of nested downbeats each with some discrete degree of strength. Though our proposal clearly resembles Liberman's in many respects, the traditional notion of a metrical grid does not provide a useful framework for an account of our phenomena.

The difficulty is that metrical grids are symbol strings, and thus can only represent the division of time into identical static intervals. The attraction itself – i.e. the adjustment of vowel onsets toward harmonic phases – would need to be accounted for outside the theory. On our view, on the other hand, metre is manifested as a real-time oscillatory process that generates attractors at specific temporal locations, on the fly. This attractor structure is directly reflected in the greater temporal stability observed with syllable beats located at the downbeats of the metre.

19.4.3 Rhythmic typology revisited

This study has demonstrated the possibility of treating relative temporal stability of one syllable beat versus another as a dependent variable for exploring the prominence structures of various languages. Temporal stability therefore provides an alternative approach for describing rhythmic typology, which has for the most part relied on isochrony as the primary diagnostic. Rather than claiming that languages vary in the level at which isochrony is maintained, we claim that languages vary in what counts as prominent. On this view, we believe that it is somewhat premature to abandon the 'stress-timed'–'syllable-timed' distinction on the basis of lack of isochrony alone. Instead, our comparison of English and Japanese suggests that there is something correct about the traditional rhythmic labels, and that the problem rather lay in using isochrony as the diagnostic for rhythm.

In English, prominence seems to correlate quite straightforwardly with stress. Stressed syllables play a central role in the rhythmic organisation of phrases, and strongly tend toward metrically strong time points. In Japanese, by contrast, there is no single linguistic factor that is primarily responsible for defining prominence. Prominence seems to be determined by several factors, which sometimes compete against each other when metre is imposed. One candidate is the initial syllables of bimoraic feet, while another is word-initial syllables with accents. In Experiment 1, foot-initial moras were also word-initial moras in each phrase. This may have led to stable and consistent performance by speakers. In

Experiment 2, however, foot-initial syllables were not always word-initial syllables. This may have led to greater overall variability in the data.

As for constructing a rhythmic typology of languages, it seems premature to make such an attempt at this time. The results do suggest that English is strongly 'stress-timed'. However, whether the particular kind of temporal organisation we observe here for Japanese will prove to be a rhythmic 'type' that deserves a name seems unclear on the basis of current data.

What do the speech-cycling results tell us about English or Japanese under conditions where speakers have not been asked to speak in a waltz-like pattern? The results suggest that although both languages showed attraction of prominent syllables to harmonic phases, the factors that made syllables prominent were quite different between the languages. These differences presumably apply just as much to ordinary, unconstrained conversation as to our artificial speech task.

Furthermore, the particular kinds of temporal organisation found in each language are not some artifact of our task, but instead are closely related to other known properties of the language. For example, the results here provide further evidence for temporal correlates of bimoraic units in Japanese (cf. Tajima 1998). Bimoraic feet have been shown to play an important role in morphophonological processes (Poser 1990), and have been demonstrated to have an effect on coarticulation of segments (Kondo and Arai 1998). These units can therefore be supported on both phonetic and phonological grounds, contrary to previous scepticism about the physical salience of Japanese feet (cf. Beckman 1995). Furthermore, the phonetic interpretation of bimoraic feet – tendency toward isochrony and stability of foot-initial syllables – shows striking resemblance to how English stress-feet are phonetically manifested. Thus, it appears that speech cycling has implications for typology as well as universals of speech rhythm.

Notes

We are grateful to Ken de Jong, Mafuyu Kitahara, Richard Wright and Bushra Zawaydeh for feedback on earlier drafts of this paper. We also thank the reviewers, particularly Mary Beckman, for their helpful comments. Special thanks are due to Fred Cummins for discussion of related issues and for the beat extractor software.

1 The vowel inversion actually leads to a difference in stress pattern. The second syllable of 'Bug-eye' bears 'secondary stress' because it contains a full vowel. By contrast, the second syllable of 'Byga' is 'stressless' since it contains a reduced vowel. In the present context, however, we focus on just the difference between 'primary stress' and 'non-primary stress' syllables.

2 The morpheme structure of some phrases deviated from this, e.g. the phrase /taka biʃa da/ consisted of one morpheme [taka biʃa] followed by the copula [da].