CHAPTER 30

The problem of speech patterns in time

Robert F. Port

One of the ways in which spoken languages differ from each other is in the temporal patterning of speech (e.g. Joos, 1948; Abramson and Lisker, 1964; Klatt, 1976; Ramus, 2002). These temporal patterns can be examined on several time scales, from very local (subsegmental and segmental) to global. The more global patterns—those lasting from a quarter second to a second or so—are most noticeable when we listen to speech in any language but our own. We may also become aware of rhythmic patterns in our own language when we hear it spoken by someone whose native language is something else. Along with other errors in foreign-accented speech, foreigners often seem to mangle the timing of our own language. Presumably there is some kind of pattern in the timing of speech gestures that accounts for our perception of rhythmic differences between languages. How might it be characterized? Human speech presents an extremely complex signal structured by many factors including the physiology of the vocal tract and constraints from the human perceptual system as well as patterns characteristic of the particular language (Fant, 1960; 1973). Which regions or time points are most relevant to rhythm perception? And, given some salient events, what kinds of durational measure between these events are most important for defining the temporal pattern itself? Whatever the answers are, these patterns in time create a sense that, for example, English, Spanish, and Chinese differ in their "speech rhythm." But what can be said objectively to justify even the use of the term "rhythm" in this context? Understanding linguistic rhythm is likely to be important, for example, for understanding how typical human linguistic fluency is achieved. Despite many attempts, however, a consistent and comprehensive framework for understanding speech timing has proven a challenge to researchers.

It seems likely that the problem of what the temporal patterns in a language really are will turn out to be partly a conceptual problem. Direct measurement of absolute time in seconds (treated as rational numbers in the computer) provides the raw empirical measures of time intervals, of course, but this is not at issue. The problem arises when we need to describe a pattern that is distributed over time.

30.1 Two conceptual frames for temporal patterns in speech

The conceptual tools we bring to any problem tend to shape our thinking along certain paths, even though the nature of this shaping may be difficult for us to see. Clarifying the conceptual tools we rely on may be a helpful step toward developing a broader, more flexible framework for description of temporal patterns. There seem to be two important a priori descriptive frames for describing speech patterns distributed over time: "symbol strings," using discrete, letter-like symbol tokens (such as the phonetic alphabet and orthographic words) whose patterns in time (i.e. transcriptions using an alphabet or orthography) serve as models for speech events, and "cycles," i.e. uniform motions around a circle, for measuring periodic time intervals. The data of speech research can be displayed graphically and measured in milliseconds, but a conceptual framework is needed to interpret the displays and measurements as linguistic or cognitive patterns over time. It seems

that the two basic descriptive schemes above present a choice of conceptual tool.

30.1.1 Symbol strings

The first and most important framework uses letter symbols as a model for the continuous gestures and sounds of phonetics and phonology. Phonetic and phonological transcriptions employ discrete, serially ordered (and thus non-overlapping) consonants and vowels as the uniform units of human speech. Thus English speakers produce sound patterns that we identify with the orthographic word *laugh* and can transcribe as the letter sequence [læf]. A transcription having these properties is assumed to be used by speakers to say the word, and used by hearers to recognize that the word has been spoken. Clearly, phonetic and phonological segments are intuitive and vivid for us, since they have most of the properties of the orthographic letters that were learned in childhood by those receiving an alphabet-based education (Ziegler and Goswami, 2005). Phonetic segments are like letters in being discrete and serial ordered, but differ from letters primarily in that they are not graphical but are hypothesized to be psychological. Thus, it is widely assumed that the nervous system needs to represent speech to itself symbolically, i.e. using some efficient, speaker-invariant way. It is the serially ordered phonological segments that play this role, just as we use the technology of writing on paper to represent and store utterances in a language. Letter-like, speaker-independent, and rate-independent symbols constitute the primary conceptual framework that has been used by most working linguists and psychologists for at least the past century (Saussure, 1916; Bloomfield, 1933; Jones, 1950; Jakobson et al., 1952; Chomsky and Halle, 1968; Liberman et al., 1968; Ladefoged, 1965; 1972; IPA, 1999). The continua of temporally overlapping speech gestures and speech acoustics are taken to be describable by an ordered sequence of discrete segmental symbols which are static (within each symbol) and whose only temporal relationships are definable in terms of the serial order of symbols (basically the relationships of "before" vs. "after" and, in the case of phonetic features, "simultaneous" vs. "non-simultaneous"). This model may be intuitively natural for us and thus attractive, but it has long been known that ordered letter-like symbols cannot provide rich enough specification of linguistic time patterns to successfully account for speech perception (Joos, 1948; Liberman et al., 1968; Dorman et al., 1979; Lisker and Abramson, 1971).

Of course, many important temporal properties of speech can be distinguished fairly well using letter-size symbols. Phonetic segments can differentiate [tæn] vs. [ænt] vs. [næt] by reordering the symbols, and much more. They can also be used to differentiate some more explicitly temporal patterns; for example, letters are useful to describe temporal patterns found in Japanese where there are several contrasts between lexical classes with the same sequence of phonetic segmental states. The phonetic alphabet can differentiate them by simply using one vs. two segmental symbols. For example, Figure 30.1a shows pronunciations of the words

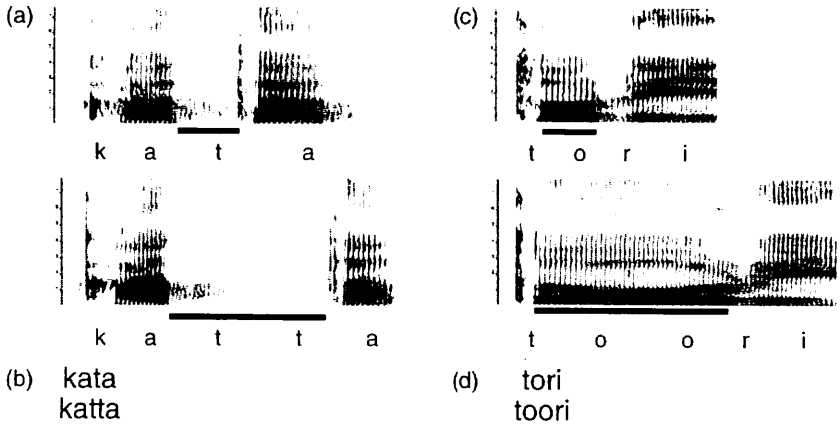

Figure 30.1 Spectrograms of Japanese minimal pairs, *kata* (shoulder) vs. *katta* (won) and *tori* (take) vs. *toori* (street) with the /t/ durations and /o/ durations marked with a black bar. The *x* axis is frequency (filtered into roughly 300 Hz bands) and *y* is time. It can be seen that both the stop and vowel segments are more than twice the duration when long as when they are short. Thanks to Kenji Yoshida for assistance with this figure.

kata and *katta* (excised from a neutral carrier sentence). The most obvious difference between the spectrograms of the two words is the duration of the [t] stop in the middle of the word (as marked by the horizontal black bar) (Dorman et al., 1979). This is normally modeled in linguistic descriptions (Vance, 1986; Tsujimura, 1995) by using one vs. two segments, as in Figure 30.1b. A similar difference between long and short vowels as in *tori* and *toori* is shown in Figure 30.1c (where the vowel is marked with black bars) and segmentally in Figure 30.1d. These segmental descriptions provide sufficient specificity for a speaker of Japanese to produce the intended words correctly. In fact, however, although the notation suggests a durational ratio between the two stops and two vowels of 2:1, the actual durational ratio tends to be closer to 3:1 in both cases (Hirata and Whiton, 2005; Hirata, 2004). But transcription using one instead of two symbols still does a reasonable job of representing these differences in duration. Later, an alternative description not depending on letters will be provided that will show why the durational ratios are about 1:3.

Many linguists and psychologists assume that everything of linguistic relevance about speech patterns in time can be captured using segment strings (e.g. Saussure, 1916; Bloomfield, 1933; Chomsky and Halle, 1968). Indeed, the observational data serving as input for most of linguistics is the phonetic transcription, a serially ordered list of symbol tokens. This seems to be the form in which speech is presented to our consciousness (Ladefoged, 1980; IPA, 1999; Bloomfield, 1933). Some linguists further assume that this list of tokens makes up a closed set and that all tokens of this set are perceptually distinct from each other (Chomsky and Halle, 1968; McCarthy, 2002), although others have disputed such a claim (Hockett, 1968; Sampson, 1977; Port and Leary, 2005). On the traditional view in linguistics, the natural and fundamental time scale of language is discrete, like letters and the integers, and not continuous in time.

A few linguists have criticized this view as reflecting a bias toward use of an alphabet (e.g. Firth, 1948; Linell, 2005), and have suggested that our segmental intuitions may result from lifelong training using letters to represent speech (Faber, 1992; Öhman, 2000; Port and Leary, 2005; Port, 2006). Phoneticians and linguists have struggled to reconcile the lack of fit between the physically observable continuous audio signals for speech and the intuitive symbol strings (Liberman et al., 1968; Fant, 1973; Ladefoged, 1984; Keating, 1984; Port, 1981). For example, the notion of phonetic segments has motivated experiments on unexpected timing cues such as those for the English voicing feature (Lisker and Abramson, 1964; Lisker, 1984) and motivated measurement by phoneticians of the durations of various segmental intervals (e.g. Klatt, 1976; Port, 1981; Ramus, Nespor and Mehler, 1999). But phoneticians generally understand that the phonetic segments of auditory phonetic transcription cannot be assigned to any specific acoustic features (see Pisoni and Levi, Chapter 1 this volume). This implies that there remains a mystery about how a consonant or vowel-like segments are related to the physical acoustic or motor forms of speech (Ladefoged, 1980; Browman and Goldstein, 1992).

An important reason for the appeal of segmental descriptions of speech is the rate invariance of this representation: a change in speaking rate need not result in a change in the phonetic or phonological transcription. This follows since strings of symbols have only serial order to encode time (although, of course, segmental symbols can be given labels like "long" and "short" or "20 ms in duration", but this is not the same as actually representing time; Lisker, 1984). Thus, distances between adjacent segments (like distances between adjacent letters) have no meaning. So the difficulty which arises is that serial order provides such a crude characterization of speech events in time that many aspects of speech timing that are critical for word specification, as well as for distinguishing between different languages, simply cannot be captured using segments alone (Port, 1986; Keating, 1985). This led to proposals for "temporal implementation rules" (Chomsky and Halle, 1968; Klatt, 1976; van Santen, 1996) which would employ durational labels like "inherent duration = 80 ms" and some arithmetic to compute "output duration targets" for each segment. But these target durations are still just symbols that need to be somehow interpreted.

The segmental model has inspired many studies of the characteristic timing patterns associated with the consonants and vowels of various languages. One product of decades of research on segmental aspects of speech is a number of generalizations about speech timing across languages (Lisker and Abramson, 1964; 1971; for reviews see Lehiste, 1970 and Klatt, 1976). For example, it seems to be generally the case that time intervals corresponding to low vowels like [a] are longer in the same context

than the same intervals for high vowels like [i] and [u] (Elert, 1964; Peterson and Lehiste, 1960). It also seems to be usually the case that the constrictions for voiceless stops and fricatives, like [t, p, s, f], are longer than the corresponding voiced consonant constrictions, like [d, b, z, v] (Elert, 1964; Lisker, 1984). Of course, there are also many temporal patterns which are unique characteristics of one language (or group of close relatives) but have not been found in other languages. Some examples include mora timing in Japanese (Port et al., 1987; Han, 1994), vowel lengthening after voiced obstruents in Arabic (Port et al., 1980), and complementary vowel and obstruent durations in Germanic languages (Elert, 1964; Lehiste, 1970; Lisker, 1984; Port and Crawford, 1989).

The important thing about most of these universal and language-specific timing patterns is that segmental models simply do not provide the conceptual tools to capture most of the temporal patterns. To use segments to describe speech, one must discard everything about time except the serial order of articulatory states. And once discarded, the information is no longer available for further analysis. But how could one retain more information about temporal patterns? One might try to measure out a time unit (e.g. a cycle) relative to which phonetic events can be located.

30.1.2 Circles in time

The second conceptual tool is the "circle in time," an isochronous cycle used as a scale for describing temporal patterns. If a temporal cycle is run at a slow rate, fractions of a time circle can be used for the specification of temporal patterns. By nesting a faster circular motion within a slower one, the slower one can be cut into halves or thirds (with frequencies twice or three times or more the frequency of the largest cycle). The European tradition of musical notation is a good illustration of this concept. Musical time is measured relative to a periodic pulse (frequently a foot tap) using nested fractions (half-notes, quarter-notes, etc.). "Meter" is the term for a pattern of nested cycles which pick out particular locations in time relative to the basic (or tactus) pulse (Handel, 1989). Repeated faster cycles (usually two or three) provide clock ticks or phase zeros within a slower cycle (Handel, 1989; Port, 2003). The standard music notation system generalizes the notion of meter to create a general method of periodic pattern notation.

Of course, because the intervals are measured relative to a regular pulse, they always represent temporal ratios, and the notated patterns are thus invariant under changes in the rate of the pulses. Any piece of music can be played with some variation in rate without changing the identity of the music. So, in music, temporal locations are in effect labeled as particular phase angles relative to nested periodic cycles. The most salient temporal targets for these locations tend to be harmonic fractions of the larger cycle. Of course, only in the notation will one find perfect nesting and cycles of ideal constant durations; actual performances will normally deviate from the formal ideal (Honing, 2002). Thus, musical notation divides the musical "measure" hierarchically into smaller integer fractions. It presumes one basic cycle, either the measure or the beat, to which the other cycles are time-locked. Most music in most cultural traditions can be approximated reasonably well using such a formal notational system for rhythm, since music traditions are usually based on some preferred metrical patterns (Seeger, 1958; Lehrdal and Jackendoff, 1983; Arom, 1991).

From mathematics, there is an alternative terminology for meter using phase angles where a single complete cycle can be described equivalently as 2π radians, 360°, or the interval {0, 1} (see Abraham and Shaw, 1983; Winfree, 2001). The cycle can then be divided into integer fractions just as in musical notation. Using a {0, 1} cycle, the onsets of a series of four quarter-notes in a four-beat musical measure would be located at phase angles of 0, 0.25, 0.50, and 0.75.

The cyclic framework for thinking about time has the important property of invariance under changes in speaking rate. Still this framework has only rarely been used to describe speech (a few cases where it has been used are Martin, 1972; Abercrombie, 1967: 97–8; Pike, 1946: 35) primarily because of the high temporal variability of speech. The problem is that the circle for measuring time obviously can have only a fixed rate of angular rotation of the cycle—just like the clock on the wall. So if some cycle-like behavior has a fixed rate, then the model applies nicely, but if the behavior exhibits moment-to-moment wobble in rate, then the framework of music notation seems inappropriate. This greatly limits its utility for describing speech.

So it seems that these two descriptive models, (1) serially ordered symbol strings and (2) phase angles of repeating cycles, provide scientists of language with our primary conceptual tools for understanding speech events in time. Both offer practical descriptive terms for certain phenomena

distributed over time. But the phenomena they describe are very different. Despite the overwhelming reliance on segmental description for the past century, this representation ignores too much temporal detail that is known to be important for speech production and perception, and offers no tools for description of rhythmically produced speech (Port and Leary, 2005). Period-based descriptions have been applied in a variety of ways over the years. The remainder of this chapter will survey a number of domains where the cycle framework has been applied, and will also describe some mechanisms that might explain the behavior of the nested cycles.

30.2 Global timing constraints: "stress-timed" vs. "syllable-timed" languages

Classical phonetic theory had little or nothing to say about possible temporal patterns in speech since it relied on letter-like segments to capture the serial order of gestures. Nevertheless, Pike (1946) boldly suggested that languages may come in two basic rhythmic styles, one based on periodic spacing of syllables and the other on periodic spacing of stressed syllables (so that unstressed syllables are constrained to fit). Abercrombie (1967) endorsed the suggestion and introduced the terms "stress-timed" (as English, German, Russian, and Arabic supposedly are) as opposed to "syllable-timed" (like French, Spanish, Telugu, and Yoruba). In the first type, some syllables are "stressed" (or emphasized) and others are not, while in the other type all syllables are equally weighted (Martin et al., 1972). This hypothesis makes predictions about the timing of various intervals that can be investigated experimentally.

Another point raised about this hypothesis is that the specific languages that were characterized as stress-timed and syllable-timed tend to differ in their constraints on the segmental structure of syllables (Dauer, 1983). Stress-timed languages, like English, tend to have some syllables that are complex, with initial and final consonant clusters and complex vowel nuclei (which tend to be the "stressed" syllables). Other syllables, the unstressed ones, tend to be much simpler, most often employing a single vowel. In contrast, typical syllable-timed languages, such as French and Spanish, tend to have much more uniform syllable types permitting few or no clusters and no complex vowel nuclei. Thus the timing regularities observed might be a consequence of mere segmental serial order patterns (Dauer, 1983; 1987).[1]

These observations have led to a great deal of research over the past few decades which has generally failed to support the predictions of isochrony (in stress-timed languages, Roach, 1982; Dauer, 1983, in syllable-timed languages, Wenk and Wioland, 1982). But, of course, perfect isochrony is only a prediction based on the notation system assuming constant-rate cycles. Another implicit assumption of the opposition of stress-timing vs. syllable-timing is that the notion of a syllable has some universal definition. In fact, no such cross-linguistic definition exists, and it is very difficult to define the notion of syllable in a universal way (Ramus, 2002).

As rigid methods of description were not supportive, attempts at statistical characterization of these speech timing types have been explored (Low and Grabe, 1995; Low et al., 2000; Ramus et al., 1999; Ramus et al., 2003). The more successful tests of the hypotheses about the typology of rhythm types have employed measures of the average variation in consonant duration (noted as ΔC in Figure 30.2) and the percentage of the total duration that was occupied by vocalic intervals (noted as %V). By this means, typical examples of stress-timed languages (English and Dutch) and syllable-timed languages (Spanish, French, Italian, and Catalan) could be separated, as shown in Figure 30.2 (from Ramus et al., 1999). The percentage of text duration that is vocalic was measured for rather small speech samples (five three-second speech samples by four speakers each) for eight languages. The results are suggestive, but have not been expanded. It is interesting that the only known example of a "mora-timed" language, Japanese, was separated nicely from both syllable-timed and stress-timed languages.

30.3 Regular "mora timing" in Japanese

Traditional teaching in the Japanese educational system has taught that each *kana* symbol (often

[1] Of course, if there were independent evidence of the two timing types, one could just as well argue that the syllable-structure constraints are a consequence of the characteristic timing type. If stress is alternated, then it would make sense for a language to invest in more complex syllables where the most attention is being paid. (Large and Jones, 1999).

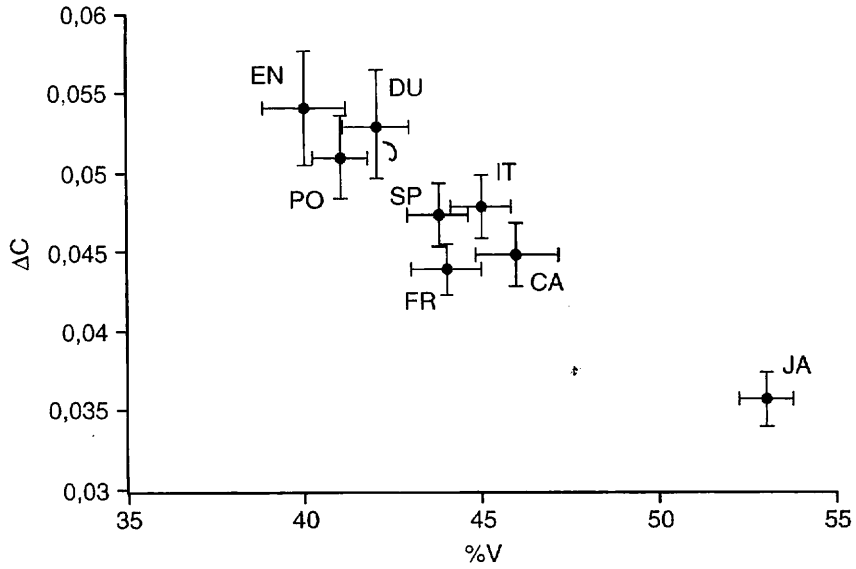

Figure 30.2 Means and standard deviations of consonant intervals in seconds plotted against the percent of overall duration that is vocalic for English (EN), Dutch (DU), Polish (PO), Spanish (SP), Italian (IT), French (FR), Catalan (CA), and Japanese (JA). Crossbars show ±1 SD across sample means. (Reproduced with permission from Ramus et. al. 1999.)

described as a syllabic writing system) takes the same amount of time to say (Homma, 1981; Beckman, 1982). This implies that word durations should come in integer-based durational ratios. Indeed, the duration of a word should be predictable from the number of moras (given a speaking rate). The speech fragment represented by a single kana symbol is called a "mora." Words can be written in this roughly syllabic writing system using an integer number of kana. But a mora is quite different from a syllable in English. Thus a consonant-vowel syllable, like *da* or *to*, is one mora, but a syllable like *hon* ("book, origin") requires two kana for its spelling and thus counts as two moras, whereas in English it is still a single syllable. Japanese has syllables with long vowels, as in *Tookyoo*, so this word counts as four moras (and requires four kana) while *Kyooto* counts as three moras (and three kana). There are also long consonants, as in *chotto* ("a little bit") (three moras, three kana) and *katta* ("won") (three moras and three kana).

Controversy arose as to whether the traditional claim is correct: are moras equal in duration? Of course, Japanese segments exhibit similar durational properties to other languages. That is, voiceless obstruents are longer than the corresponding voiced ones, low vowels are longer than high vowels, [s] is longer than the other fricatives, flaps are much shorter than other consonants, and so forth (Port et al., 1980; Beckman, 1982). This would lead one to expect that mora durations could not be constant. How could *ri* be the same duration as *sa* if the segments exhibit universal trends? In fact, they are not the same (Beckman, 1982; Port et al., 1987; Warner and Arai, 2001). Nevertheless, some researchers have continued to insist that the traditional claim is partly correct by showing that there are compensatory adjustments of neighboring moras when a mora is too long or too short (Port et al., 1987; Han, 1994). Supporters of "mora timing" point to evidence which looks at a series of neighboring moras. They note that, at least for certain styles of speech, word duration is tightly correlated with the number of moras, and that segments adjacent to a short mora (such as the long stop or vowel in Figure 30.1) are typically lengthened, which brings word duration closer to that expected for the number of moras.

Still the regularity of mora duration has remained controversial for several reasons. One is that as speaking rate is increased in Japanese, the regularity of mora timing gets much weaker (Warner and Arai, 2001). A second reason is that many phenomena about mora timing might be attributed to segmental constraints, although this account cannot explain the observed compensatory lengthening and shortening effects. Whatever the degree to which Japanese speakers may constrain their speech timing to regularize moras, it will be seen in the next section that humans

everywhere engage in *some* styles of speech that involve entrainment to periodic patterns.

30.4 Deliberate metrical production

There are a number of contexts in which speakers of most languages talk in a way that exhibits periodicity. The first is when reciting a list. When a person produces a familiar list of items (e.g. the numbers, the alphabet, the days of the week), the fluent pattern is to space the stressed syllables roughly equally in time (Quené and Port, 2005). Repetitive speech that is chanted by a single speaker (e.g. a commercial chant, a tour guide's text) or by a group of speakers (e.g. communal prayer, chanting at a political rally) also seem to be invariably periodic.

Poetry presents a somewhat confusing case. In literate communities, poetry seems to be more strongly associated with the written culture, and has metrical structures that need not be defined in continuous time at all but in terms of serially ordered patterns of various kinds (Boomsliter and Creel, 1977). But some poetry styles, such as so-called doggerel verse (i.e. humorous verse such as the limerick), are normally performed aloud using meter based on periodic temporal intervals, although little research has been done on such performances. Apparently all humans are able to constrain their speech to fit some externally specified periodicities. It seems that almost all human communities have traditions of periodic speaking styles (Merriam, 1964; List, 1963). These styles and their characteristic rhythm patterns seem to be easily picked up by children by 3 or 4 years of age, and are evidently entertaining to perform and listen to throughout life. Typically, there is periodic repetition of some pulse in the speech (sometimes reinforced by coupled periodic motions of the hands, feet or trunk), and prominent events in the speech are approximately aligned in time with these pulses (Port, 2003).

30.4.1 Perceptual centers

These periodic speech patterns raise an important question whose answer is not obvious: If many patterns are periodic and can be modeled with an oscillator, what region in a continuous speech event counts as phase zero? Where exactly is the beat? What serves as the origin of the cycle or the place where a person would locate, say, a tap (Allen, 1972a; 1972b)? To take a concrete example, if a speaker regularizes the spacing of a pair of alternated monosyllables, as in *ba, spa, ba, spa*, etc., will the [b]-release and [s]-onset be equally spaced (since they are the first sound emitted for each syllable), or will the onsets of the two [a]s, or something else? One can imagine a variety of possibilities. It might be the loudness maximum of the vowel, the location of a pitch maximum, the onset of a motor gesture, or even a gesture velocity peak. It is not known if the answer is uniform across languages, but with English-speaking subjects the answer seems to be approximately the onset of the vowel (Morton et al., 1976; Patel et al., 1999). The so-called "perceptual center" or "P-center" of a stressed syllable is close to the onset of the vowel [a] for both syllables. Thus, for *ba*, the beat occurs right at the onset of the vowel, and for *spa*, the beat moves slightly to the "left" of the vowel onset showing that the [s] in *spa* has some influence on the effective location of phase zero. This notion of a perceptual center can be approximated automatically by bandpass filtering the speech signal to include energy between roughly 100 Hz and 800 Hz (the F1 region), smoothing this energy envelope in time (integrated by about a 30 ms window) and then locating a peak in the first derivative (Scott, 1993). The reason that the first formant frequency region is most important is presumably because the first formant is where most of the acoustic energy is concentrated (Fant, 1960), and the onset is most important because the auditory nerve responds most strongly to onsets of energy (Delgutte, 1997; Kato et al., 1998).

30.5 Harmonic timing effect

As noted above, there are many situations where speakers talk as if in time to a metronome, placing stressed syllable onsets at roughly equal intervals. Rhythmic production of speech is certainly not the normal way to speak, but it is something every speaker performs for short periods from time to time. In addition, if English speakers are asked to repeat a short piece of text over and over, there is a strong tendency to divide the repetition cycle into equal-interval fractions, i.e. to nest several shorter cycles within the longer cycle. Thus the onsets of stressable syllables tend to migrate toward integer fractions of the whole cycle of repeated speech (Cummins and Port, 1998; Tajima and Port, 2003). For example, if a speaker repeats aloud a phrase like *Two thirty-five* at least five times, there is a strong tendency to locate the vowel onset of each instance of

five at one of only three locations relative to the vowel onsets of each successive repetition beginning with *Two*, as suggested by example (1). In the speaker's first or second reading of each pattern, irregular timings may result (where the beat for *five* may occur at many positions in the *Two-Two* cycle), but if the speaker continues to repeat the phrase, very quickly one of the three patterns, as in example (1a), (1b), or (1c), will become stable (Cummins and Port, 1998).

1. Three stable ways to repeat the phrase *Two thirty-five*, using boldface to suggest stress.

1a. ***Two thirty-five***, ***Two thirty-five***, ... (2 beats)

1b. ***Two thir-ty five***, ***Two thir-ty five***, ... (3 beats fast)

1c. ***Two thir-ty-five*** [rest], ***Two thir-ty-five*** [rest], ... (3 beats slow)

This has been termed the "harmonic timing effect" (Cummins and Port, 1998). To see the degree of the preference for these three patterns by English speakers, see Figure 30.3. In this experiment, participants were given a short piece of text to repeat, similar to *Two thirty-five* above. A two-toned metronomic auditory signal was presented, a high tone followed by a low tone. They were instructed to produce the first syllable (like *Two*) in time with the high tone and to produce the last syllable (like *five*) in time with the low tone. The phase lag between the high and low tones (relative to the high-high interval) was randomly varied from 0.20 to 0.70, producing a flat distribution of phase-lagged stimuli. Each metronome pattern was repeated for eight to ten cycles per trial (fewer at the slower rates) and subjects tried to hit the target pattern for each cycle. However, the frequency histogram of almost 8,000 measured phase lags in Figure 30.3 shows a strong bias toward one of the three rhythm patterns shown above in example (1). These results suggest the three formal rhythm patterns of the two- and three-beat measures shown in Figure 30.4. But the important question is: what mechanisms could cause speakers to show such strong biases toward these particular temporal patterns?

The data suggest that the formal, simple time locations illustrated in Figure 30.4 are really attractor basins in time. In fact, the attractor basins probably look much like Figure 30.3 turned upside down. These attractors encourage production of temporal patterns that are sufficiently optimal, in some sense, that speakers

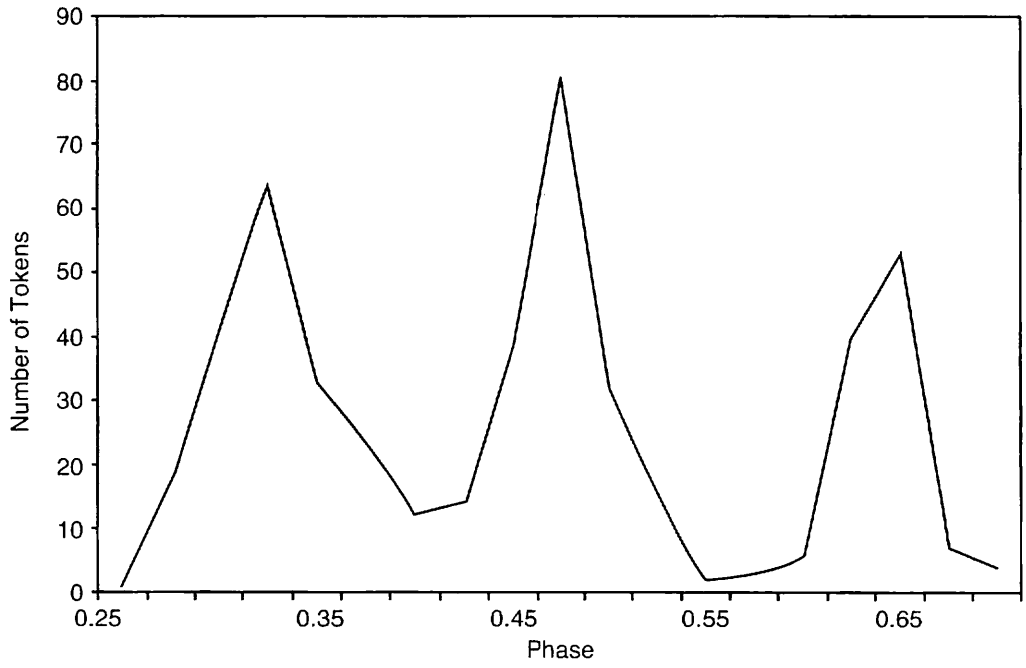

Figure 30.3 Participants were instructed to align the first syllable in a short phrase (*Two thirty-five* or *Beg for a dime*, etc.) with the first metronome pulse and the last syllable with the second pulse. The uniformly distributed target phases (as indicated by auditory pulses) could only be imitated by the participants with a bias showing three strong modes near 0.33, 0.5, and 0.66. Some lags they found very difficult to perform, such as the regions near 0.4 and 0.55. See Cummins and Port (1998) for more details. (Figure reproduced with permission from Port and Leary, 2005.)

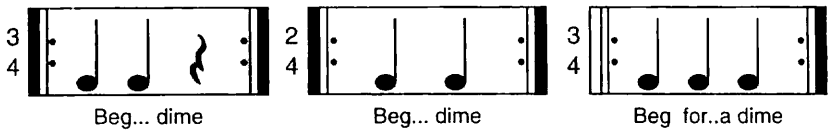

Figure 30.4 The three peaks in Figure 30.3 suggest the three musical notation patterns shown here (using the phrase *Beg for a dime* rather than *Two thirty-five*). When the final syllable occurs near one-third, there is a rest on the third of three beats. When it occurs near one-half, there are two beats to the cycle. And when it occurs near two-thirds, the syllable *for* finds itself on the second of three beats and begins to sound somewhat stressed. (Figure reproduced with permission from Cummins and Port, 1998.)

cannot help gravitating toward these particular phase relationships. This notion of attraction shows why a dynamical model, based on limit cycles, is an improvement over the rigid circle model discussed above. A dynamical model only claims that speakers will exhibit attractors at certain phase angles; and attractors can be demonstrated in a number of ways, such as the inability to accurately imitate a pattern containing a target away from an attractor but a fairly good ability to target the attractor itself.

But is this kind of timing at the phrase level really a linguistic phenomenon? One reason for such a conclusion is that speakers of different languages entrain speech to metronomes in different ways (cf. Zawaydeh et al., 2002; Tajima and Port, 2003). Some prefer for linguistics to deal only with phenomena that can be described with serially ordered symbols (Chomsky and Halle, 1968), implying that linguistics is defined by the familiar conceptual model of letters. To insist on this is to allow our conceptual tools to restrict our domain of study.

30.6 Modeling rhythm with dynamical systems

Further conceptual frameworks, beyond segments and circles in time, can be developed employing additional mathematics. Models using relaxation oscillators already exist which can produce attractors at the preferred phase angles. It is known how to construct simple dynamical models that behave periodically in ways that exhibit meter-like patterns by producing attractors at harmonic fractions. These mathematical models of rhythmic behavior can also be interpreted as models of the behavior of some structures of the central nervous system (e.g. Abraham and Shaw, 1983; Kelso, 1995; Saltzman and Munhall, 1989; Guenther, 1995). One kind of mechanism that has been proposed to account for the patterns in example (1) and Figure 30.3 would employ two coupled oscillator cycles, one of which tracks the cycle of the phrase as a whole and the other cycles either two or three times faster. Phase zero of each oscillator is an attractor for syllable onsets (i.e. for energy onsets), thereby providing target times for stressed syllable beats within the phrase cycle (Port, 2003). The faster oscillators are themselves attracted to the phase zero of the slow oscillator (thus keeping them at constant phase relative to each other), and also provide targets within the long cycle for stressed syllable onsets. The faster oscillators provide target phase angles at either one-half or at one-third and two-thirds of the longer cycle. Coupling between the oscillators means that the equation for the instantaneous phase of each includes the instantaneous phase of the other oscillator as a parameter (Abraham and Shaw, 1983; Large and Jones, 1999; Port, 2003). This would prevent their relative phase from shifting very much, and thus provide a periodic target time for location of each of the stressed syllable beats. Given equations for coupled oscillators, the phase zeros of both these oscillators become attractors that draw the auditory beats of the stressed syllables toward them (Large and Jones, 1999; Port, 2003). The only strong attractors (for English speakers, at least) were, as shown in example (1), at one-third, one-half, and two-thirds of the phase cycle (Cummins and Port, 1998), although other languages may find additional stable divisions (Zawaydeh et al., 2002; Tajima and Port, 2003). This implies that 2:1 and 3:1 oscillators are easy (or familiar) to English speakers, while other ratios (such as 5:1, 7:1, etc.) are much more difficult (and less familiar). It is possible that further model development along these lines will be productive.

30.7 Phrase edge timing phenomena

Another prominent durational effect is associated, most noticeably, with larger edges in

speech: phrases, paragraphs, and so on. It was noted early on that the vowels in the last syllable of a phrase (whether the phrase is a whole sentence or just a one-syllable word) are lengthened. Recent research suggests that this effect is much more general, and occurs in the middle of utterances such that the degree of slowing down is roughly proportional to the strength of the boundary (Byrd et al., 2000; Byrd and Saltzman, 2003). Thus the word boundary after *boys* in *Look at the boys on the field* will exhibit less slowing down than *If you see the boys, tell them to come home* and much less than *Hello boys, come on home.* This stretching or slowing appears to be achieved by decelerating all aspects of speech articulation at such boundaries. Thus consonants are also lengthened near (and especially before) a phrasal boundary (Byrd and Saltzman, 2003). The work by Byrd and Saltzman is formulated in terms of dynamical models of speech that are compatible with the dynamical models of speech rhythm discussed above.

30.8 Conclusion

The segment model for speech has dominated the thinking of phoneticians, psychologists, and phonologists for a century. But it has long been known to be inadequate to account for speech timing. These inadequacies have been ignored by many linguists because it was considered that intuitive descriptions of speech were of central relevance to linguistics. Phoneticians have turned in recent years to dynamical models of speech production and perception in order to address these inadequacies. Dynamical models for motor and perceptual cycles account for much, although we still do not know how to describe the audible rhythmic patterns of most speech in most languages. But at least some new conceptual tools are available that offer greater flexibility.

Acknowledgments

The author is grateful to Ken de Jong, Mark van Dam, and Kenji Yoshida for contributions to this essay.

References

Abercrombie, D. (1967) *Elements of General Phonetics.* Aldine, Chicago.

Abraham, R., and Shaw, C. (1983) *Dynamics: The Geometry of Behavior*, part 1. Aerial Press, Santa Cruz, CA.

Allen, G. (1972a) The location of rhythmic stress beats in English: an experimental study, I. *Language and Speech*, 15: 72–100.

Allen, G. (1972b) The location of rhythmic stress beats in English: an experimental study, II. *Language and Speech*, 15: 179–95.

Arom, S. (1991) *African Polyphony and Polyrhythm: Musical Structure and Methodology*, MIT Press, Cambridge, MA.

Beckman, M. (1982) Segment duration and the 'mora' in Japanese. *Phonetica*, 39: 113–135.

Bloomfield, L. (1933) *Language*, Holt, Rinehart & Winston, New York.

Boomsliter, P., and Creel, W. (1977) The secret springs: Housman's outline on metrical rhythm and language. *Language and Style*, 10: 296–323.

Browman, C., and Goldstein, L. (1992) Articulatory phonology: an overview. *Phonetica*, 49: 155–80.

Byrd, D., Kaun, A., Narayanan, S., and Saltzman, E. (2000) Phrasal signatures in articulation. In M. Broe and J. Pierrehumbert (eds), *Papers in Laboratory Phonology V.* Cambridge, Cambridge University Press, 70–87.

Byrd, D., and Saltzman, E. (2003) The elastic phrase: modeling the dynamics of boundary-adjacent lengthening. *Journal of Phonetics*, 31: 149–80.

Chomsky, N., and Halle, M. (1968) *The Sound Pattern of English.* Harper & Row, New York.

Cummins, F., and Port, R. (1998) Rhythmic constraints on stress timing in English. *Journal of Phonetics*, 26: 145–71.

Dauer, R. (1983) Stress-timing and syllable-timing reanalyzed. *Journal of Phonetics*, 11: 51–62.

Dauer, R. (1987) Phonetic and phonological components of language rhythm. Paper presented at the 11th International Congress of Phonetic Sciences, Tallinn, Estonia.

Delgutte, B. (1997) Auditory neural processing of speech. In W. J. Hardcastle and J. Laver (eds), *The Handbook of Phonetic Sciences.* Oxford, Blackwell, 507–38.

Dorman, M., Raphael, L., and Liberman, A. (1979) Some experiments on the sound of silence in phonetic perception. *Journal of the Acoustical Society*, 65: 1518–32.

Elert, C.-C. (1964) *Phonological Studies of Quantity in Swedish.* Stockholm, Almqvist & Wiksell.

Faber, A. (1992) Phonemic segmentation as epiphenomenon: evidence from the history of alphabetic writing. In P. Downing, S. Lima, and M. Noonan (eds), *The Linguistics of Literacy*, pp. 111–34. Amsterdam, Benjamins.

Fant, G. (1960) *The Acoustical Theory of Speech Production.* The Hague, Mouton.

Fant, G. (1973) *Speech Sounds and Features.* MIT Press, Cambridge, MA.

Firth, J. R. (1948) Sounds and prosodies. *Transactions of the Philological Society*, 127–52.

Guenther, F. (1995) Speech sound acquisition, coarticulation and rate effects in a neural network model of speech production. *Psychological Review*, 102: 594–621.

Han, M. (1994) Acoustic manifestations of mora timing in Japanese. *Journal of the Acoustical Society of America*, 96: 73–82.

Handel, S. (1989) *Listening: An Introduction to the Perception of Auditory Events*, MIT Press, Cambridge, MA.

Hirata, Y. (2004) Effects of speaking rate on the vowel length distinction in Japanese. *Journal of Phonetics*, 32: 565–89.

Hirata, Y. and Whiton, J. (2005) Effects of speaking rate on the single/geminate stop distinction in Japanese. *Journal of the Acoustical Society of America*, 118: 1647–60.

Hockett, C. (1968) *The State of the Art*. The Hague, Mouton.

Homma, Y. (1981) Durational relationship between Japanese stops and vowels. *Journal of Phonetics*, 9: 273–81.

Honing, H. (2002) Structure and interpretation of rhythm and timing. *Tijdschrift voor Muziektheorie*, 7: 227–32.

IPA (1999) *Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet*. Cambridge University Press, Cambridge, MA.

Jakobson, R., Fant, G., and Halle, M. (1952) *Preliminaries to Speech Analysis: The Distinctive Features*, MIT Press, Cambridge, MA.

Jones, D. (1950) *The Phoneme: Its Nature and Use*. Cambridge University Press, Cambridge.

Joos, M. (1948) Acoustic phonetics. *Language Monograph*, Linguistic Society of America, 23.

Kato, H., Tsuzaki, M., and Sagisaka, Y. (1998) Acceptibility for temporal modification of single vowel segments in isolated words. *Journal of the Acoustical Society of America*, 104: 540–9.

Keating, P. (1984) Phonetic and phonological representation of stop consonant voicing. *Language*, 60: 286–319.

Keating, P. (1985) Universal phonetics and the organization of grammars. In V. Fromkin (ed.), *Phonetic Linguistics: Essays in Honor of Peter Ladefoged*, pp. 115–32. Academic Press, New York.

Kelso, J. A. S. (1995) *Dynamic Patterns: The Self-Organization of Brain and Behavior*. MIT Press, Cambridge, MA.

Klatt, D. H. (1976) Linguistic uses of segmental duration in English: acoustic and perceptual evidence. *Journal of the the Acoustical Society of America*, 59: 1208–21.

Ladefoged, P. (1965) The nature of general phonetic theories. *Georgetown University Monograph 18, Language and Linguistics*, 27–42.

Ladefoged, P. (1972) *A Course in Phonetics*. Orlando, FL, Harcourt Brace Jovanovich.

Ladefoged, P. (1980) What are linguistic sounds made of? *Language*, 56: 485–502.

Ladefoged, P. (1984) 'Out of chaos comes order': physiological, biological and structural patterns in phonetics. In M. P. R. V. D. Broeke and A. Cohen (eds), *Proceedings of the Tenth International Congress of Phonetic Sciences*, 83–95. Dordrecht, Foris.

Large, E. W., and Jones, M. R. (1999) The dynamics of attending: how we track time-varying events. *Psychological Review*, 106: 119–159.

Lehiste, I. (1970) *Suprasegmentals*. MIT Press, Cambridge, MA.

Lehrdahl, F., and Jackendoff, R. (1983) *A Generative Theory of Tonal Music*. MIT Press, Cambridge, MA.

Liberman, A. M., Delattre, P., Gerstman, L., and Cooper, F. (1968) Perception of the speech code. *Psychological Review*, 74: 431–61.

Linell, P. (2005) *The Written Language Bias in Linguistics*. Oxford, Routledge.

Lisker, L. (1984) 'Voicing' in English: a catalogue of acoustic features signalling /b/ vs. /p/ in trochees. *Language and Speech*, 29: 3–11.

Lisker, L., and Abramson, A. (1964) A cross-language study of voicing in initial stops: acoustical measurements. *Word*, 20: 384–422.

Lisker, L., and Abramson, A. (1971) Distinctive features and laryngeal control. *Language*, 47: 767–85.

List, G. (1963) The boundaries of speech and song. *Ethnomusicology*, 7: 1–16.

Low, E., and Grabe, E. (1995) Prosodic patterns in Singapore English. In *Proceedings of the XIIIth* International Congress of Phonetic Sciences, Stockholm, Vol 3: 636–9.

Low, E., Grabe, E., and Nolan, F. (2000) Quantitative characterizations of speech rhythm: syllable-timing in Singapore English *Language and Speech*, 43: 377–401.

Martin, J. (1972) Rhythmic (hierarchical) versus serial structure in speech and other behavior. *Psychological Review*, 79: 487–509.

McCarthy, J. (2002) *A Thematic Guide to Optimality Theory*, Cambridge University Press, Cambridge.

Merriam, A. (1964) *The Anthropology of Music*, Northwestern University Press, Chicago.

Morton, J., Marcus, S., and Frankish, C. (1976) Perceptual centers (P-centers). *Psychological Review*, 83: 405–8.

Öhman, S. E. G. (2000) Expression and content in linguistic theory. In M. Gustafsson and L. Hertzberg (eds), *The Practice of Language*. Dordrecht, Kluwer Academic.

Patel, A., Lofquist, A., and Naito, W. (1999) The acoustics and kinematics of regularly timed speech: a database and method for the study of the P-center problem. Paper presented at the 14th International Congress of Phonetic Sciences, San Francisco, Vol 1: 405–8.

Peterson, G. E., and Lehiste, I. (1960) Duration of syllable nuclei in English. *Journal of the Acoustical Society of America*, 32: 693–703.

Pike, K. (1946) *The Intonation of American English*. University of Michigan Press, Ann Arbor, MI.

Port, R. (1981) Linguistic timing factors in combination. *Journal of the Acoustical Society of America*, 69: 262–74.

Port, R. (1986) Invariants in phonetics. In D. Klatt and J. Perkell (eds), *Invariance and Variability in the Speech Processes*, pp. 540–58. Erlbaum, Hillsdale, NJ.

Port, R. (2003) Meter and speech. *Journal of Phonetics*, 31: 599–611.

Port, R. (2006) The graphical basis of phones and phonemes. In Ocke-Schwen Bohn and Murray J. Munro M. (eds), *Language Experience in Second Language Speech Learning: In honor of James Emil Flege*.Amsterdam, Benjamins 349–65.

Port, R. F., Al-Ani, S., and Maeda, S. (1980) Temporal compensation and universal phonetics. *Phonetica*, 37: 235–52.

Port, R. and Crawford, P. (1989) Pragmatic effects on neutralization rules. *Journal of Phonetics*, 16: 257–82.

Port, R., Dalby, J., and O'dell, M. (1987) Evidence for mora timing in Japanese. *Journal of Acoustical Society*, 81: 1574–85.

Port, R. F., and Leary, A. (2005) Against formal phonology. *Language*, 81: 927–64.

Quené, H., and Port, R. (2005) Effects of timing regularity and metrical expectancy on spoken word perception. *Phonetica*, 62: 1–13.

Ramus, F. (2002) Acoustic correlates of linguistic rhythm: perspectives. Paper presented at Speech Prosody 2002, Aix-en-Provence.

Ramus, F., Dupoux, E., and Mehler, J. (2003) The psychological reality of rhythm classes: perceptual studies. Paper presented at International Congress of Phonetic Sciences, "Speech Prosody". Barcelona.

Ramus, F., Nespor, M., and Mehler, J. (1999) Correlates of linguistic rhythm in the speech signal. *Cognition*, 73: 265–92.

Roach, P. (1982) On the distinction between 'stress-timed' and 'syllable-timed' languages. In D. Crystal (ed.), *Linguistic Controversies*, pp. 73–9. Arnold, London.

Saltzman, E., and Munhall, K. (1989) A dynamical approach to gestural patterning in speech production. *Ecological Psychology*, 1: 333–82.

Sampson, G. (1977) Is there a universal phonetic alphabet? *Language*, 50: 236–59.

Saussure, F. D. (1916) *A Course in General Linguistics*. Philosophical Library, New York.

Scott, S. K. (1993) *P-centers in Speech: An Acoustic Analysis*. Unpublished doctoral thesis, University College, London.

Seeger, C. (1958) Descriptive and prescriptive music writing. *The Musical Quaterly*, Vol 44: 184–95.

Tajima, K., and Port, R. (2003) Speech rhythm in English and Japanese. In J. Local, J., Ogden, R., and Temple, R. (eds), *Phonetic Interpretation: Papers in Laboratory Phonology*, pp. 317–34. Cambridge University Press, Cambridge.

Tsujimura, N. (1995) *An Introduction to Japanese Linguistics*. Blackwell, Oxford.

van Santen, J. P. H. (1996) Segmental duration and speech timing. In Y. Sagisaka, N. Campbell, and N. H-iguchi (eds), *Computing Prosody: Computational Models for Processing Spontaneous Speech*. Springer, New York.

Vance, T. (1986) *An Introduction to the Phonology of Japanese*. State University of New York Press, Albany, New York.

Warner, N., and Arai, T. (2001) Japanese mora timing: a review. *Phonetica*, 58: 53–87.

Wenk, B. J., and Wioland, F. (1982) Is French really syllable-timed? *Journal of Phonetics*, 10: 193–216.

Winfree, A. (2001) *The Geometry of Biological Time*. Springer, New York.

Zawaydeh, B., Tajima, K., and Kitahara, M. (2002) Discovering Arabic rhythm through a speech cycling task. In D. Parkinson and E. Benmamoun (eds), *Perspectives on Arabic Linguistics*, pp. 39–58. Amsterdam, Benjamins.

Ziegler, J., and Goswami, U. (2005) Reading acquisition, developmental dyslexia and skilled reading across languages: A psycholinguistic grain size theory. *Psychological Bulletin*, 131: 3–29.