3

Rhythm

3.1 Introduction

The comparative study of spoken and musical rhythm is surprisingly underdeveloped. Although hundreds of studies have explored rhythm within each domain, empirical comparisons of linguistic and musical rhythm are rare. This does not reflect a lack of interest, because researchers have long noted connections between theories of rhythm in the two domains (e.g., Selkirk, 1984; Handel, 1989). The paucity of comparative research probably reflects the fact that specialists in one domain seldom have the time to delve into the intricacies of the other. This is regrettable, because cross-domain work can provide a broader perspective on rhythm in human cognition. One goal of this chapter is to equip researchers with conceptual and empirical tools to explore the borderland between linguistic and musical rhythm. As we shall see, this is a fertile area for new discoveries.

Before embarking, it is worth addressing two overarching issues. The first is the definition of rhythm. The term "rhythm" occurs in many contexts besides speech and music, such as circadian rhythms, oscillations in the brain, and the rhythmic calls of certain animals. In most of these contexts, "rhythm" denotes periodicity, in other words, a pattern repeating regularly in time. Although periodicity is an important aspect of rhythm, it is crucial to distinguish between the two concepts. The crux of the matter is simply this: Although all periodic patterns are rhythmic, not all rhythmic patterns are periodic. That is, periodicity is but one type of rhythmic organization. This point is especially important for understanding speech rhythm, which has had a long (and as we shall see, largely unfruitful) association with the notion of periodicity. Thus any definition of rhythm should leave open the issue of periodicity. Unfortunately, there is no universally accepted definition of rhythm. Thus I will define rhythm as the systematic patterning of sound in terms of timing, accent, and grouping. Both speech and music are characterized by systematic temporal, accentual, and phrasal patterning. How do these patterns compare? What is their relationship in the mind?

The second issue is the very notion of rhythm in speech, which may be unfamiliar to some readers. One way to informally introduce this concept is to consider the process of learning a foreign language. Speaking a language with native fluency requires more than mastering its phonemes, vocabulary, and grammar. One must also master the patterns of timing and accentuation that characterize the flow of syllables in sentences. That is, each language has a rhythm that is part of its sonic structure, and an implicit knowledge of this rhythm is part of a speaker's competence in their language. A failure to acquire native rhythm is an important factor in creating a foreign accent in speech (Taylor, 1981; Faber, 1986; Chela-Flores, 1994).

The following two sections (3.2 and 3.3) give overviews of rhythm in music and speech, respectively, focusing on issues pertinent to cross-domain comparisons. (Such comparisons are made within each section where appropriate.) These overviews motivate a particular way of looking at rhythmic relations between speech and music. This new perspective is introduced in the final section of the chapter, together with empirical evidence spanning acoustic, perceptual, and neural studies.

3.2 Rhythm in Music

The following discussion of rhythm in music focuses on music that has a regularly timed beat, a perceptually isochronous pulse to which one can synchronize with periodic movements such as taps or footfalls. Furthermore, the focus is on music of the Western European tradition, in which beats are organized in hierarchies of beat strength, with alternation between stronger and weaker beats. This form of rhythmic organization has been the most widely studied from a theoretical and empirical standpoint, and is also the type of rhythm most often compared with speech, either implicitly or explicitly (Pike, 1945; Liberman, 1975; Selkirk, 1984).

It is important to realize, however, that this is just one way in which humans organize musical rhythm. It would be convenient if the rhythmic structure of Western music indicated general principles of rhythmic patterning. Reality is more complex, however, and only a comparison of different cultural traditions can help sift what is universal from what is particular. To illustrate this point, one can note musical traditions in which rhythm is organized in rather different ways than in most Western European music.

One such tradition involves the Ch'in, a seven string fretless zither that has been played in China for over 2,000 years (van Gulik, 1940). The musical notation for this instrument contains no time markings for individual notes, indicating only the string and type of gesture used to produce the note (though sometimes phrase boundaries are marked). The resulting music has no sense of a beat. Instead, it has a flowing quality in which the timing of notes emerges from the gestural dynamics of the hands rather than from an explicitly regulated temporal scheme. The Ch'in is just one of many examples of unpulsed music from around the globe, all of which show that the mind is capable of organizing temporal patterns without reference to a beat.

Another tradition whose rhythms are quite different from Western European music is Balkan folk music from Eastern Europe (Singer, 1974; London, 1995). This music has salient beats, but the beats are not spaced at regular temporal intervals. Instead, intervals between beats are either long or short, with the long interval being 3/2 the length of the shorter one. Rhythmic cycles are built from repeating patterns of long and short intervals, such as S-S-S-L, S-S-L-S-S (note that the long element is not constrained to occur at the end of the cycle). One might think that such an asymmetric structure would make the music difficult to follow or synchronize with. In fact, listeners who grew up with this music are adept at following these complex meters (Hannon & Trehub, 2005), and much of this music is actually dance music, in which footfalls are synchronized to the asymmetric beats.

As a final example of a rhythmic tradition with a different orientation from Western European music, Ghanian drumming in West Africa shows a number of interesting features. First, the basic rhythmic reference is a repeating, non-isochronous time pattern played on a set of hand bells (Locke 1982; Pantaleoni, 1985). Members of a drum ensemble keep their rhythmic orientation by hearing their parts in relation to the bell, rather than by focusing on an isochronous beat. Furthermore, the first beat of a rhythmic cycle is not heard as a "downbeat," in other words, a specially strong beat (as in Western music); if anything, the most salient beat comes at the *end* of the cycle (Temperley, 2000). Finally, this music emphasizes diversity in terms of the way it can be heard. As different drums enter, each with its own characteristic repeating temporal pattern, a polyrhythmic texture is created that provides a rich source of alternative perceptual possibilities depending on the rhythmic layers and relationships one chooses to attend to (Locke 1982; Pressing, 2002). This is quite different from the rhythmic framework of most Western European music, in which the emphasis is on relatively simple and perceptually consensual rhythmic structures. One possible reason for this difference is that Western music has major preoccupations in other musical dimensions (such as harmony), and a relatively simple rhythmic framework facilitates complex explorations in these other areas. Another reason may be that tempo in Western European music is often flexible, with salient decelerations and accelerations of the beat used for expressive purposes. A fairly simple beat structure may help a listener stay oriented in the face of these temporal fluctuations (cf. Temperley, 2004).

Thus it would be an error to assume that the rhythmic structure of Western European music reflects basic constraints on how the mind structures rhythmic patterns in terms of production or perception. As with every musical tradition, the rhythmic patterns of Western European music reflect the historical and musical concerns of a given culture. On the other hand, a comparative perspective reveals that certain aspects of rhythm in Western European music (such as a regular beat and grouping of events into phrases) are also found in numerous other cultures, which suggests that these aspects reflect widespread cognitive proclivities of the human mind.

The discussion below relies at times on one particular melody to illustrate various aspects of rhythmic structure in Western music. This is the melody of a children's song, indexed as melody K0016 in a database of Bohemian folk melodies (Schaffrath, 1995; Selfridge-Feld, 1995). Figure 3.1 shows the melody in Western music notation and in "piano roll" notation with each tone's pitch plotted as a function of time (the melody can be heard in Sound Example 3.1).

The melody was chosen because it is historically recent and follows familiar Western conventions, yet is unlikely to be familiar to most readers and is thus free of specific memory associations. It also illustrates basic aspects of rhythm in a simple form. Beyond this, there is nothing special about this melody, and any number of other melodies would have served the same purpose.

3.2.1 The Beat: A Stable Mental Periodicity

The phenomenon of a musical beat seems simple because it is so familiar. Almost everyone has tapped or danced along to music with a beat. A regular beat is widespread in musical cultures, and it is worth considering why this might be so. One obvious function of a beat is to coordinate synchronized movement, such as dance. (The relationship between dance and music is widespread

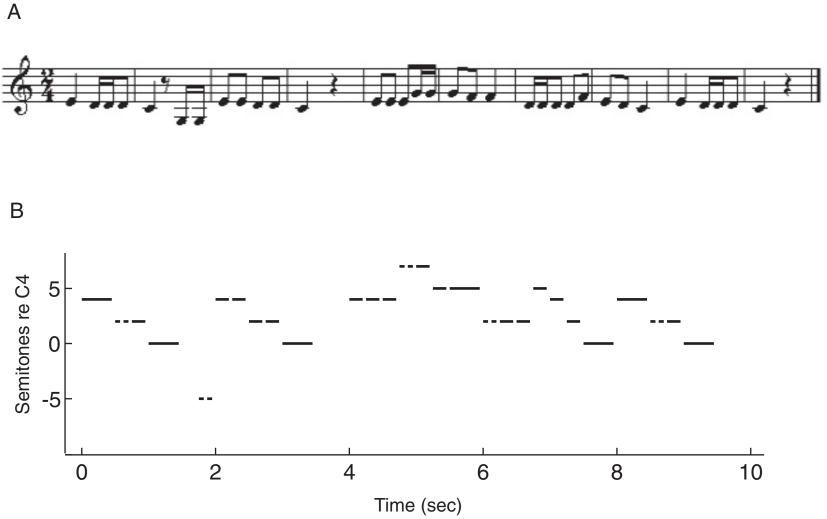

Figure 3.1 A simple melody (K0016) in (A) music notation and (B) piano roll format. In (B), the y-axis shows the semitone distance of each pitch from C4 (261.63 Hz).

in human societies; indeed, some cultures do not even have separate terms for music and dance.) A second obvious function of a beat is to provide a common temporal reference for ensemble performance. Indeed, a cross-cultural perspective reveals that ensemble music without a periodic temporal framework is a rare exception. Perlman (1997) points to one such exception in Javanese music known as *pathetan,* noting that "Except for certain isolated phrases, *pathetan* has no unifying metric framework. . . . Rhythmic unison is not desired, and the musicians need not match their attacks with the precision made possible by a definite meter" (p. 105). A detailed discussion of pathetan by Brinner (1995:245–267) suggests that it is an exception that proves the rule: without a metric frame, players substitute close attention to a lead melodic instrument (typically a *rebab* or bowed lute) in order to coordinate and orient their performance. Thus when periodicity in ensemble music is withdrawn, its functional role is filled in other ways.

From a listener's perspective, perception of a beat is often linked to movement in the form of synchronization to the beat. For many people, this synchronization is a natural part of musical experience requiring no special effort. It may come as a surprise, then, that humans are the only species to spontaneously synchronize to the beat of music. Although synchrony is known from other parts of the animal kingdom, such as the chorusing of frogs or the synchronized calls of insects (Gerhardt & Huber 2002, Ch. 8; Strogatz, 2003), human synchronization with a beat is singular in a number of respects (see Chapter 7, section 7.5.3, for further discussion of this point). Of course, beat perception does not automatically cause movement (one can always sit still), but the human uniqueness of beat synchronization suggests that beat perception merits psychological investigation. Research in music cognition has revealed several interesting facts about beat perception.

First, there is a preferred tempo range for beat perception. People have difficulty following a beat that is faster than every 200 ms and slower than every 1.2 seconds. Within this range, there is a preference for beats that occur roughly every 500–700 ms (Parncutt, 1994; van Noorden & Moelants, 1999). It is interesting to note that this is the same range in which people are the most accurate at making duration judgments, in other words, they neither overestimate nor underestimate the duration of temporal intervals (Eisler, 1976; cf. Fraisse, 1982). Furthermore, this is the range in which listeners are the most accurate in judging slight differences in tempo (Drake & Botte, 1993). It is also interesting to note that in languages with stressed and unstressed syllables, the average duration between stressed syllables has been reported to be close to or within this range (Dauer, 1983; Lea, 1974, described in Lehiste, 1977).

Second, although people usually gravitate toward one particular beat tempo, they can tap at other tempi that are simple divisors or multiples of their preferred tapping rate (e.g., at double or half their preferred rate; Drake, Jones, & Baruch, 2000). For example, consider Sound Example 3.2, which presents K0016 along with two different indications of the beat. Both are perfectly possible, and it is likely that most people could easily tap at either level depending on whether they focus on lower or higher level aspects of rhythmic structure. Drake, Jones, and Baruch (2000) have shown that people vary in the level they synchronize with in music, and that their preferred level correlates with their spontaneous tapping rate. Furthermore, although individuals naturally gravitate to one particular level, they can move to higher or lower levels if they wish (e.g., by doubling or halving their tapping rate) and still feel synchronized with the music. Thus when speaking of "the beat" of a piece, it is important to keep in mind that what a listener selects as the beat is just one level (their *tactus*) in a hierarchy of beats.

Third, beat perception is robust to moderate tempo fluctuations. In many forms of music, the overall timing of events slows down or speeds up within phrases or passages as part of expressive performance (Palmer, 1997). People are still able to perceive a beat in such music (Large & Palmer, 2002) and synchronize to it (Drake, Penel, & Bigand, 2000), indicating that beat perception is based on flexible timekeeping mechanisms.

Fourth, there is cultural variability in beat perception. Drake and Ben El Heni (2003) studied how French versus Tunisian listeners tapped to the beat of French versus Tunisian music. The French tapped at a slower rate to French music than to Tunisian music, whereas the Tunisians showed the opposite pattern. Drake and Ben Heni argue that this reflects the fact that listeners can extract larger-scale structural properties in music with which they are familiar. These findings indicate that beat perception is not simply a passive response of the auditory system to physical periodicity in sound: It also involves cultural influences that may relate to knowledge of musical structure (e.g., sensitivity to how notes are grouped into motives; cf. Toiviainen & Eerola, 2003).

Fifth, and of substantial interest from a cognitive science standpoint, a perceived beat can tolerate a good deal of counterevidence in the form of accented events at nonbeat locations and absent or weak events at beat locations, in other words, syncopation (Snyder & Krumhansl, 2001). For example, consider Sound Examples 3.3 and 3.4, two complex temporal patterns studied by Patel, Iversen, et al. (2005) with regard to beat perception and synchronization. The patterns begin with an isochronous sequence of 9 tones that serves to indicate the beat, which has a period is 800 ms. After this "induction sequence," the patterns change into a more complex rhythm but with the same beat period. Participants were asked to synchronize their taps to the isochronous tones and then continue tapping at the same tempo during the complex sequence. Their success at this task was taken as a measure of how well they were able to extract a beat from these sequences. In the "strongly metrical" (SM) sequences (Sound Example 3.3), a tone occurred at every beat position. In the "weakly metrical" (WM) sequences, however, about 1/3 of the beat positions were silent (Sound Example 3.4). (NB: The SM and WM sequences had exactly the same set of interonset intervals, just arranged differently in time; cf. Povel & Essens, 1985.) Thus successful beat perception and synchronization in WM sequences required frequent taps at points with no sound.

All participants were able to synchronize with the beat of the SM sequence: Their taps were very close in time to the idealized beat locations. (In fact, taps typically preceded the beat by a small amount, a finding typical of beat synchronization studies, indicating that beat perception is anticipatory rather than reactive.) Of greater interest was performance on the WM sequences. Although synchronization was not as accurate as with the SM sequences as measured by tapping variability, most participants (even the musically untrained ones) were able to tap to the beat of these sequences, *though from a physical standpoint there was little periodicity at the beat period.* That is, most people tapped to the silent beats as if they were physically there, illustrating that beat perception can tolerate a good deal of counterevidence.

The above facts indicate that beat perception is a complex phenomenon that likely has sophisticated cognitive and neural underpinnings. Specifically, it involves a mental model of time in which periodic temporal expectancies play a key role (Jones, 1976). This may be one reason why it is unique to humans.

Beat perception is an active area of research in music cognition, in which there has long been an interest in the cues listeners use to extract a beat. Temperley and Bartlette (2002) list six factors that most researchers agree are important in beat finding (i.e., in inferring the beat from a piece of music). These can be expressed as preferences:

1. For beats to coincide with note onsets
2. For beats to coincide with longer notes
3. For regularity of beats
4. For beats to align with the beginning of musical phrases
5. For beats to align with points of harmonic change
6. For beats to align with the onsets of repeating melodic patterns

Because beat perception is fundamental to music and is amenable to empirical study, it has attracted computational, behavioral, and neural approaches (e.g., Desain, 1992; Desain & Honing, 1999; Todd et al., 1999; Large, 2000; Toiviainen & Snyder, 2003; Hannon et al., 2004; Snyder & Large, 2005; Zanto et al., 2006) and has the potential to mature into a sophisticated branch of music cognition in which different models compete to explain a common set of behavioral and neural data. Its study is also attractive because it touches on larger issues in cognitive neuroscience. For example, synchronization to a beat provides an opportunity to study how different brain systems are coordinated in perception and behavior (in this case, the auditory and motor systems). A better understanding of the mechanisms involved in beat perception and synchronization could have applications for physical therapy, in which synchronization with a beat is being used to help patients with neuromotor disorders (such as Parkinson's disease) to initiate and coordinate movement (Thaut et al., 1999; cf. Sacks, 1984, 2007).

3.2.2 Meter: Multiple Periodicities

In Western European music, beats are not all created equal. Instead, some beats are stronger than others, and this serves to create a higher level of periodicity in terms of the grouping and/or accentuation of beats. For example, the beats of a waltz are grouped in threes, with an accent on the first beat of each group, whereas in a march beats are grouped into twos or fours, with primary accent on the first beat (in a four-beat march, there is secondary accent on the third beat).

Waltzes and marches are but two types of meter in a broad diversity of meters used in Western European music, but they serve to illustrate some general features of meter in this tradition. First, the meters of Western music are dominated by organization in terms of multiples of two and three in terms of how many beats constitute a basic unit (the measure), and how many subdivisions of each beat there are. For example, a waltz has three beats per measure, each of which can be subdivided into two shorter beats, whereas a march has two (or four) beats per measure, each of which can also be subdivided into two beats. Many other possibilities exist, for example two beats per measure, each of which is subdivided into three beats.[1] The key point is that meter typically has at least one level of subdivision below the beat (London, 2002, 2004:34), in addition to periodicity above the beat created by the temporal patterning of strong beats. One way to represent this is via a metrical grid, which indicates layers of periodicity using rows of isochronous dots. One of these rows represents the tactus, with the row above this showing the periodic pattern of accentuation above the tactus. Other rows above or below the tactus show other psychologically accessible levels of periodicity (Figure 3.2 shows a metrical grid for K0016).

Thus one should be able to tap to any of these levels and still feel synchronized with the music. (The use of dots in metrical grids indicates that meter concerns the perceptual organization of points in time, which in physical terms would correspond to the perceptual attacks of tones; Lerdahl & Jackendoff, 1983.) In grid notation, the relative strength of each beat is indicated by the number of dots above it, in other words, the number of layers of periodicity it participates in. Dots at the highest and lowest level must fall within the "temporal envelope" for meter: Periodicities faster than 200 ms and slower than ~4–6 s

[1] This corresponds to a time signature of 6/8, in contrast to a waltz, which has a time signature of 3/4. As one can see, although 6/8 = 3/4 in mathematical terms, these ratios refer to rather different forms of organization in a musical context.

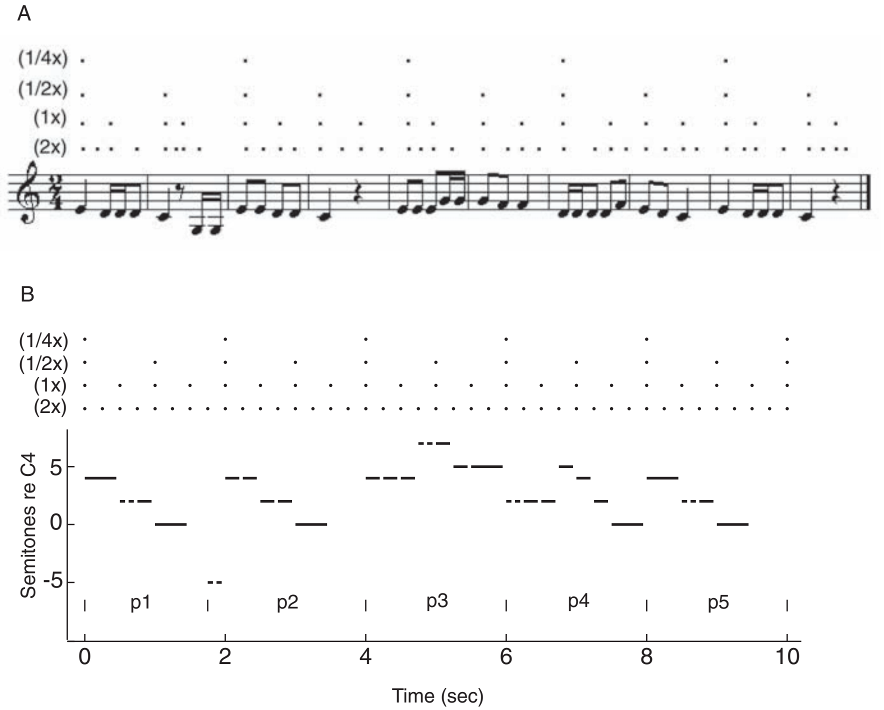

Figure 3.2 Metrical structure of K0016. A typical tactus is shown by the metrical level labeled 1x. Phrase boundaries are indicated below the piano roll notation (p1 = phrase 1, etc.).

are unlikely to be spontaneously perceived as part of a metric framework. (Note that the upper end of this envelope is substantially longer than the ~1.2 second limit for following a beat mentioned in section 3.2.1. That shorter limit refers to beat-to-beat intervals, whereas 4–6 s refers to the highest metrical levels, and is likely to be related to our sense of the psychological present (cf. London, 2002, 2004:30).[2]

Before moving on, the relationship between accent and meter should be discussed. This is an important relationship, because strong beats are perceptually accented points in the music. This kind of accent does not always rely on

[2] An apparent exception occurs in long rhythmic cycles of Indian classical music (e.g., cycles of 16 beats or more), in which the strong accent on the first beat of each cycle can be separated by 10 seconds or more, yet plays an important perceptual role in the music. However, this may be an exception that proves the rule, as explicit counting of the beats by the audience is part of the listening tradition in this music. That is, conscious effort is expended in order to keep track of where the music is in its long metrical cycle.

physical cues such as intensity or duration (note, for example, that all tones in K0016 are of equal intensity), and emerges from the detection of periodicity at multiple time-scales. There are of course many physical (or "phenomenal") accents in music due to a variety of factors, including duration, intensity, and changes in melodic contour. There are also "structural accents" due to salient structural points in the music, for example, a sudden harmonic shift or the start of a musical phrase (Lerdahl & Jackendoff, 1983). The interplay of different accent types is one of the sources of complexity in music (Jones, 1993), particularly the interplay of metrical accents with off-beat phenomenal or structural accents. For example, syncopation in music illustrates the successful use of phenomenal accents "against the grain" of the prevailing meter. This raises a key point about the musical metrical grid, namely that it is a *mental* pattern of multiple periodicities in the mind of a listener, and not simply a map of the accentual structure of a sequence. This point will become relevant in the discussion of metrical grids in language.

The influence of musical meter on behavior, perception, and brain signals has been demonstrated in a number of ways. Sloboda (1983) had pianists perform the same sequence of notes set to different time signatures (in music, the time signature indicates the grouping and accentuation pattern of beats, i.e., the meter). The durational patterning of the performances differed substantially depending on the meter, and in many cases a given pianist did not even realize they were playing the same note sequence in two different meters. A demonstration of meter's effect on synchronization comes from Patel, Iversen, et al. (2005), who showed that tapping to a metrical pattern differs from tapping to a simple metronome at the same beat period. Specifically, taps to the first beat of each metric cycle (i.e., the "downbeats" in the strongly metrical sequences of Sound Example 3.3) were closer to the physical beat than taps on other beats. Importantly, these downbeats (which occurred every four beats) were identical to other tones in terms intensity and duration, so that the influence of downbeats on tapping was not due to any physical accent but to their role in creating a four-beat periodic structure in the minds of listeners.

In terms of meter's influence on perception, Palmer and Krumhansl (1990) had participants listen to a sequence of isochronous tones and imagine that each event formed the first beat of groups of two, three, four or six beats. After a few repetitions, a probe tone was sounded and participants had to indicate how well it fit with the imagined meter. The ratings reflected a hierarchy of beat strength (cf. Jongsma et al., 2004). Turning to neural studies, Iversen, Repp, and Patel (2009) had musically trained participants listen to a metrically ambiguous repeating two-note pattern and mentally impose a downbeat in a particular place. Specifically, in half of the sequences they imagined that the first tone was the downbeat, and in the other half they imagined that the second tone was the downbeat. Participants were instructed not to move or to engage in motor imagery. Measurement of brain signals from auditory regions using magnetoencephalography (MEG) revealed that when a note was interpreted as the downbeat, it evoked an increased amount of neural activity in a particular frequency band (beta, 20–30 Hz) compared to when it was not a downbeat (even though the tones were physically identical in the two conditions).[3] A control experiment showed that the pattern of increased activity closely resembled the pattern observed when the note in question was in fact physically accented (Figure 3.3). These results suggest that the perception of meter involves the active shaping of incoming signals by a mental periodic temporal-accentual scheme.

3.2.3 *Grouping: The Perceptual Segmentation of Events*

Grouping refers to the perception of boundaries, with elements between boundaries clustering together to form a temporal unit. This can be illustrated with K0016. In listening to this melody, there is a clear sense that it is divided into phrases, schematically marked in Figure 3.4.

The perceptual boundaries of the first two phrases are marked by silences (musical rests). Of greater interest are the boundaries at the end of the third and the fourth phrases, which are not marked by any physical discontinuity in the tone sequence, but are nevertheless salient perceptual break points.

As emphasized by Lerdahl and Jackendoff (1983), grouping is distinct from meter, and the interaction of these two rhythmic dimensions plays an important role in shaping the rhythmic feel of music. For example, anacrusis, or upbeat, is a rhythmically salient phenomenon involving a slight misalignment between grouping and meter, in other words, a phrase starting on a weak beat (such as phrase 2 of K0016).

Psychological evidence for perceptual grouping in music comes from a number of sources. Memory experiments show that if a listener is asked to indicate whether a brief tone sequence was embedded in a previously heard longer tone sequence, performance is better when the excerpt ends at a group boundary in the original sequence than when it straddles a group boundary (Dowling, 1973; Peretz, 1989). This suggests that grouping influences the mental chunking of sounds in memory. Further evidence for grouping comes from studies that show how grouping warps the perception of time. For example, clicks placed near phrase boundaries in musical sequences perceptually migrate to those boundaries and are heard as coinciding with them (Sloboda & Gregory, 1980; Stoffer,

[3] Neural activity in the beta frequency band has been associated with the motor system, raising the possibility that meter perception in the brain involves some sort of coupling between the auditory and motor system, even in the absence of overt movement.

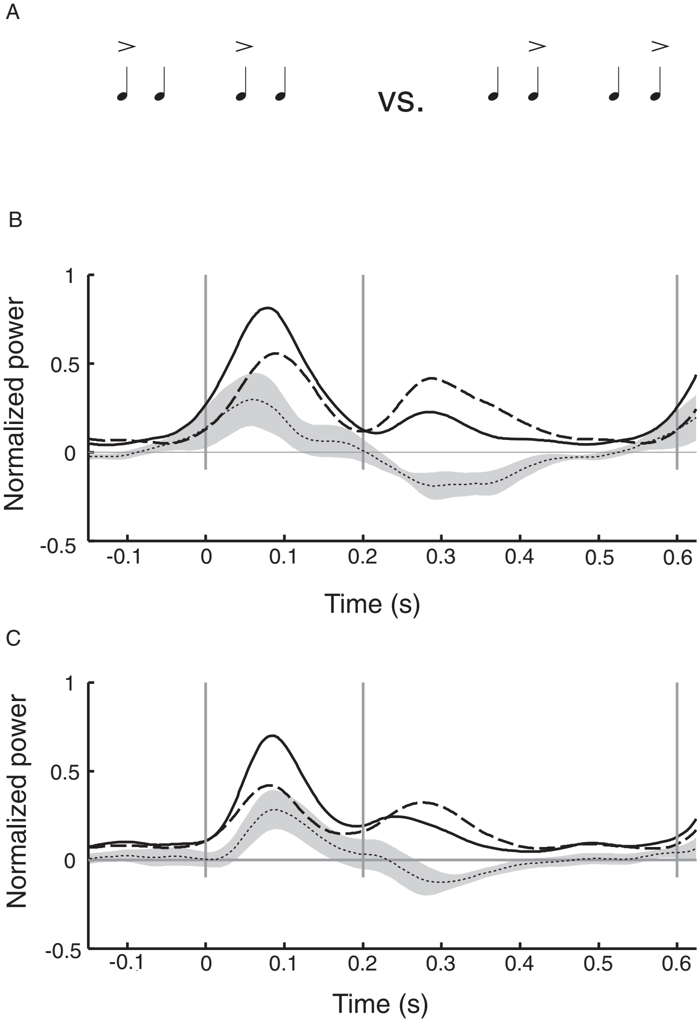

Figure 3.3 (A) Repeating two-note rhythmic pattern, in which the listener imagines the downbeat on either the first tone (left) or second tone (right). (B) Evoked neural responses (measured over auditory brain regions) to the two-tone pattern subjectively interpreted in two different ways, in other words, with the downbeat on tone 1 versus tone 2. (The onset times of tones 1 and 2 are indicated by thin, vertical, gray lines at 0 and 0.2 s). The solid and dashed black lines show across-subject means for the two imagined beat conditions (solid = beat imagined on tone 1, dashed = beat imagined on tone 2). Data are from the beta frequency range (20–30 Hz). The difference is shown by the dotted line, with shading indicating 1 standard error. (C) Evoked neural responses in the beta frequency range to a two-tone pattern physically accented in two different ways, with the accent on tone 1 (solid line) versus tone 2 (dashed line).

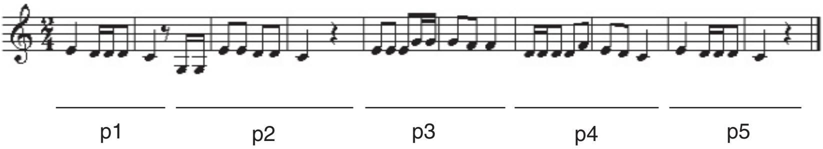

Figure 3.4 K0016 segmented into melodic phrases (p1 = phrase 1, etc.).

1985). More evidence for perceptual warping based on grouping comes from a study by Repp (1992a), in which participants repeatedly listened to a computer-generated isochronous version of the opening of a Beethoven minuet. The task was to detect a lengthening in 1 of 47 possible positions in the music. Detection performance was particularly *poor* at phrase boundaries, probably reflecting an expectation for lengthening at these points (Repp further showed that these were the points at which human performers typically slowed down to mark the phrase structure). Finally, in a gating study of recognition for familiar melodies in which successively longer fragments of tunes were heard until they were correctly identified, Schulkind et al. (2003) found that identification performance was highest at phrase boundaries.[4] Thus there is abundant evidence that grouping plays a role in musical perception.

What cues do listeners use in inferring grouping structure in music? Returning again to K0016, the ends of phrases 3 and 4 are marked by local durational lengthening and lowering of pitch. It is notable that these cues have been found to be important in the prosodic marking of clause endings in speech (Cooper & Sorensen, 1977). Even infants show sensitivity to these boundary cues in both speech and music (Hirsh-Pasek et al., 1987; Krumhansl & Jusczyk, 1990; Jusczyk & Krumhansl, 1993). For example, infants prefer to listen to musical sequences in which pauses are inserted after longer and lower sounds rather than at other locations, presumably because in the former case the pauses coincide with perceptual boundaries. Of course, there is much more to grouping than just these two cues. Deliège (1987) found that salient changes in intensity, duration, pitch, and timbre can all play a role in demarcating the edges of groups. Another factor that is likely to be important is motivic repetition, for example, a repeating pattern of

[4] Schulkind et al. (2003) conducted an interesting analysis of melodic structure that helps suggest why phrase boundaries might be important landmarks in melody recognition. They examined the temporal distribution of pitch and temporal accents in melodies, in which pitch accents were defined as notes that were members of the tonic triad or points of contour change, and temporal accents were defined as relatively long notes and metrically accented notes. They found that accent density was higher at phrase boundaries than within phrases. (For those unfamiliar with the concept of a tonic triad, it is explained in Chapter 5). Thus the edges of phrases are structural slots that attract accents of various kinds.

the same overall duration and internal durational patterning. When these different cues are in conflict, people can disagree about where they hear grouping boundaries (Peretz, 1989). The interaction of different factors in grouping perception is a topic that draws continuing interest, because data on the perceived segmentation of pieces is relatively easy to collect (Clarke & Krumhansl, 1990; Deliège et al., 1996; Frankland & Cohen, 2004; Schaefer et al., 2004).

Grouping plays a prominent role in modern cognitive theories of music, in which it is conceived of as hierarchical, with lower level groups nested within higher level ones. For example, a theoretical analysis of grouping in K0016 would add layers above and below the phrase layer. Below the phrase layer, each phrase would be parsed into smaller groups (motives); above the phrase layer, phrases would be linked into higher level structures. For example, one might unite Phrases 1 and 2 into a group, followed by a group consisting of Phrases 3 and 4, and a final group coincident with Phrase 5. One of the most developed theoretical treatments of hierarchical grouping in music is that of Lerdahl and Jackendoff (1983), who propose certain basic constraints on grouping structure such as the constraint that a piece must be fully parsed into groups at each hierarchical level, and that boundaries at higher levels must coincide with those at lower levels. Evidence for multiple layers of grouping structure in music comes from research by Todd (1985), who showed that the amount of lengthening at a given phrase boundary in music is predicted by the position of that boundary in a hierarchical phrase structure of a piece.

The hierarchical view of grouping structure in music shows strong parallels to theories of prosodic structure in modern linguistic theory, notably the concept of the "prosodic hierarchy" (Selkirk, 1981, Nespor & Vogel, 1983). The prosodic hierarchy refers to the organization of sonic groupings at multiple levels in speech, ranging from the syllable up to the utterance. A key conceptual point made by all such theories is that these groupings are not simple reflections of syntactic organization. To take a well-known example, consider the difference between the syntactic bracketing of a sentence in 3.1a versus its prosodic phrasal bracketing 3.1b (Chomsky & Halle, 1968):

(3.1a) This is [the cat [that caught [the rat [that stole [the cheese]]]]]

(3.1b) [This is the cat] [that caught the rat] [that stole the cheese]

Prosodic grouping reflects a separate phonological level of organization that is not directly determined by syntactic structure. Instead, other linguistic factors play an important role, such as the semantic relations between words and the desire to place focus on certain elements (Marcus & Hindle, 1990; Ferreira, 1991). Furthermore, there are thought to be purely rhythmic factors such as a tendency to avoid groups that are very short or very long, and a tendency to balance the lengths of groups (Gee & Grosjean, 1983; Zellner Keller, 2002). The prosodic grouping structure of a sentence is by no means set in stone: There are differences among individuals in terms of how they group the words of the same sentence, and the grouping structure of a sentence can vary with speech rate (Fougeron & Jun, 1998). Nevertheless, grouping is not totally idiosyncratic, and psycholinguists have made good progress in predicting where speakers place prosodic boundaries in a sentence based on syntactic analyses of sentences (Watson & Gibson, 2004).

Although Example 3.1 above only shows one level of prosodic phrasing, modern theories of the prosodic hierarchy posit multiple levels nested inside one another. Theories vary in the number of levels they propose (Shattuck-Hufnagel & Turk, 1996),[5] so for illustrative purposes only one such theory is discussed here. Hayes (1989) posits a five-level hierarchy comprised of words, clitic groups, phonological phrases, intonational phrases, and utterances. Figure 3.5 shows a prosodic hierarchy for a sentence according to this theory, with the syntactic structure also shown for comparison. (Note that a clitic group combines a lexical word that has a stressed syllable with an adjacent function word—an unstressed syllable—into a single prosodic unit. See Hayes, 1989, for definitions of other units.)

One form of evidence offered for the existence of a given level in the prosodic hierarchy is a systematic variation in the realization of a phonemic segment that depends on prosodic structure at that level. For example, Hayes (1989) discusses /v/ deletion in English speech as an example of a rule that operates within the clitic group. Thus it is acceptable to delete the /v/ in American English when saying, "Will you [save me] a seat?" because "save me" is a clitic group. (That is, if you listen carefully to an American English speaker say this phrase rapidly, "save" is often acoustically realized as "say," though it is intended as—and heard as—"save".) In contrast, the /v/ is not deleted when saying "[save] [mom]" because [save] and [mom] are two separate clitic groups. Other evidence that has been adduced for prosodic constituents includes preferences for interruption points between, rather than within, constituents (Pilon, 1981), and speeded word spotting at the boundaries of constituents (Kim, 2003).

Evidence that the prosodic hierarchy has multiple levels comes from phonetic modifications of speech that vary in a parametric fashion with the height of the prosodic boundary at the phoneme's location (this corresponds to the number of coincident prosodic boundaries at that point, as higher level boundaries are always coincident with lower level ones). For example, Cho and Keating (2001) showed that in Korean, the voice-onset time of stop consonants is larger at higher prosodic boundaries, and Dilley, Shattuck-Hufnagel, and Ostendorf (1996) have shown that the amount of glottalization of word-onset vowels is greater at higher level boundaries.

[5] In a survey of 21 typologically different languages, Jun (2005) found that all languages had at least one grouping level above the word, and most had two.

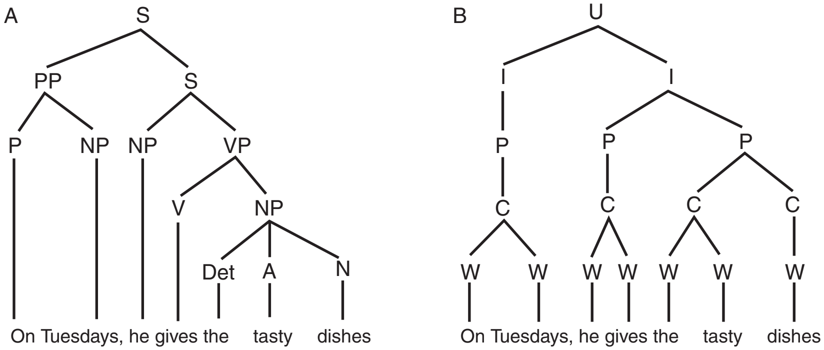
A
S
PP
S
P
NP
NP
VP
V
NP
Det
A
N
On Tuesdays, he gives the tasty dishes
B
U
I
I
P
P
P
C
C
C
C
W
W
W
W
W
W
W
On Tuesdays, he gives the tasty dishes

Figure 3.5 (A) Syntactic and (B) prosodic hierarchy for a sentence of English. Abbreviations for (A): S = sentence, PP = prepositional phrase, NP = noun phrase, VP = verb phrase, Det = Determiner, A = adjective, N = Noun, V = Verb. Abbreviations for (B): U = utterance, I = Intonation phrase, P = phonological phrase, C = clitic group, W = word. Adapted from Hayes, 1989.

Another phenomenon that supports the notion of hierarchical grouping in speech is variation in perceived juncture between words (Jun, 2003). In connected speech, words are acoustically run together and the silent intervals that do occur (e.g., due to stop consonants) are not necessarily located at word boundaries (cf. Chapter 2, section 2.3.3, subsection "A Brief Introduction to the Spectrogram"). Nevertheless, words are perceived as separated from one another. Unlike with written language, however, the perceived degree of spacing is not identical between each pair of words: Rather, some word boundaries seem stronger than others. For example, the sentence in 3.1c below contains juncture markings from a system devised by Price et al. (1991). In this system, a researcher listens repeatedly to a given sentence and places numerals from 0 to 6 between each pair of words to indicate the degree of perceived separation between them. A "break index" of 0 indicates the weakest perceived juncture, in other words, between the words of a clitic group. At the opposite extreme, a break index of 6 indicates the end of a sentence.

(3.1c) Only 1 one 4 remembered 3 the 0 lady 1 in 1 red 6.

Wightman et al. (1992) studied the relationship between these break indices and speech duration patterns in a large speech corpus, and found a correlation between perceived boundary strength and amount of lengthening of the syllable preceding the boundary (cf. Gussenhoven & Rietveld, 1992).[6] This finding is

[6] More specifically, the lengthening was confined to the syllabic rime (the vowel and following consonants). Interestingly, this asymmetric expansion of the syllable due to prosodic boundaries differs from temporal changes in a syllable due to stress (Beckman et al., 1992).

strikingly reminiscent of the research on music by Todd (1985) described above. Another parallel to music is that durational lengthening interacts with pitch and amplitude cues in determining the perceived strength of prosodic boundaries (Streeter, 1978; de Pijper & Sanderman, 1994).

In conclusion, grouping is a fundamental rhythmic phenomenon that applies to both musical and linguistic sequences. In both domains, the mind parses complex acoustic patterns into multiple levels of phrasal structure, and music and language share a number of acoustic cues for marking phrase boundaries. These similarities point to shared cognitive process for grouping across the two domains, and indicate that grouping may prove a fruitful area for comparative research. As discussed in section 3.5, empirical work is proving these intuitions correct.

3.2.4 Durational Patterning in Music

Up to this point, the discussion of musical rhythm has been concerned with points and edges in time: beats and grouping boundaries. A different set of issues in rhythm research concerns how time gets filled, in other words, the durational patterning of events.

Duration Categories in Music

In music, the durational patterning of events is typically measured by the time intervals between event onsets within a particular event stream: This defines a sequence of interonset intervals (IOIs). For example, the sequence of IOIs between the tones of a melody defines the durational patterning of that melody. Typically, durations tend to be clustered around certain values reflecting the organization of time in music into discrete categories. Fraisse (1982) pointed out that two categories that figure prominently in Western musical sequences are short times of 200–300 ms and long times of 450–900 ms (cf. Ross, 1989). He argued that these two duration categories were not only quantitatively different but also different in terms of their perceptual properties: Long intervals are perceived as individual units with distinct durations, whereas short intervals are perceived collectively in terms of their grouping patterns rather than in terms of individual durations.

There is empirical evidence that durations in musical rhythms are perceived in terms of categories (Clarke, 1987; Schulze, 1989). For example, Clarke (1987) had music students perform a categorical perception experiment on rhythm. Participants heard short sequences of tones in which the last two tones had a ratio that varied between 1:1 and 1:2. Listeners had to identify the final duration ratio as one or the other of these, and also had to complete a task that required them to discriminate between different ratios. The results showed a steep transition in the identification function, and increased discrimination when stimuli were near the boundary versus within a given region. (Clarke also found that the location of the boundary depended on the metrical context in which the sequence was perceived, providing another example of the influence of meter on perception; cf. section 3.2.2.)

In speech, the duration of basic linguistic elements (such as phonemes and syllables) is influenced by a number of factors. For example, there are articulatory constraints on how fast different sounds can be produced, which creates different minimum durations for different sounds (Klatt, 1979). There are also systematic phonological factors that make some sounds longer than others. For example, in English, the same vowel tends to be longer if it occurs before a final stop consonant that is voiced rather than unvoiced (e.g., the /i/ in "bead" vs. "beet"), and this difference influences the perception of the final stop as voiced or voiceless (Klatt, 1976). A simple phonological factor that influences syllable duration is the number of phonemes in the syllable: Syllables with more phonemes tend to be longer than those with fewer phonemes (e.g., "splash" vs. "sash"; Williams & Hiller, 1994). Atop these sources of variation are other sources including variations in speaking style (casual vs. clear), and variations in speech rate related to discourse factors, such as speeding up near the end of a sentence to "hold the floor" in a conversation (Schegloff, 1982; Smiljanic & Bradlow, 2005). Given all these factors, it is not surprising that the durations of speech elements do not tend to cluster around discrete values. Instead, measurements of syllable or phoneme duration typically reveal a continuous distribution with one main peak. For example, Figure 3.6a shows a histogram of syllable durations for a sample of spoken English.

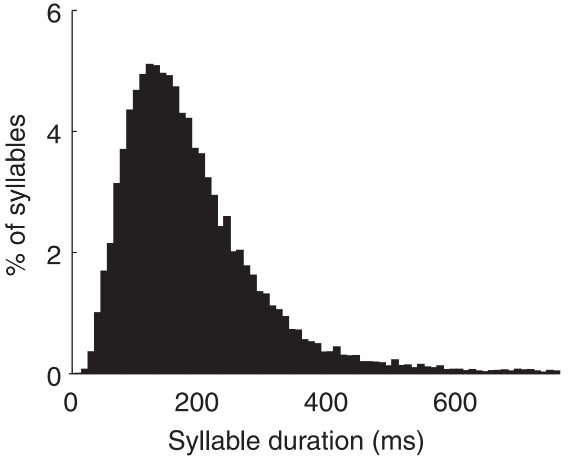

Figure 3.6a Histogram of syllable durations in a corpus of spontaneous speech in American English. Data are from approximately 16,000 syllables. Mean syllable duration = 191 ms, sd = 125 ms. Syllables with duration >750 ms are not shown (<1% of total). Histogram bin size = 10 ms. Analysis based on data from Greenberg, 1996.

Having said this, it is important to note that durational categories do occur in some languages. For example, there are languages with phonemic length contrasts in which the same word can mean entirely different things when a short versus long version of the same vowel or consonant is used. In some languages, such as Estonian, there can even be three-way length contrasts. For example, "sata" can mean three entirely different things ("hundred," "send," and "get") depending on the length of the first /a/. It would be interesting to study length contrasts in a given vowel phoneme and examine the amount of temporal variability within each duration category in connected speech. This could be compared to temporal variability of a given duration category in music, to see whether the perceptual system has a similar tolerance for within-category variability in the two domains.[7]

Expressive Timing in Music

If the perceptual system cared only about musical durations as a sequence of discrete categories, then computer renditions of musical pieces based on exact renderings of music notation would be perfectly acceptable to listeners. Although such mechanical performances do occur in some settings (e.g., rhythm tracks in some modern popular music), in other contexts, such as the classical piano repertoire, such performances are rejected as unmusical. Not surprisingly then, physical measurements of human performances reveal considerable deviations from notated durations. For example, Figure 3.6b shows a histogram of IOIs, all of which represent realizations of notes with the *same notated duration* (an eighth note or quaver) from a famous pianist's rendition of Schumann's Träumerei (Repp, 1992b).[8] Had the piece been performed by a machine, all of these IOIs would be a single value. Instead, considerable variation is seen. The crucial fact about this variation is that it is not "noise": It largely represents structured variation related to the performer's interpretation of the piece (Palmer, 1997; Ashley, 2002). For example, Repp (1992b) studied several famous pianists' renderings of Träumerei and found that all showed slowing of tempo at structural boundaries, with the amount of slowing proportional to the importance of the boundary (cf. Todd, 1985). At a finer timescale, Repp found that within individual melodic phrases there was a tendency to accelerate at the beginning and slow near the end, with the pattern of IOIs following a smooth parabolic function. Repp speculated that this pattern may reflect principles of

[7] In doing this research, it would be important to be aware of durational lengthening in the vicinity of phrase boundaries in both speech and music: If there are different degrees of preboundary lengthening in the two domains, then events near boundaries should be excluded from the analysis as this would be confounded with variability measures.

[8] I am grateful to Bruno Repp for providing me with this data.

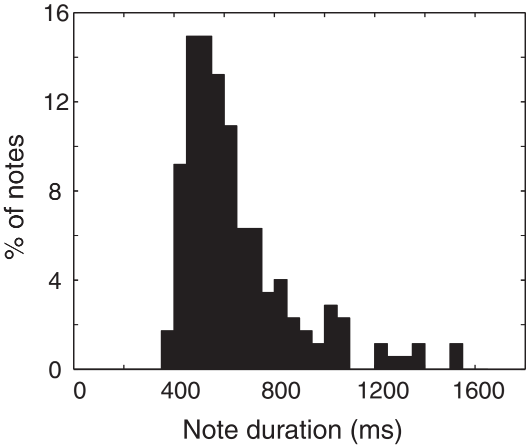

Figure 3.6b Histogram of durations of eighth notes from a performance of Schumann's Träumerei by Claudio Arrau. The large values in the right tail of the histogram are due to phrase-final ritards. Data are from approximately 170 eighth notes. Mean note duration = 652 ms, sd = 227 ms. Notes with duration > 1,600 ms are not shown (<1% of total). Histogram bin size = 50 ms.

human locomotion, in other words, a musical allusion to physical movement (cf. Kronman & Sundberg, 1987).

The above paragraph focuses on the role of IOIs in expressive timing. IOIs are the basis of "expressive timing profiles," time series that show the actual pattern of event timing versus the idealized pattern based on notated duration. Although studies of these profiles have dominated research on expressive timing, it is important not to overlook another aspect of expressive timing, namely, articulation. Although IOI refers to the time interval between the onsets of successive tones, articulation refers to the time between the *offset* of one tone and the onset of the next. If there is little time between these events (or if the tones overlap so that the offset of the prior tone occurs after the onset of the following tone, which is possible in piano music), this is considered "legato" articulation. In this type of articulation, one tone is heard as flowing smoothly into the next. In contrast, staccato articulation involves a salient gap between offset and onset, giving the tones a rhythmically punctuated feel. In addition to IOI and articulation patterns, another important cue to musical expression is the patterning of tone intensity.

Due to the fact that timing, articulation and intensity in music can be measured with great precision using modern technology (e.g., using pianos with digital interfaces, such as the Yamaha Disklavier), expression has been a fruitful area of research in studies of music production. There has also been some research on expressive features in perception. For example, listeners can reliably identify performances of the same music as expressive, deadpan (mechanical), or exaggerated (Kendall & Carterette, 1990), and can identify the performer's intended emotion on the basis of expressive features (Gabrielsson & Juslin, 1996).

Palmer (1996) has shown that musically trained listeners can identify a performer's intended metrical and phrase structure on the basis of expressive cues. One clever demonstration of the perceptual importance of expressive timing was provided by Clarke (1993), who used naturally performed short melodies. For each melody, Clarke extracted its expressive timing profile, manipulated it, and then reimposed it on a mechanical performance of the melody, thus creating a Frankensteinian melody with structure and expression mismatched. For example, in one condition the original note-by-note expressive timing profile was shifted several notes to the right. Musicians judged the originals versus the mismatched melodies in terms of the quality of performance, and favored the originals. Thus listeners are sensitive to the way expressive timing aligns with the structure of musical passages.

Expressive timing in music has an interesting relationship to prosodic structure in speech. Just as a musical passage played by different performers will have different expressive timing patterns, the same sentence spoken by different speakers will have a different temporal patterning of syllables and phonemes. In the past, researchers have suggested that these individualistic aspects of performance are "normalized away" in memory for musical and spoken sequences, arguing that the abstract memory representation favors a less detailed, more categorical structure (Large et al., 1995; Pisoni, 1997). More recent research, however, suggests that listeners retain some temporal information in memory for speech and music (Bradlow et al., 1999; Palmer et al., 2001). For example, Palmer et al. (2001) familiarized listeners with particular performances of short melodic sequences, and then later tested the ability to recognize these performances against other performances of the same sequences. The different performances were generated by a pianist who produced the same short melodic sequences as part of longer melodies that differed in their metrical structure (3/4 vs. 4/4 time). As a result of the differing metrical structure, the same melodic sequence was produced with different patterns of articulation and intensity. For each such melodic sequence, both musicians and nonmusicians were able to recognize the original version they had heard when presented with it versus another version. Furthermore, even 10-month-old infants discriminated between familiar and unfamiliar performances, orienting longer toward the former. Palmer et al. relate these findings to research in speech perception showing that listeners retain stimulus-specific acoustic properties of words along with abstract linguistic properties (Luce & Lyons, 1998).

Another line of research relating timing in music to speech prosody concerns "tempo persistence." Jungers et al. (2002) had pianists alternate between listening to short melodies and sight-reading different short melodies. The participants were told to attend to both the heard and performed melodies for a later memory test. In fact, the real question of interest was the relationship between the tempo of the heard and performed melodies. The heard melodies occurred in blocks of slow and fast tempi, and Jungers et al. found that the tempo of performed melodies was influenced by the tempo of heard melodies: The pianists played more slowly after slow melodies and faster after fast melodies. A similar experiment using spoken sentences rather than melodies showed a similar tempo persistence effect in speech. These findings are reminiscent of research on "accommodation" in sociolinguistics, which has shown that when people of different social backgrounds meet, their speech becomes more alike (cf. Giles et al., 1991).

In an interesting follow-up study, Dalla Bella et al. (2003) studied tempo persistence across modalities. Listeners (both musicians and nonmusicians) alternated between hearing melodies and reading sentences aloud. The musicians showed a tempo persistence effect: They spoke faster after hearing faster melodies. However, the nonmusicians showed no such effect. Furthermore, when the musicians did the reverse experiment (in which they alternated between hearing sentences and sight-reading melodies), there was no evidence of tempo persistence. Dalla Bella et al. suggest that musicians may be better than nonmusicians at beat extraction in music, and that this may drive the effect of tempo persistence seen in their first study. Following this logic, I would suggest that the lack of an effect in their second study indicates that speech perception does not involve extraction of a beat.

3.2.5 The Psychological Dimensions of Musical Rhythm

The interactions of beat, meter, accent, grouping, and expressive timing make musical rhythm a psychologically rich phenomenon (and this is just within the confines of Western European music!). Some idea of this richness is suggested by the work of Gabrielsson, who has conducted studies in which a variety of rhythms are compared and classified by listeners using similarity judgments and adjective ratings (reviewed in Gabrielsson, 1993). Statistical techniques such as multidimensional scaling and factor analysis are used to uncover the perceptual dimensions involved in the experience of musical rhythms. This research has revealed an astonishingly large number of dimensions (15), which group broadly into those concerned with structure (e.g., meter, simplicity vs. complexity), motion (e.g., swinging, graceful), and emotion (e.g., solemnity vs. playfulness). Although much of the cognitive science of musical rhythm focuses on structural issues, it is important to keep the links to motion and emotion in mind, for these connections are part of what distinguishes musical rhythm from speech rhythm, a point to which I will return at the end of the discussion of rhythm in speech.

3.3 Rhythm in Speech

Although the study of rhythm in poetry has a long history, dating back to ancient Greek and Indian texts, the study of rhythm in ordinary language is a relatively recent endeavor in linguistics. Researchers have taken at least three approaches to this topic. The first approach is typological, and seeks to understand the rhythmic similarities and differences among human languages. The driving force behind this work has been idea that linguistic rhythms fall into distinct categories. For example, in one widespread typological scheme (discussed in the next section), English, Arabic, and Thai are all members of a single rhythmic class ("stress-timed languages"), whereas French, Hindi, and Yoruba are members of a different class ("syllable-timed languages"). As is evident from this example, membership in a rhythmic class is not determined by the historical relationship of languages; rhythm can unite languages that are otherwise quite distant both historically and geographically.

The second approach to speech rhythm is theoretical, and seeks to uncover the principles that govern the rhythmic shape of words and utterances in a given language or languages. This research, which includes an area called "metrical phonology," seeks to bring the study of the linguistic rhythm in line with the rest of modern linguistics by using formalized rules and representations to derive the observed rhythmic patterning of utterances.

The third approach is perceptual, and examines the role that rhythm plays in the perception of ordinary speech. One prominent line of research in this area concerns the perceptual segmentation of words from connected speech. Another, smaller line of research examines the effects of rhythmic predictability in speech perception.

The goal of this part of the chapter is to introduce each of these areas and make comparisons to musical rhythm when appropriate. Before commencing, it is worth introducing a concept that occurs in each section: the notion of prominence in speech. In many languages, it is normal to produce the syllables of an utterance with differing degrees of prominence. This is true even when a sentence is said with no special emphasis on any particular word. For example, when speaking the following sentence, note how the syllables marked by an x are more prominent than their neighbors:

```
             x           x              x   x              x               x
(3.2) She wrote all her novels with a blue pen that she inherited from her aunt.
```

The most important physical correlates of prominence are duration, pitch movement, vowel quality, and loudness.[9] Prominence in speech raises many

[9] Perceived loudness incorporates both physical intensity and the distribution of energy across different frequency bands, in other words, "spectral balance." The latter may be

interesting questions. How many different degrees of prominence can listeners reliably distinguish (Shattuck-Hufnagel & Turk, 1996)? Do languages differ in the extent to which they rely on particular acoustic cues to prominence in production and perception (Berinstein, 1979; Lehiste & Fox, 1992)? Empirical data on these issues is still relatively sparse, and we will not delve into them here. Instead, most sections below treat prominence as a binary quantity referred to as "stress," following the tradition of much work on speech rhythm. An exception occurs in section 3.3.2, where degrees of prominence are discussed in the context of modern linguistic theories of speech rhythm.

Before embarking on the following sections, a word should be said about the concept of stress in linguistics. Stress is recognized as one aspect of word prosody in human languages; tone and lexical pitch accent are two other aspects. Just as not all languages have lexical tone (cf. Chapter 2 for a discussion of tone languages) or lexical pitch accent,[10] not all languages have lexical stress, in other words, a systematic marking of certain syllables within a word as more prominent than others. Importantly, these three aspects of word prosody are not mutually exclusive. For example, there are tone languages with stress (e.g., Mandarin) and without it (e.g., Cantonese), and pitch-accent languages with or without stress (e.g., Swedish and Japanese, respectively; Jun, 2005). Thus in the discussion below, it should be kept in mind that stress is a widespread but not universal feature of human language.

3.3.1 *Rhythmic Typology*

Four approaches to rhythmic typology are described below. Behind all of these approaches is a common desire to understand the relationships of the world's linguistic rhythms.

Periodicity and Typology

The most influential typology of language rhythm to date is based on the notion of periodicity in speech. This typology has its roots in the work of Kenneth Pike (1945), who proposed a theory of speech rhythm based on a dichotomy between languages in terms of syllable and stress patterns. He dubbed certain languages (such as Spanish) "syllable-timed," based on the idea that syllables

a more salient and reliable cue than the former (Sluijter & van Heuven, 1996, Sluijter et al., 1997).

[10] In a pitch-accent language, a word can have entirely different meaning depending on its pitch pattern. The difference between a tone language and a pitch-accent language is that in the former there is a prescribed pitch for each syllable, whereas in pitch-accent languages a certain syllable of a word may have lexical specification for pitch (Jun, 2005).

mark off roughly equal temporal intervals. These stood in contrast to "stress-timed" languages such as English, which were characterized by roughly equal temporal intervals between stresses. To illustrate stress-timed rhythm, Pike invited the reader to "notice the more or less equal lapses of time between the stresses in the sentence":

```
           x         x             x            x
(3.3) The teacher is interested in buying some books.
```

Pike then asked the reader to compare the timing of stresses in the above sentence with the following one, and notice the similarity "despite the different number of syllables" (p. 34):

```
       x   x           x      x
(3.4) Big battles are fought daily.
```

Pike argued that in stress-timed languages the intervals between stressed syllables (referred to as "feet") were approximately equal despite changing numbers of syllables per foot. To achieve evenly timed feet, speakers would stretch or compress syllables to fit into the typical foot duration. Pike believed that learning the rhythm of a language was essential to correct pronunciation. He noted, for example, that Spanish speakers learning English "must abandon their sharp-cut syllable-by-syllable pronunciation and jam together—or lengthen where necessary—English vowels and consonants so as to obtain rhythm units of the stress-timing type" (p. 35).

Abercrombie (1967:34–36, 96–98) went further than Pike and proposed a physiological basis for stress versus syllable timing. This bold step was based on a specific hypothesis for how syllables are produced. Abercrombie believed that each syllable was associated with a contraction of muscles associated with exhalation (the intercostal muscles of the rib cage), and that some contractions were especially strong: These latter contractions produced stressed syllables. He referred to these two types of contractions as "chest pulses" and "stress pulses" (thus only some chest pulses were stress pulses; cf. Stetson 1951). Abercrombie proposed that in any given language, one or the other kind of pulse occurred rhythmically. He then equated rhythm with periodicity: "Rhythm, in speech as in other human activities, arises out of the periodic recurrence of some sort of movement. . . " (p. 96). Furthermore, he claimed that "as far as is known, every language in the world is spoken with one kind of rhythm or with the other" (p. 97), naming English, Russian, and Arabic as examples of stress-timed languages, and French, Telugu, and Yoruba as examples of syllable-timed languages. Just as Pike had done, he noted that a language could not be both stress-timed and syllable-timed. Because there are variable numbers of syllables between stresses, equalizing the duration of interstress intervals meant that "the rate of syllable succession has to be continually adjusted, in order to fit varying numbers of syllables into the same time interval."

It is hard to overestimate the impact of Pike and Abercrombie on the study of rhythm in speech. The terms "stress-timed" and "syllable-timed" have become part of the standard vocabulary of linguistics. A third category, "mora-timing," is also in standard use, and is used to describe the rhythm of Japanese speech. The mora is a unit that is smaller than the syllable, usually consisting of a consonant and vowel, but sometimes containing only a single consonant or vowel. Ladefoged (1975:224) stated that "each mora takes about the same length of time to say," thus arguing for the rough isochrony of morae.[11] Since the publication of Abercrombie's book, many languages have been classified into one of these two categories (Dauer, 1983; Grabe & Low, 2002), and many research studies have examined the issue of isochrony in speech. In this sense, the stress versus syllable-timed theory of speech rhythm has been very fruitful. It provided a clear, empirically testable hypothesis together with a physiological justification.

In another sense, however, the theory has been an utter failure. Empirical measurements of speech have failed to provide any support for the isochrony of syllables or stresses (see references in Bertinetto, 1989).[12] To take just a few examples from the many papers that have tested the isochrony hypothesis, Dauer (1983) showed that English stress feet grow in duration with increasing number of syllables, rather than maintaining the even duration necessary for isochrony (cf. Levelt, 1989:393). Roach (1982) compared English, Russian, and Arabic to French, Telugu, and Yoruba and demonstrated that the former stress-timed languages could not be discriminated from the latter syllable-timed ones on the basis of the timing of interstress intervals. Finally, Beckman (1982) and Hoequist (1983) showed that morae are not of equal duration in Japanese.

Given that the notion of periodicity in ordinary speech was empirically falsified over 20 years ago, why do the labels of stress-timing, syllable-timing, and mora-timing persist? One reason may be that it matches subjective intuitions about rhythm. For example, Abercrombie himself (1967:171) noted that the idea of isochronous stress in English dates back to the 18th century. Another reason is suggested by Beckman (1992), who argues that this tripartite scheme persists because it correctly groups together languages that are perceived as rhythmically similar, even if the physical basis for this grouping is not clearly understood (and is not isochrony of any kind).

[11] As an example, Ladefoged points out that the word *kakemono* (scroll) takes about the same amount of time to say as "nippon" (Japan), and attributes this to the fact that both words contain four morae: [ka ke mo no] and [ni p po n].

[12] Abercrombie's theory of syllables as rooted in chest pulses has also been falsified. It should be noted that Abercrombie was a pioneering scientist who established one of the first laboratories devoted to basic research in phonetics (in Edinburgh). His ideas about speech rhythm are but a tiny slice of his work, and though wrong, stimulated a great deal of research.

The key point of the current section, then, is that periodicity, which plays such an important role in much musical rhythm, is not part of the rhythm of ordinary speech. The next section explores a different approach to speech rhythm, one that sets aside notions of isochrony.

Phonology and Typology

The fact that speech is not isochronous should not lead us to discard the idea of speech rhythm. That is, research can move forward if one thinks of rhythm as systematic timing, accentuation, and grouping patterns in a language *that may have nothing to do with isochrony.* One productive approach in this framework is the phonological approach to rhythmic typology. The fundamental idea of this approach is that the rhythm of a language is the *product* of its linguistic structure, not an organizational *principle* such as stress or syllable isochrony (Dauer, 1983; cf. Dasher & Bolinger, 1982). In this view, languages are rhythmically different because they differ in phonological properties that influence how they are organized as patterns in time. One clear exposition of this idea is that of Dauer (1983, 1987), who posited several factors that influence speech rhythm.

The first factor is the diversity of syllable structures in a language.[13] Languages vary substantially in their inventory of syllable types. For example, English has syllables ranging from a single phoneme (e.g., the word "a") up to seven phonemes (as in "strengths"), and allows up to three consonants in onset and coda. In sharp contrast, languages such as Japanese (and many Polynesian languages) allow few syllable types and are dominated by simple CV syllables. Romance languages such as Spanish and French have more syllable types than Japanese or Hawaiian but avoid the complex syllables found in languages such as English and Dutch, and in fact show active processes that break up or prevent the creation of syllables with many segments (Dauer, 1987).

The diversity of syllables available to a language influences the diversity of syllable types in spoken sentences. For example, Dauer (1983) found that in a sample of colloquial French, over half the syllable tokens had a simple CV structure, whereas in a similar English sample, CV syllables accounted for only about one-third of the syllable tokens. These differences are relevant to rhythm because syllable duration is correlated with the number of phonemes per syllable, suggesting that sentences of English should have more variable syllable durations (on average) than French sentences.

[13] Syllables are generally recognized as having three structural slots: the onset consonant(s), the nucleus (usually occupied by a vowel), and the following consonants (referred to as the coda). A syllable with one consonant in the onset and none in the coda is represented by CV, whereas CCVC means two consonants in the onset and one in the coda, and so on.

The second factor affecting speech rhythm is vowel reduction. In some languages, such as English, unstressed syllables often have vowels that are acoustically centralized and short in duration (linguists commonly refer to this sound as "schwa," a neutral vowel sounding like "uh"). In contrast, in other languages (such as Spanish) the vowels of unstressed syllables are rarely if ever reduced, contributing to a less variable pattern of vowel duration between stressed and unstressed syllables.

The third rhythmic factor proposed by Dauer is the influence of stress on vowel duration. In some languages, stress has a strong effect on the duration of a vowel in a syllable. For example, one recent measurement of spoken English finds that vowels in stressed syllables are about 60% longer than the same vowels in unstressed syllables (Greenberg, 2006). In contrast, studies of Spanish suggest that stress does not condition vowel duration to the same degree (Delattre, 1966).

Dauer suggested that languages traditionally classified as stress-timed versus syllable-timed differ in the above phonological features, with stress-timed languages using a broader range of syllable types, having a system of reduced vowels, and exhibiting a strong influence of stress on vowel duration. This nicely illustrates the perspective of speech rhythm as a product of phonology, rather than a causal principle (e.g., involving periodicity).[14]

Dauer's proposal leads to testable predictions. Specifically, the three factors she outlines (diversity in syllable structure, vowel reduction, and the influence of stress on vowel duration) should all contribute to a greater degree of durational variability among the syllables of stress-timed versus syllable-timed utterances. Surprisingly, there is little published data on durational variability of syllables in sentences of stress versus syllable-timed languages. One reason for this may be that the demarcation of syllable boundaries in speech is not always straightforward. Although people generally agree on how many syllables a word or utterance has, there can be disagreement about where the boundaries between syllables are, even among linguists. For example does the first "l" in the word "syllable" belong to the end of the first syllable or to the beginning of the second syllable, or is it "ambisyllabic," belonging to both syllables? Although it is true that syllable measurements are subject to decisions that may vary from one researcher to the next, this should not impede empirical research: It simply means that measurements should be accompanied by an indication of where each syllable boundary was placed. I return to this point below.

[14] Dauer also suggested that stress- and syllable-timed languages had a different relationship between stress and intonation: In the former, stressed syllables serve as turning points in the intonation contour, whereas in the latter, intonation and stress are more independent. As intonation is not discussed in this chapter, this idea will not be pursued here.

Before turning to another phonological approach to speech rhythm, it is worth noting that the phonological properties listed by Dauer do not always co-occur. Thus Dauer argued against the idea of discrete rhythmic classes and for the notion of a rhythmic continuum. In support of this idea, Nespor (1990) has noted that Polish has complex syllable structure but no vowel reduction (at normal speech rates), and Catalan has simple syllable structure but does have vowel reduction. Thus there is currently a debate in the field of speech rhythm as to whether languages really do fall into discrete rhythm classes or whether there is a continuum based on the pattern of co-occurrence of rhythmically relevant phonological factors (cf. Arvaniti, 1994; Grabe & Low, 2002). Only further research can resolve this issue, particularly perceptual research (as discussed below in section 3.3.1, subsection "Perception and Typology").

I now turn briefly to a different phonological theory of speech rhythm, proposed by Dwight Bolinger (1981). Although Bolinger focused on English, his ideas are quite relevant to typological issues. The foundation of Bolinger's theory is the notion that there are two distinct sets of vowels in English: full vowels and reduced vowels. By "reduced" vowels Bolinger does not simply mean vowels in unstressed syllables that are short and acoustically centralized (i.e., a phonetic definition). He argues for a phonological class of reduced vowels in English, which behave differently from other vowels. Bolinger places three vowels in this class, an "ih"-like vowel, and "uh"-like vowel, and a "oh"-like vowel (more similar to "uh" than to the full vowel "o"). Phonetically all of these vowels occur in the central region of vowel space, near the schwa vowel /ə/ of English (see Figure 2.19: Bolinger's "ih" and "oh" vowel are not shown in that figure, but the former would occur just to the left and up from /ə/, and the latter would occur just to the right and up from /ə/). Bolinger (1981:3–9) presents arguments to support the notion that these vowels are a phonologically distinct subclass, in other words, that they behave in certain ways that full vowels do not. Space limitations prevent a detailed discussion of these arguments. Here I will focus on two claims Bolinger makes about full and reduced vowels that are relevant for speech rhythm.

First, he claims that syllables containing full and reduced vowels tend to alternate in English sentences. Second, he claims that there is a "lengthening rule" such that "when a long syllable is followed by a short one, the short one borrows time from it and makes it relatively short" (p. 18). (By a "long" syllable, he means a syllable with a full vowel, and by a "short" syllable, he means a syllable with a reduced vowel; there is no claim for a particular duration ratio between the two types of syllables.) To illustrate this rule, Bolinger offers the following example (note that the first sentence is from an ad for a special type of soap):

(3.5) Gets out dirt plain soap can't reach.
L L L L L L L

(3.6) Takes a-way the dirt that com-mon soaps can nev-er reach
L^- S L^- S L^- S L^- S L^- S L^- S L

In the example above, I have indicated the shortened L's of the second sentence by L^- (after Faber, 1986). The point of this example is that each L^- of sentence 3.6 is shorter than the L's of sentence 3.5, and this occurs (according to Bolinger) because each S "borrows time" from the preceding L. Note that sentence 3.6 has strict alternation between L and S. This is a special case: Bolinger makes no claims for strict alternation, only a claim for a tendency (thus sequences such as L L S S S L S L L . . . are perfectly possible). I suspect Bolinger chose the sentences in 3.5 and 3.6 as examples because he felt that each (L^- S) pair in sentence 3.6 is not terribly different in duration from each L in sentence 3.5: This is suggested by his graphical placement of the L's in the two sentences above one another, in his original text. However, the durational equivalence of L and (L^- S) is not part of Bolinger's claim. This is an important point. Bolinger's theory may be relevant to the subjective impression of isochrony (because of the rough alternation of L and S and the lengthening rule), but it has no isochrony principle.

Faber (1986) argues that Bolinger's theory is superior to stress-timing theory when it comes to teaching the rhythm of English to foreign students (cf. Chela-Flores, 1994). He also points out that Bolinger's theory can be used to explain characteristic timing patterns that stress-timing theory cannot account for, such as why "cart" is shorter in:

(3.7) Ask Mr. Carter
L - S

than in:

(3.8) Ask Mr. Cartwright.
L L

Or why "man" is shorter in:

(3.10) Have you seen the manor?
L- S

than in:

(3.10) Have you seen the manhole?
L L

Bolinger's theory of speech rhythm is distinct from the theory outlined by Dauer in that it deals not just with the variability syllable duration but with the patterning of duration. Specifically, Bolinger argues that the characteristic rhythm of English is due to the rough *alternation* of syllables with full and reduced vowels, and to the way full vowels *change* duration when intervening reduced syllables are added. This is already enough to suggest a basis for typological distinctions between languages. For example, one might test the idea that stress-timed languages have more contrast in adjacent vowel durations than do syllable timed languages, and that stress-timed languages have lengthening rules of the type suggested by Bolinger for English (Bolinger himself does not suggest these ideas, but they are an obvious corollary of his work). If Bolinger had stopped here, he would already have made a valuable contribution to speech rhythm research. Bolinger's theory has one further component, however, that represents a fundamental divergence from the theory outlined by Dauer.

Once again focusing on English, Bolinger suggested that just as there are two kinds of vowels (full and reduced), there are also two kinds of rhythm. The first is the rhythmic patterning already described, in other words, the rough alternation of long and short syllables and the lengthening rule. Above this level, however, is a second level of rhythmic patterning concerned with temporal relations between accents cued by pitch. Note that this idea entails the notion that syllabic rhythm is fundamentally about duration and does not rely on pitch as a cue. In other words, "there is a basic level of temporal patterning that is independent of tonal patterning" (Bolinger 1981:24, citing Bruce, 1981). Bolinger argues that this temporal patterning would be observable even in speech spoken on a monotone. Speech is not spoken on a monotone, however, and Bolinger argues that syllables accented by pitch form a second level of rhythmic patterning in which the fundamental rule is *a tendency to separate pitch accents so that they do not occur too closely together in time.* The mechanism for avoiding "accent clash" is to move adjacent accents away from each other, a phenomenon sometimes called "stress-shift" in English. (One oft-cited example of stress shift is when "thirtéen" becomes "thírteen mén"; Liberman & Prince, 1977.) The term "stress-shift" is somewhat unfortunate, because there is evidence that what is shifting is pitch accent, not syllable duration or amplitude (Shattuck-Hufnagel et al., 1994).

The idea that speech rhythm involves temporal patterning at two distinct linguistic levels merits far more empirical research than it has garnered to date. I will return to this idea in section 3.3.4.

Duration and Typology

Until very recently, the measurement of duration has had a largely negative role in the study of speech rhythm, namely in falsifying claims for the periodicity of stresses or syllables. The insights of the phonological approach, however, have created a new positive role for durational measurements. A key feature of this work has been the abandonment of any search for isochrony, and a focus on durational correlates of phonological phenomena involved in speech rhythm. Ramus and colleagues (1999), inspired by the insights of Dauer, examined the durational pattering of vowels and consonants in speech, based on ideas about how syllable structure should influence this patterning. For example, languages that use a greater variety of syllable types (i.e., stress-timed languages) are likely to have relatively less time devoted to vowels in sentences than languages dominated by simple syllables, due to the frequent consonant clusters in the former languages. By similar reasoning, the durational variability of consonantal intervals in sentences (defined as sequences of consonants between vowels, irrespective of syllable or word boundaries) should be greater for languages with more diverse syllable structures. This latter point is schematically illustrated in 3.11, in which boundaries between syllables are marked with a dot and consonantal intervals are underlined:

(3.11a) $\underline{C}V.\underline{CCC}V\underline{C.C}V.\underline{C}V.\underline{C}V\underline{CC}$ "stress-timed" language

(3.11b) $\underline{C}V.\underline{C}V.\underline{C}V\underline{C.C}V.\underline{C}V.\underline{C}V\underline{C.C}V$ "syllable-timed" language

Note how the greater diversity of syllable types in 3.11a leads to greater variation in the number of consonants between vowels (likely to translate into greater durational variability of consonantal intervals) as well as a lower vowel to consonant ratio (likely to translate into a lower fraction of utterance duration spent on vowels).

These ideas were borne out by empirical measurements. Figure 3.7 (from Ramus et al., 1999) shows a graph with percent of duration occupied by vowels

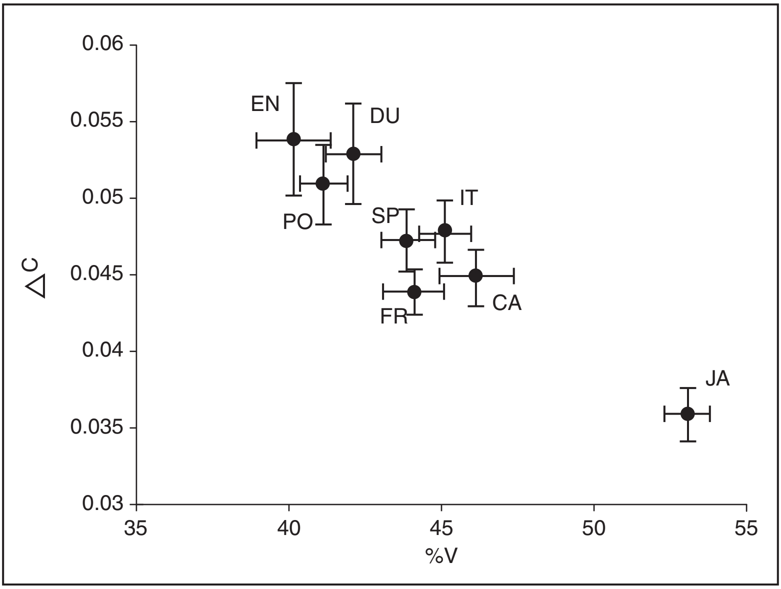

Figure 3.7 Percentage of sentence duration occupied by vowels versus the standard deviation of consonantal intervals within sentences for 8 languages. (CA = Catalan, DU = Dutch, EN = English, FR = French, IT = Italian, JA = Japanese, PO = Polish, SP = Spanish.) Error bars show +/- 1 standard error. From Ramus, Nespor, & Mehler, 1999.

(%V) versus consonantal interval variability (ΔC) within sentences in eight languages. (The data for each language came from 20 sentences read by four speakers, i.e., five sentences per speaker.)

What is interesting about this graph is that languages traditionally classified as stress-timed (English and Dutch) have low %V and high ΔC values, and occupy a different region of the graph than languages traditionally classified as syllable timed (French, Italian, and Spanish). Furthermore, Japanese, which linguists place in a different rhythmic category (mora-timed) is isolated from the other languages. (The location of Polish and Catalan in this graph is discussed in the next section, on perception.) This demonstrated an empirical correlate of traditional linguistic rhythmic classes, and has inspired other researchers to examine more languages in this framework. One interesting study is that of Frota and Vigário (2001), who examined the rhythm of Brazilian Portuguese versus European Portuguese (henceforth BP and EP). Linguists had often claimed that these two varieties were rhythmically different, with EP being stress-timed, and BP being syllable-timed or having mixed rhythmic characteristics. This makes Portuguese a fascinating topic for speech rhythm research, because one can study sentences with exactly the same words but spoken with different rhythms. (British English and Singapore English provide another such opportunity, because the former is stress-timed and the latter has been described as syllable-timed; see Low et al., 2000.) Frota and Vigário compared sentences spoken by European and Brazilian speakers of Portuguese, and found that EP had significantly a higher ΔC and lower %V than BP, as predicted by Ramus et al.'s findings.[15]

One important question about this line of research concerns the perceptual relevance of ΔC and %V. Ramus et al. focused on these measures because of their interest in the role of rhythm in infant speech perception. There is evidence that newborns and young infants can discriminate languages that belong to different rhythmic classes (Mehler et al., 1988, 1996; see also the next section). Mehler and colleagues (1996) have argued that this ability helps bootstrap language acquisition: Once a given rhythmic class is detected, class-specific acquisition mechanisms can be triggered that direct attention to the units that are relevant for segmenting words from connected speech (e.g., stresses in the case of English, syllables in the case of French, as discussed in section 3.3.3, subsection "The Role of Rhythm in Segmenting Connected Speech"). For this theory to work, infants must have some basis for discrimi-

[15] Although Ramus et al. (1999) related differences in ΔC and %V to syllable structure, Frota and Vigário (2001) point out that in the BP/EP case, differences in these variables are driven by vowel reduction, because syllable structures are similar in the two varieties. See Frota and Vigário (2001) for details. Frota and Vigário also provide a very useful discussion of the ΔC parameter and the need to normalize this variable for overall sentence duration/speech rate (see also Ramus, 2002a).

nating rhythmic class. Thus Ramus et al. (1999) sought an acoustic correlate of rhythmic class that would require minimal knowledge about linguistic units. ΔC and %V are two such parameters, because one only need assume that the infant can distinguish between vowels and consonants (see Galves et al., 2002, for an acoustic correlate of ΔC that does not even require segmentation into vowels and consonants).

One may ask, however, if ΔC and %V are directly relevant to the perception of speech rhythm, or if they are simply correlated with another feature that is more relevant to rhythm perception. That is, one could argue that these measures are global statistics reflecting variability in syllable structure, and are not themselves the basis of rhythm perception in speech (cf. Barry et al., 2003). A more promising candidate for perceptual relevance may be variability in syllable duration, which is likely to be correlated with variability in syllable structure and with vowel reduction. Because the syllable is widely regarded as a fundamental unit in speech rhythm, and because both adults and infants are sensitive to syllable patterning (e.g., van Ooyen et al., 1997), it would be worth examining the corpus of sentences used by Ramus et al. for syllable duration variability to see if this parameter differentiates traditional rhythmic classes. This would also be a straightforward test of Dauer's ideas, as the phonological factors she outlines imply that syllable duration variability should be higher in sentences of stress-timed than of syllable-timed languages.

Surprisingly, there has been little empirical work comparing sentence-level variability in syllable duration among different languages. As noted in the previous section, this may reflect the difficulties of assigning syllable boundaries in connected speech. From a purely practical standpoint, it is easier to define phoneme boundaries, using criteria agreed upon by most phoneticians (e.g., Peterson & Lehiste, 1960). However, this should not stop research into syllabic duration patterns, because these patterns are likely to be perceptually relevant. To illustrate both the feasibility and the challenges of a syllable-based approach, examples 3.11c and d below show a sentence of English and French segmented at syllable boundaries (the segmentations were done by myself and Franck Ramus, respectively). Periods indicate syllable boundaries that we felt were clear, whereas square brackets indicate phonemes that seemed ambiguous in terms of their syllabic affiliation. In the latter case, one must decide where to place the syllable boundary. For example, if the phoneme sounds ambisyllabic then the boundary can be placed in the middle of the phoneme, or if it sounds like it has been resyllabified with the following vowel, the boundary can be placed before the phoneme.

(3.11c) The .last .con.cert .gi.ven .at .the .o[p]era .was .a .tre.men.dous .suc.cess

(3.11d) Il. fau.dra .beau.coup .plus .d'ar.gent .pour .me.ne[r] à .bien .ce .pro.jet

It is likely that different researchers will vary in how they make these judgment calls. Nevertheless, this is not an insurmountable problem for rhythm research.

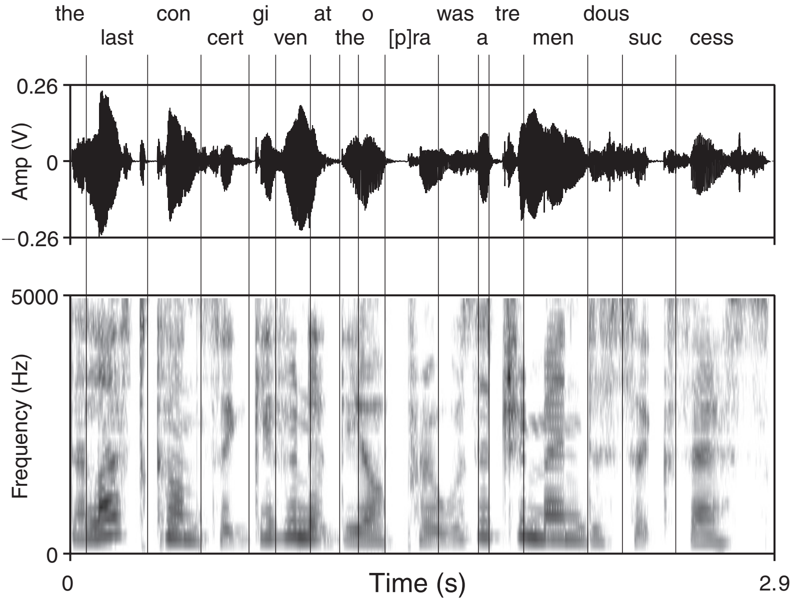

Figure 3.8a A sentence of British English segmented into syllables. (Note that "opera" is pronounced "opra" by this speaker.)

In fact, if different researchers define syllable boundaries in slightly different ways but nevertheless converge on the rhythmic differences they find between languages, this is strong evidence that the observed differences are robust.[16] Figure 3.8a and 3.8b show my markings of syllable boundaries in the waveform and spectrograms of these two sentences (the sentences can be heard in Sound Examples 3.5a and b; note that in sentence 3.5a, "opera" is pronounced "opra").

For these sentences, the variability of syllable durations as measured by the coefficient of variation (the standard deviation divided by the mean) is .53 for the English sentence and .42 for the French sentence. Making similar measurements on all the English and French sentences in the Ramus database yields the data in Figure 3.8c. As can be seen, on average English sentences have more variable syllable durations than do French sentences (the difference is statistically significant, $p < 0.01$, Mann-Whitney U test). It would be interesting to

[16] Individual researchers who are comparing syllable duration patterns across two languages can also handle the problem of syllable boundary identification by making all such decisions in a manner that is conservative with regard to the hypothesis at hand. For example, if comparing languages A and B with the hypothesis that syllables are more variable in duration in the sentences of language A, then any judgment calls about syllable boundaries should be made in such a way as to work against this hypothesis.

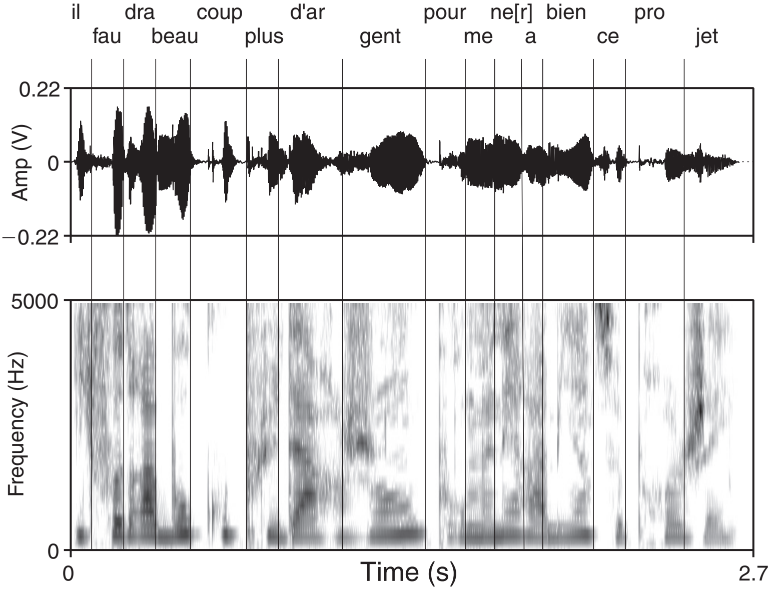

Figure 3.8b A sentence of French segmented into syllables.

have similar variability measurements for numerous languages that have been classified as stress- versus syllable-timed: Would these measurements divide the languages into their traditional rhythmic classes? (See Wagner & Dellwo, 2004, for a promising start.)

Turning now to the ideas of Dwight Bolinger, recall Bolinger's claim that syllables containing full and reduced vowels tend to alternate in English. This leads to an empirical prediction, namely that the durational contrast between adjacent vowel durations in English sentences should be greater than in languages of a different rhythmic class, such as French or Spanish. In fact, there is research supporting this prediction, though it was not inspired by Bolinger's work but by an interest in the role that vowel reduction plays in the rhythm of stress- versus syllable-timed languages. Low, Grabe, and Nolan (2000) set out to explore the idea that vowel reduction contributes to the impression of stress-timing via its impact on vowel duration variability in sentences. They tested this idea by examining vowel duration patterning in a stress-timed versus a syllable-timed variety of English (British vs. Singapore English). Crucially, they developed an index of variability that was sensitive to the patterning of duration. Their "normalized pairwise variability index" (nPVI) measures the degree of contrast between successive durations in an utterance. An intuition for the nPVI can be gained by examining Figure 3.9, which schematically depicts two sequences of events of varying duration (the length of each bar corresponds to the duration of the event).

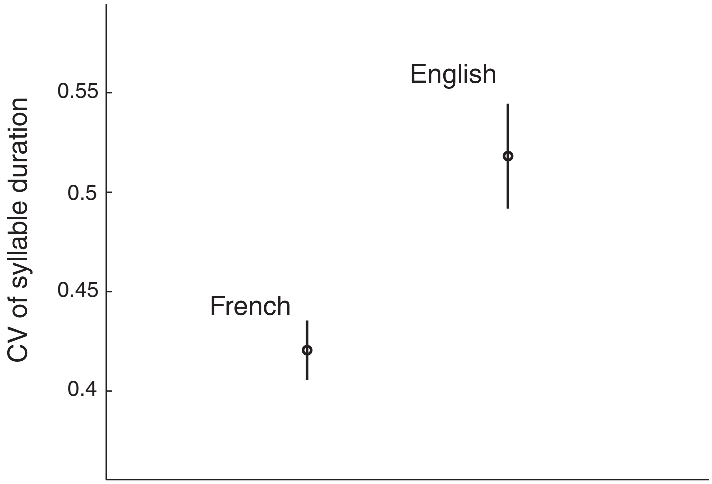

Figure 3.8c The coefficient of variation (CV) of syllable duration in 20 English and 20 French sentences. Error bars show +/- 1 standard error.

In sequence A, neighboring events (e.g., events 1 and 2, 2 and 3) tend to have a large contrast in duration, and hence the sequence would have a large nPVI. Now consider sequence B, which has the same set of durations as sequence A, arranged in a different temporal order. Now neighboring events tend to have low contrast in duration, giving the sequence a low nPVI value. Hence the two sequences have a sharp difference in durational contrastiveness, even though they have exactly the same overall amount of durational variability, for

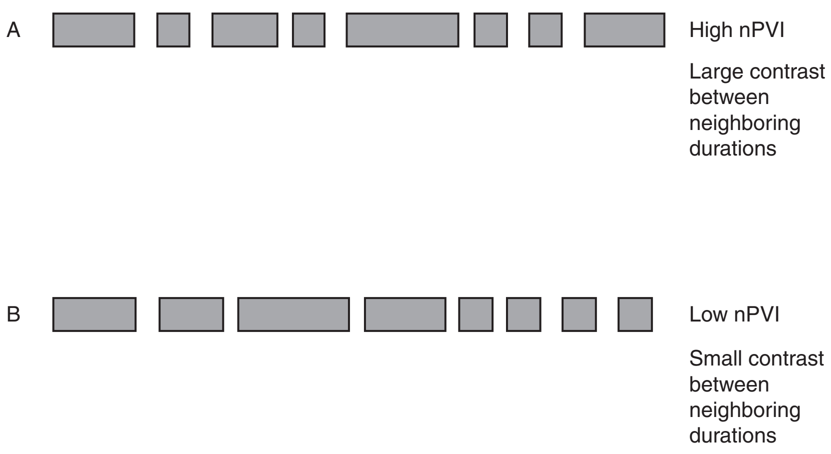
A
High nPVI
Large contrast between neighboring durations
B
Low nPVI
Small contrast between neighboring durations

Figure 3.9 Schematic of sequences of events with varying duration, to illustrate the nPVI (longer bars = longer durations). See text for details.

example, as measured by the standard deviation of durations. (See this chapter's appendix 1 for the nPVI equation.)

Because the nPVI is fundamentally a measure of contrast, the use of the term "variability" in its name is somewhat unfortunate, as variability and contrast are not necessarily correlated, as shown in Figure 3.9. In fact, it is quite possible to have two sequences A and B in which the variability of durations in A is greater than B, but the nPVI of durations is greater in B than of A (an example is given in section 3.5.1). Thus a better term for this measure might have been the "normalized pairwise contrastiveness index."

I have delved into the details of the nPVI because it has proven quite fruitful in the study of speech rhythm and in the comparative study of linguistic and musical rhythm (discussed in 3.5.1). Grabe and Low (2002) have used the nPVI to examine the patterning of vowel durations in sentences of a number of languages, and have shown that several languages traditionally classified as stress-timed (such as German, Dutch, British English, and Thai) have a larger vocalic nPVI than a number of other languages traditionally classified as syllable timed (such as French, Italian, and Spanish). This supports Bolinger's idea that durational alternation of vowels is important to stress-timed rhythm.[17] Inspired by this work, Ramus (2002b) measured the vowel nPVI for all eight languages in his database and found the results shown in Figure 3.10.

Figure 3.10 plots the nPVI for vocalic intervals against the rPVI for intervocalic intervals (i.e., consonantal intervals). (The rPVI, or "raw pairwise variability index," is computed in the same way as the nPVI but without the normalization term in the denominator; cf. this chapter's appendix 1. Grabe and Low [2002] argue that normalization is not desirable for consonantal intervals because it would normalize for cross-language differences in syllable structure.) Focusing on the nPVI dimension, the stress-timed languages (English and Dutch) are separated from the syllable-timed languages (Spanish, Italian, and French), which provides additional support for Bolinger's ideas.[18] Furthermore,

[17] The correlation between nPVI and rhythm class is not perfect, however. Grabe and Low (2002) found a high vowel nPVI value for Tamil, a language which has been classified as syllable-timed (cf. Keane, 2006). It should be noted that Grabe and Low's (2002) results should be considered preliminary because only one speaker per language was studied. Subsequent work has applied the nPVI to cross-linguistic corpora with fewer languages but more speakers per language (e.g., Ramus, 2002b; Lee & Todd, 2004; Dellwo, 2004).

[18] It should be noted that both Grabe and Low (2002) and Ramus (2002b) measured the nPVI of vocalic intervals, defined as vowels and sequences of consecutive vowels irrespective of syllable and word boundaries, whereas Bolinger's arguments are focused on individual vowels. This is not a serious problem because most vocalic intervals are individual vowels due to the strong tendency for vowels to be separated by consonants in speech. For example, in the database used by Ramus (eight languages, 160 sentences),

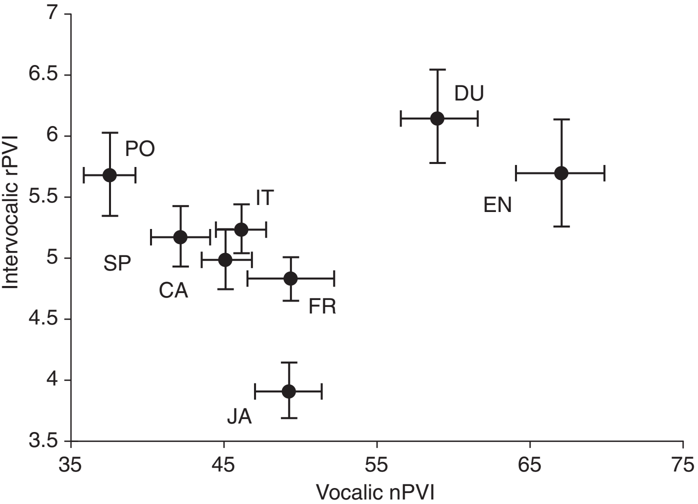

Figure 3.10 Vocalic nPVI versus Consonantal (intervocalic) rPVI for sentences in eight languages. (CA = Catalan, DU = Dutch, EN = English, FR = French, IT = Italian, JA = Japanese, PO = Polish, SP = Spanish.) Error bars show +/- 1 standard error. From Ramus, 2002a.

Polish is now far from the stress-timed languages, which is interesting because there is perceptual evidence that Polish is rhythmically different from these languages due to its lack of vowel reduction (see section 3.3.1, subsection "Perception and Typology"). Japanese is similar to French in terms of nPVI, however, suggesting that nPVI alone is not enough to sort languages into traditional rhythmic classes. Adding a second dimension of rPVI for consonantal intervals, however, does segregate out Japanese, which has very low durational contrast between successive consonantal intervals. This suggests that at least two phonetic dimensions may be needed to capture differences between rhythmic classes (see also Ramus et al., 1999).

One interesting linguistic application of the nPVI has been to the ontogeny of speech rhythm. It has been claimed that the rhythm of English-speaking children is syllable-timed in contrast to the stress-timed rhythm of adult speech (Allen & Hawkins, 1978). Grabe et al. (1999) conducted an nPVI study that

there are 2,725 vowels, out of which 2,475 (91%) are singletons, in other words, a single vowel flanked by a consonant on either side (or if the vowel is the first or last phoneme of the sentence, a single vowel flanked by a following or preceding consonant, respectively). Thus it is likely that nPVI measurements based on individual vowels would produce qualitatively similar results as those based on vocalic intervals (cf. Patel et al., 2006).

supported this claim. They measured the nPVI of vowels in the speech of English versus French speaking 4-year-olds and their mothers. They found that English children had significantly lower nPVI values than their mothers, whereas French children resembled their mothers in having a low nPVIs. That is, both English and French children spoke with a syllable-timed rhythm (though the nPVI of the English children was already larger than that of their French counterparts). It would be interesting to track nPVI as a function of age in English and French children, to study the developmental time course of speech rhythm in the two languages.

All nPVI studies to date have focused on a single rhythmic layer in language: the temporal patterning of vowels or consonants. In the spirit of Bolinger's idea that rhythm may involve multiple levels of temporal organization, it would be worth using the nPVI to explore the relationship of durational patterns at various rhythmically relevant levels in speech (cf. Asu & Nolan, 2006). For example, within English sentences, one could compute the nPVI of interstress intervals (ISIs) relative to the nPVI of syllable durations by measuring both of these quantities in each sentence and then taking the ratio of the former to the latter. It may be that the subjective impression of isochrony in English arises in part from a lower durational contrast between ISIs than between syllables, which would make this ratio significantly less than 1. I return to this idea in section 3.3.4.

This section has reviewed a few different acoustic correlates of speech rhythm. Due to the success of this work, it seems certain that more such correlates will be proposed and explored in the future (e.g., Gut, 2005). Ultimately, the usefulness of such measures will depend on whether they group together languages that are perceived as rhythmically similar and divide languages perceived as rhythmically different. Perceptual studies are thus fundamental to research on rhythmic typology, and it is to such studies that we turn next.

Perception and Typology

All typological theories of language rhythm are ultimately rooted in perception. In the past, linguists have defined rhythm categories (such as stress vs. syllable timing) based on their auditory impressions of languages, and then researchers have sought to identify phonological and acoustic correlates of these classes. The recent success in finding durational correlates of traditional rhythm classes is a testament to the intuition of linguists in their aural rhythmic judgments. However, it is also apparent that the old categorization system has its shortcomings. For example, some languages straddle different categories (e.g., Polish and Catalan, see above), and many languages do not fit neatly into any of the existing categories (Grabe & Low, 2002). Thus the old system is cracking at the seams, and a new science of rhythm classification is called for. Such a science must have as its foundation a body of perceptual data that provides a measure of the rhythmic similarities and differences between languages. These data will allow researchers to construct a perceptual map of language rhythms and determine to what extent the rhythms of human languages fall into distinct clusters (vs. forming a continuum). It will also help suggest new avenues for empirical research into the acoustic foundations of speech rhythm.

Fortunately, perceptual work on the rhythmic differences between languages has already begun. An innovative study by Ramus and Mehler (1999) devised a method for studying the perception of speech rhythm posited on the idea that if a listener can tell two languages apart when the only cues are rhythmic, then the languages belong to distinct rhythmic classes. Speech resynthesis techniques were used to selectively remove various phonetic differences between languages and focus attention on rhythm. Sound Examples 3.6 and 3.7 illustrate Ramus and Mehler's technique on a sentence of English and Japanese. Each sentence is presented in four versions, which convert the original sentence to an increasingly abstract temporal pattern of vowels and consonants. In the first transformation, each phoneme is replaced by a particular member of its class: all fricatives replaced by /s/, vowels by /a/, liquids (l & r) by /l/, plosives by /t/, nasals by /n/, and glides by /ai/ (a condition they called "saltanaj," pronounced "sal-tan-ai"). The original intonation of each sentence is preserved. In the second transformation, all consonants are replaced by /s/ and all vowels by /a/ (a condition they call "sasasa"). In the final transformation, the voice pitch is flattened to a monotone, leaving the temporal pattern of vowels and consonants as the only difference between the languages (a condition the authors refer to as "flat sasasa"). Ramus and Mehler found that French adults could discriminate between English and Japanese in all three conditions, supporting the hypothesis that the rhythms of English and Japanese are indeed perceptually distinct.

Focusing on the flat sasasa transformation, Ramus et al. (2003) also tested French adults' ability to discriminate the rhythms of English, Polish, Spanish, and Catalan. The results indicated that Polish could be discriminated from the other languages, whereas Catalan could not be discriminated from Spanish, though it was distinct from English and Polish. (Recall that on phonological grounds, Polish and Catalan seemed intermediate between stress-timed and syllable-timed languages; cf. section 3.3.1, subsection "Phonology and Typology.") These perceptual data suggest that Polish does belong in a separate rhythmic category than English, whereas Catalan belongs in the same category as Spanish. This finding has implications for maps of the acoustic correlates of speech rhythm, such as Figure 3.7. In that map, Polish clustered with stress-timed languages, indicating that a different acoustic dimension is needed to separate perceived rhythmic classes. Indeed, Ramus et al. (1999, 2003) have noted that Polish can be separated out from all other languages in their original study on a dimension that measures the variability of vowel duration in a sentence, ΔV, because Polish has a very low vowel duration variability compared to all other languages in their sample.[19] Thus perceptual work on speech has already suggested that if one wishes to preserve the notion of rhythm classes in language, at least four classes are needed: stress-timed, syllable-timed, mora-timed (represented by Japanese), and one other yet-to-be-named category represented by Polish.[20]

Another line of perceptual research concerned with rhythmic typology has focused on newborns and infants. This choice of subjects may seem surprising, but these studies are motivated by the idea that very young humans are sensitive to speech rhythm and use it to guide learning of fine-grained sound patterns of language (Mehler et al., 1996). Nazzi et al. (1998) studied newborn rhythm perception using low-pass filtered speech. This removes most of the phonetic information but preserves syllable, stress, and pitch patterns. They showed that French newborns are able to discriminate English from Japanese, but not English from Dutch, suggesting that the latter are members of the same rhythmic class. They also showed that the newborns could discriminate English and Dutch from Spanish and Italian, but not English and Spanish from Dutch and Italian, suggesting that the former pairings more accurately capture perceptual rhythmic classes (cf. Nazzi et al., 2000, for converging findings with 5-month-old infants). These findings support the authors' hypothesis that babies can discriminate languages *only* when they belong to different rhythmic classes, a notion that they dub the "rhythm hypothesis" for language acquisition. If this is true, then the ears of infants may be particularly important instruments in mapping human speech rhythms in future research.[21]

The studies of Ramus, Nazzi, and colleagues raise a number of points for future research on the perception of speech rhythm. First, it is important to design stimuli and tasks that focus attention on those aspects of speech rhythm that play a role in normal speech perception. For example, a danger of the flat sasasa condition it that when a language with a highly variable syllable structure (such as English) is compared to a language dominated by simple syllables (such as Japanese), a salient perceptual difference between the resulting flat

[19] Polish also has the lowest vocalic nPVI of all languages in the Ramus et al. (1999) database (Ramus, 2002a; cf. Figure 3.10).

[20] It would be desirable for future research on rhythmic classes to suggest new names for rhythmic classes, as the current names (stress-timed, syllable-timed, and mora-timed) are implicitly bound up with the (failed) notion of isochrony.

[21] One possible confound in the elegant study of Nazzi et al. (1998) is the presence of intonation, which may have played a role in the infants' discrimination. Indeed, Ramus et al. (2000) found that French newborns could distinguish Dutch from Japanese using resynthesized saltanaj speech, but that their discrimination ability was much weaker when the original F0 contours of the sentences were replaced by the same artificial contours (Ramus, 2002b). He also notes that intonation can be removed entirely using flat sasasa resynthesis, but that the resulting sound patterns are problematic for use with newborns and infants, who may find them boring or distressing.

sasasa stimuli is the more frequent occurrence of long-duration /s/ sounds in the former stimulus (which result from transforming consonant clusters into single, long /s/ sounds). Thus discrimination could simply be based on listening for frequent long /s/ sounds rather than on attending to temporal structure.

Second, research on the perceptual taxonomy of language rhythms should not only be based on discrimination tasks, but should also incorporate similarity judgments. Musical studies of rhythmic similarity provide a good model for such research (Gabrielsson, 1973, 1993). In this research, rhythms are presented in a pairwise fashion and listeners rate their perceived similarity using a numerical scale. The resulting ratings are studied using multidimensional scaling to uncover perceptual dimensions used by listeners in classifying rhythms. This paradigm could easily be adapted to study speech rhythm, using low-pass filtered speech with minimal pitch variation as stimuli. Such studies should be sensitive to the idea that the important perceptual dimensions for rhythm may be relational, for example, a high contrastiveness between successive syllable durations while simultaneously having a lower durational contrastiveness between interstress intervals (cf. the end of section 3.3.1, subsection "Duration and Typology").

Finally, a fundamental issue for all future studies of rhythmic typology is the extent to which perceived rhythmic similarities and differences between languages depend on the native language of the listener. The theory of stress, syllable, and mora timing was proposed by native English speakers, and it is an open question whether speakers of other languages perceive rhythmic cues in the same way that English speakers do. For example, it has recently been demonstrated that French listeners have some difficulty distinguishing nonsense words that differ only in the location of stress, whereas Spanish listeners have no such difficulty (Dupoux et al., 2001). This likely reflects the fact that Spanish has contrastive stress: Two words can have the same phonemes but a different stress pattern, and this can change the meaning of the word entirely (e.g., *sábana* vs. *sabána,* which mean "sheet" and "savannah" respectively; cf. Soto-Faraco et al., 2001). French does not have this property, and Dupoux et al. suggest that this difference is responsible for the "stress deafness" they found in their French listeners. Results such as this raise a fundamental question: Is there a single map of perceived rhythmic similarities and differences among languages, or does the geography of the map differ according to the native language of the listener? Only empirical work can resolve this issue, but it seems a real possibility that native language influences the perception of rhythmic relations between languages.

3.3.2 *Principles Governing the Rhythmic Shape of Words and Utterances*

For those interested in comparing rhythm in language and music, it is important to be familiar with a branch of theoretical linguistics known as "metrical phonology." Metrical phonology deals with speech rhythm, but it does so in a manner quite different from the approaches described so far. First and foremost, rhythmic prominence is treated as hierarchical. That is, prominence is incrementally assigned *at each level* of the prosodic hierarchy according to systematic principles. For example, in a given theory it may be the case that all syllables begin with a basic amount of prominence, then the lexically stressed syllable of each word (or clitic group) is assigned an additional degree of prominence, then a phrase-level prominence is added to a particular word of a phrase (e.g., the "nuclear stress rule" in English), and so on. In this view, prominence is not simply a binary phonetic feature called "stress" that syllables either have or do not. Rather, prominence is an acoustic projection of the hierarchical prosodic structure of an utterance, and as such, has several degrees that serve to indicate a syllable's position in an utterance's rhythmic hierarchy (Halle & Vergnaud, 1987; Halle & Idsardi, 1996; Shattuck-Hufnagel & Turk, 1996).[22]

One of the clearest expositions of metrical phonology is in Selkirk's 1984 book, *Phonology and Syntax: The Relation Between Sound and Structure*. One goal of this book is to show how one can go from a string of words to a representation of the syllabic prominence pattern of the spoken utterance in a rule-governed fashion. The relative prominence of syllables is represented using a "metrical grid" that treats each syllable as a point in abstract time (Figure 3.11), meaning that prominence patterns are considered without regard to their exact timing.

Two aspects of the linguistic metrical grid, introduced by Liberman (1975), embody "the claim that the rhythmic organization of speech is quite analogous to that of music" (Selkirk, 1984:9). First, as noted above, prominence is treated hierarchically, analogously to hierarchical theories of musical meter (Cooper & Meyer, 1960; Lerdahl & Jackendoff, 1983). Above the basic level of the syllable are several other levels. The second level marks stressed syllables, and is the level of the basic "beat," in analogy to the tactus in music (Selkirk, 1984:10, 40). The third level marks the primary lexical stress of each word, and the fourth level marks the main accent of each phrase. This "text-to-grid" assignment of

[22] An early conceptual link between hierarchical theories of linguistic rhythm and theories of musical structure was made by Jackendoff (1989), who noted a structural equivalence between one type of prosodic tree structure used to depict hierarchical prominence relations in language and a type of tree used by Lerdahl and Jackendoff (1983) to indicate the relative structural importance of events in a span of musical notes. Jackendoff speculated that the coincidence of these two formalisms might reflect the fact that language and music use different specializations of general-purpose mental principles for assigning structure to temporal patterns, in other words, principles that parse sound sequences into recursive hierarchies of binary oppositions of structural importance. As noted by Jackendoff, however, prosodic trees have been largely abandoned in theories of speech rhythm.

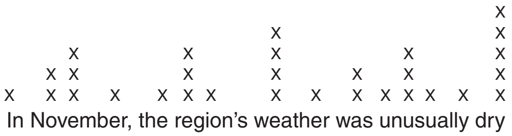

Figure 3.11 A metrical grid for a sentence of English. From Selkirk, 1984.

beats provides the input on which rhythmic principles operate. These principles, which represent the second link to musical meter, amount to a tendency to alternate between stronger and weaker elements at *each level* of the hierarchy. The principles are enforced by rules that can add, delete, and move beats to make the pattern at each level more congruous with an alternating pattern. For example, a rule of "beat addition" might add a beat at the second level to avoid a long series of unstressed syllables. At the third level, a rule of "beat movement" might shift the primary stress/accent of word to avoid the adjacency of primary lexical stress/accent (as when "thirtéen" becomes "thírteen mén"; Liberman and Prince, 1977, Shattuck-Hufnagel et al.,1994; Grabe & Warren, 1995). The ideal goal is that strong beats at any given level are separated by no more than two weak beats at that level (the "principle of rhythmic alternation"). Thus metrical phonology derives the prominence pattern of a sentence using ideas directly inspired by theories of musical meter.

The notion that speech has multiple rhythmically relevant levels is an interesting abstract similarity between rhythm in language and music, because a fundamental property of musical meter is the existence of perceptually salient temporal pattering on multiple timescales (cf. section 3.2.2). Furthermore, just as musical meter involves at least one psychologically accessible rhythmic level below the beat and one or two levels above it, metrical phonology proposes rhythmic levels below and above the "beat" of stressed syllables. That is, both theories concern the patterning of time intervals at several timescales.

Although the picture painted by metrical phonology is an elegant one, it should be noted that its claims are by no means universally accepted by speech scientists (Cooper & Eady, 1986), and that the patterns of prominence it proposes are typically constructed from the intuitions of linguists rather than from acoustic and perceptual data collected in laboratory settings. However, there is hope that the field can be put on an empirical footing. For example, there is phonetic evidence for four degrees of prominence in speech (at least in stress-timed languages), corresponding to reduced vowels, full vowels, stressed syllables, and accented syllables (Terken & Hermes, 2000). For the current purposes, metrical phonology is interesting because it draws attention to a number of issues in which comparisons of linguistic and musical rhythm are instructive. One of these issues (multiple layering in rhythmic structure) leads to ideas for empirical comparative studies of rhythm in speech and music, and is discussed further in section 3.3.4. Two other issues are discussed below.

Differences Between Linguistic and Musical Metrical Grids

Although the hierarchies posited by metrical phonology were inspired by Western music, some very important differences between the meters of music and language are readily apparent. Most notably, temporal periodicity in musical meter is much stricter than anything found in speech, and this difference has dramatic cognitive consequences. The regular periodicities of music allow meter to serve as a mental framework for sound perception, such that an event can be perceived as metrically prominent even if is physically quite weak, as in syncopated rhythms. By comparison, the prominences of language are not regular enough to allow for anything as abstract as syncopation. As a result, linguistic metrical grids are not abstract periodic mental patterns (like musical metrical grids) but are simply maps of heard prominences, full of temporal irregularities. For example, Dauer (1983) reports that the average interstress interval in speech was around 450 ms, with a standard deviation of approximately 150 ms. Dividing the standard deviation by the mean yields a coefficient of variation of about 33%. This variability is markedly different from music, in which perceived beats occur in a much more evenly spaced fashion. For example, when tapping to music, adults show a coefficient of variation of about 5%. Thus "metrical grids" in language should perhaps be called "prominence grids," to avoid the implication of an abstract mental periodicity.

Questioning the Principle of Rhythmic Alternation in Speech

Setting aside questions of temporal periodicity, one can ask if speech and music share a more abstract similarity in terms of a tendency to arrange prominences in alternating patterns of strong and weak elements. If so, this might suggest a basic cognitive relationship between rhythm in language and music. Evidence in favor of a principle of alternation in language comes from studies showing that English speakers adjust prominence patterns to make them more regular. For example, Kelly and Bock (1988) had speakers pronounce nonsense words embedded in sentences, such as:

(3.12a) The full teplez decreased.

(3.12b) Throw the teplez badly.

The focus of interest was whether speakers stressed the first or second syllable of the nonsense word. Overall, speakers tended to stress the first syllable, in accordance with a general trend in English for disyllabic nouns to have initial stress. However, this tendency was significantly weaker when the nonsense word was preceded by a stressed syllable (as in sentence 3.12a), as if speakers wanted to spread out the occurrence of stresses. Further evidence for alternation of stress patterns in speech production comes from Cutler (1980), who examined sentences in which speakers inadvertently omitted a syllable, such as:

```
        x       x           x       x
(3.13a) Next we have this bicential rug
```

versus the intended sentence:

```
        x       x              x       x
(3.13b) Next we have this bicentennial rug
```

Much more often than chance, the errors served to shorten a long run of unstressed syllables, thus tending to promote the alternation of stressed and unstressed syllables.

Although these findings seem to support a positive principle of alternation, it is also possible that they reflect the action of a negative principle that seeks to break up clusters of prominent syllables or clusters of nonprominent syllables (i.e., "stress clashes" and "stress lapses," Nespor & Vogel, 1989). Some support for the latter view comes from the observation that the regularizing tendencies reported by Kelly and Bock (1988) and Cutler (1980) are actually rather weak. In the former study, subjects placed initial stress on the target nonsense word in the majority of cases, whether or not the immediately preceding syllable was stressed. The presence of a prior stressed syllable simply lowered the proportion of initial stress from 80% to 70%, suggesting only a mild tendency to maintain an alternating stress pattern in speech. Similarly, Cutler's study is based on collecting relatively rare syllable omission errors, meaning that speakers usually manage quite well with irregular prominence patterns.

Thus at the current time it is impossible to rule out the hypothesis that the tendency to alternate stronger and weaker syllables in speech is the result of nonrhythmic forces that seek to keep prominences at a comfortable distance from each other. In fact, research on Greek suggests that the *alternation* of prominence may not even be a universal pattern for human languages, because Greek tolerates long sequences of unstressed syllables (Arvaniti, 1994). Thus it may be that the only universal principle regarding prominence patterns in language is that prominences that are too close together are subject to linguistic mechanisms for clash avoidance (see Arvaniti, 1994, for evidence from Greek; and Nespor, 1990, for references to studies of clash avoidance mechanisms in numerous languages). The reason such mechanisms exist may ultimately be rooted in the mechanics of articulation. Stressed syllables tend to be made with larger jaw movements than unstressed syllables (de Jong, 1995), and it may be biomechanically advantageous to avoid crowding larger jaw movements together when speaking at the fast rates typical of normal conversation.

3.3.3 *The Perception of Speech Rhythm*

The role of rhythm in speech perception has been the focus of at least four different lines of research. Two of these have obvious conceptual connections to musical rhythm: the study of perceived isochrony in speech and the investigation of the role of rhythmic predictability in perception. The third line has investigated the role of speech rhythm in the perceptual segmentation of words from connected speech. Although not obvious at first, this research is in fact quite pertinent to comparative perceptual studies of rhythm in language and music. The final (and most recent) line of work concerns the role that rhythm plays in perception of nonnative accents. Although no conceptual link has yet been made between this work and music research, it is briefly described because it is a promising new area for empirical work on speech rhythm.

The Perception of Isochrony in Speech

As noted in section 3.3.1 (subsection "Periodicity and Typology"), the idea that linguistic rhythm involves regular temporal intervals (e.g., between stresses or syllables) has received no empirical support from measurements of speech. However, all such measurements have been based on data from speech production, in other words, on waveforms or spectrograms of spoken utterances. In an influential paper, Lehiste (1977) made the interesting suggestion that periodicity may be stronger in perception than in production. That is, the ear may ignore or compensate for surface irregularities in judging periodicity in speech. She based this idea on empirical work in which she examined the ability of listeners to identify the longest or shortest interstress interval (ISI) in short sentences of four ISIs, and to do the same task on nonspeech analogs of the sentences in which the stresses were replaced by clicks and the speech by noise. She found that listeners performed better in the nonlinguistic condition, and suggested that if listeners had difficulty judging ISI duration differences in speech, this would lead to a sense that ISIs were similar in duration, in other words, an impression of isochrony. (Of course, it may not be speech per se that makes small duration differences difficult to detect; it could be that the presence of semantic meaning in language preoccupies the listener's attention so that fine duration judgments are difficult. Thus an important control condition for future studies of this sort is to use a language with which the listener is unfamiliar.)

Lehiste went on to study the just noticeable difference (JND) in duration for sequences of four noise-filled intervals, reasoning that this would establish a conservative estimate of the JNDs for ISIs in speech. She used three basic reference durations in her noise sequences (300, 400, and 500 ms). In each sequence, three of the four intervals had the same duration, and the fourth was increased or decreased in nine 10-ms steps. She found that reliable judgments identifying one interval as longer or shorter than the others required changes of between

30 and 100 ms. She argued that JNDs for ISIs in speech are no better than this and are likely to be worse, and thus that physical measurements of isochrony need to take this "perceptual tolerance" into account (cf. Kristofferson, 1980).

Lehiste's work is interesting because it raises the possibility that listeners hear more isochrony than is really there in speech. Some evidence offered in favor of this argument comes from Donovan and Darwin (1979), who had individuals listen to English sentences and then imitate the timing of each sentence's stress pattern by tapping. The subjects also performed this task with sequences of noises whose timing mimicked the stress pattern of sentences. The critical finding was that when imitating speech, subjects tapped with less temporal variability than the actual timing of stressed syllables, whereas when imitating noise they did not show this pattern.

Although these findings are intriguing, further work has suggested that this paradigm may be flawed. Scott et al. (1985) replicated the findings of Donovan and Darwin for English, but also found that subjects showed regularization of tapping to French (which is not considered to have periodic stress), as well as to garbled speech. This suggests that the observed regularization may be a side consequence of the greater difficulty of remembering acoustically complex stimuli versus noise patterns.

Thus Lehiste's ideas merit further investigation, because they raise the important issue of how perceived timing patterns relate to the physical intervals measured in speech. Nevertheless, there is nothing in Lehiste's work that supports the idea that speech is perceived as isochronous under ordinary circumstances. As noted in section 3.3.2 (subsection "Differences Between Linguistic and Musical Metrical Grids"), the variability of ISI durations in speech is on the order of 33%. For a 500-ms average ISI, this is 150 ms, which is above the threshold for subjective isochrony suggested by Lehiste.

Lehiste's ideas do point in one interesting direction in terms of comparative studies of speech and music, namely a direct comparison of the threshold for detecting temporal irregularities in perceptually isochronous sequences (e.g., a repeating syllable "ta ta ta ta") versus a repeating musical sound of equivalent acoustic complexity.[23] If speech can tolerate more durational variability than music and still sound isochronous, this raises interesting questions about different mechanisms for time perception in the two domains.

One could also examine the threshold for detecting tempo change in isochronous sequences in speech and music. Current data suggest that for nonmusician listeners, the threshold for tempo change detection in sequences of musical sounds is 5%–8% (Drake & Botte, 1993; cf. Rivenez et al., 2003). Would the threshold be higher if speech sounds were used?

[23] Such a study will have to be careful to try and match acoustic properties of the onset of the spoken and musical sound. For example, if a musical sound with a sharp attack is used, such as a piano tone, then a speech sound with a plosive onset (such as /ta/) should be used rather than one with a gradual onset (such as /la/).

The Role of Rhythmic Predictability in Speech Perception

A number of researchers have argued that the ability to predict the location of stressed syllables in English is perceptually beneficial (e.g., Martin, 1972; Shields et al., 1974; Cutler & Foss, 1977). The reasoning behind this idea is based on certain assumptions: Stressed syllables carry important semantic information, and a listener's attention is limited, so that it is useful to expend attentional resources on those points in time where stresses occur. Thus the ability to anticipate stress location can help guide attention in an efficient manner. This idea suggests a point of contact between rhythm perception in speech and music, because there are theories linking rhythm and attention in music psychology (Jones, 1976; Large & Jones, 1999; Barnes & Jones, 2000). In order to determine if this is really a meaningful parallel, however, two questions must be answered. First, is there evidence that rhythmic predictability plays an important role in speech perception? Second, are the mechanisms for rhythmic prediction similar in speech and music?

The best evidence for a role of rhythmic predictability in speech perception comes from studies using phoneme-monitoring tasks. In these experiments, listeners are told to listen to one sentence at a time and press a button when they hear a target phoneme (such as /d/). Cutler and Darwin (1981) conducted a study in which sentences were recorded that had high, low, or neutral emphasis on a given target word. For example, sentences 3.14a and b below were used to record high versus low emphasis on the word "dirt" (in the sentences below, the word bearing the main emphasis of the sentence is italicized):

(3.14a) She managed to remove the *dirt* from the rug, but not the grass stains.

(3.14b) She managed to remove the dirt from the *rug*, but not from their clothes.

Cutler and Darwin then spliced the neutral version of the target word into high and low emphasis sentences, so that the target phoneme /d/ (and the rest of the word that began with this phoneme) were acoustically identical in the two cases. A faster reaction time to the target phoneme in high-emphasis sentences would indicate that the *prediction* of stress was influencing speech processing. This is precisely what was found: Listeners were reliably faster in detecting the target phoneme in high-stress sentences. Of particular interest is that this difference persisted even when fundamental frequency variation was removed from the two sentence types, suggesting that patterns of duration and amplitude were sufficient to predict the upcoming stress.

Cutler and Darwin's study focused on target words that either did or did not bear the main contrastive stress of the entire sentence. That is, they were not studying the perception of just any stressed syllable, but of a particularly salient stressed syllable in a sentence. Pitt and Samuel (1990) conducted a study in which the context manipulation was not so extreme: They used sentences that predicted stress or nonstress at a target point due to rhythmic and syntactic factors, for example, the first syllable of the word “permit” in:

(3.15a) The guard asked the visitor if she had a permit to enter the building.

(3.15b) The waiter decided he could not permit anyone else in the restaurant.

In sentence 3.15a, the context leads one to expect stressed syllable at the target location, both because the syntax of the sentence predicts a noun (a word category that tends to start with a stressed syllable in English) and for the rhythmic reason that the prior stress in the sentence is quite far away (the first syllable of “visitor”). In sentence 3.15b, the context leads one *not* to predict a stressed syllable, both because the syntax predicts a verb (a word category that tends to start with a weak syllable in English), and because the prior stress is quite nearby (on “could” or “not”).

Like Cutler and Darwin, Pitt and Samuel used a splicing technique to ensure that the physical target word was the same in the two contexts, and asked listeners to respond when they heard a target phoneme (e.g., /p/ in the above example). Unlike Cutler and Darwin, however, they found no significant difference in reaction time to the target phoneme as function of the preceding context. Thus it appears that although rhythm may help listeners predict sentence level emphasis, it does not play a strong role in predicting lexical stress, even when reinforced by syntax. This casts some doubt on the idea that rhythm plays an important role in guiding attention to the majority of stressed syllables in spoken sentences. Clearly, more work is needed to determine to what extent stress is predictable under normal circumstances.

Even if a significant role for rhythmic prediction in speech is demonstrated, however, it is quite possible that the mechanisms that underlie rhythmic prediction in speech and music are quite different. In music, rhythmic predictability reflects the periodic structure of temporal intervals. In speech, the basis for rhythmic predictability (e.g., predicting when a stress will occur) is unlikely to involve periodic time intervals, because there is no evidence that such intervals exist in normal speech. A first step in studying the basis of rhythmic prediction in speech would be to study the stimuli used by Cutler and Darwin (1981), especially those stimuli in which fundamental frequency variation was removed. What temporal and/or amplitude patterns helped guide listeners’ expectations in that study?

Thus at the current time, the hypothesis that rhythmic predictability in speech confers an advantage by guiding attention to semantically important parts of utterances is not well supported by empirical evidence. In music, it is clear that rhythmic predictability has an adaptive value: it allows the formation of a temporal expectancy scheme that plays an important role in musical perception (e.g., beat perception), and it guides the coordination of ensemble performance and the synchronization of movements in dance. Because speech does not have a regular beat, what functional role would rhythmic predictability play? One idea suggested by Lehiste (1977) is that it plays a role in signaling phrase boundaries in speech. Specifically, she suggested that one method speakers have for disambiguating syntactically ambiguous sentences is by signaling a structural boundary via lengthening of an interstress interval (ISI). For example, she studied speakers' productions of sentences such as "The old men and women stayed at home," which is syntactically ambiguous (either just the men were old or both the men and women were old). She found that when speakers said this sentence in such a way to make one or the other interpretation clear, the sequence "men and women" was very different in duration, being substantially longer when a syntactic boundary was intended between "men" and "women." Furthermore, she conducted a follow-up study in which the same sentences were resynthesized using a monotone, and the duration of the critical ISI was manipulated by uniformly expanding the duration of the phonemes within it, so that the relative durations of segments remained the same. She found that listeners were able to perceive the intended meaning solely on the basis of the length of the critical ISI, suggesting that ISI duration can signal a phrase boundary.

A subsequent study by Scott (1982) set out to test Lehiste's hypothesis against the more conventional notion that phrase boundaries are signaled by phrase-final lengthening. She found evidence for a weak version of Lehiste's hypothesis, in that there appeared to be some cases in which listeners used ISI patterns, but others in which they relied on traditional phrase-final lengthening. Nevertheless, the evidence was suggestive enough for this line of research to merit further study. A key conceptual point, however, is that evidence that ISI duration plays a role in creating perceived boundaries in speech is not equivalent to evidence for isochrony. The expectation for how long an ISI should be need not be based on an expectation for isochrony, but could be based on expectations for how long a given ISI should be given the number (and type) of syllables within it and the current speech rate (cf. Campbell, 1993). According to this view, a prosodic break is more likely to be heard when an ISI is significantly longer than expected, and rhythmic predictability is simply implicit knowledge of the statistical relation between ISI duration and the number and type of syllables in an ISI. This would allow a functional role for rhythmic predictability without any recourse to notions of isochrony.

The Role of Rhythm in Segmenting Connected Speech

To a native listener, spoken sentences consist of a succession of discrete words, yet this perception is an illusion. As pointed out in Chapter 2, word boundaries in language do not map in any simple way onto acoustic breaks in the speech signal, and as anyone who has listened to sentences in a foreign language can attest, it is far from obvious where the word boundaries in connected speech are. This problem is particularly relevant for infants, who are constantly faced with multiword utterances (van de Weijer, 1999) and who do not have the benefit of an existing vocabulary to help them identify where one word ends and the next begins.

A substantial body of research in psycholinguistics indicates that the rhythmic properties of a language assist a listener in segmenting speech. Work on English, for example, has pointed to a segmentation strategy based on stress: Listeners expect strong syllables to be word-initial. This likely reflects the predominance of words with initial stress in the English lexicon (Cutler & Carter, 1987), and manifests itself in a number of different ways in perception. For example, Cutler and Butterfield (1992) showed that when listeners missegment speech, they tend to place word boundaries before stressed syllables, as when "by loose analogy" is misheard as "by Luce and Allergy." Furthermore, when English speakers are asked to spot real monosyllabic words embedded in larger polysyllabic nonsense words, they find it easier when the real word does not straddle two stressed syllables. "Mint," for example, is easier to spot in "mintef" than in "mintayf," presumably because in the latter word the strong second syllable "tayf" triggers segmentation, thus splitting "mint" into two parts (Cutler & Norris, 1988). Cutler (1990) has dubbed the strategy of positing a word onset at each strong syllable the "metrical segmentation strategy."

Research on segmentation in other languages has revealed that stress-based segmentation is by no means universal. French and Spanish speakers, for example, favor syllabically based segmentation (Mehler et al., 1981; Pallier et al., 1993), whereas Japanese speakers favor moraic segmentation (Otake et al., 1993). Thus segmentation relies on units that are phonologically important in the native language. One striking finding of this cross-linguistic research is that the native language's segmentation strategies are applied *even when listening to a foreign language,* showing that segmentation tendencies are not simply a reaction to a particular speech rhythm, but a perceptual habit of a listener (see Cutler, 2000, for a review). One possibility suggested by Cutler is that this habit is a residue of early language learning, when rhythmic segmentation played an important role in bootstrapping lexical acquisition.

The relevance of this research to comparative studies of language and music is that it shows that experience with a language's rhythm leaves a permanent influence on a listener in terms of segmenting speech patterns, whether or not these patterns come from the native language. From this observation it is but one step to ask if experience with the native language influences how one segments nonlinguistic rhythmic patterns. This question is taken up in section 3.5.2 below.

The Role of Rhythm in the Perception of Nonnative Accents

When a person listens to their native language, s/he usually has a keen sense of whether or not it is being spoken with a native accent. Recent research on speech rhythm has taken advantage of this fact by having listeners judge the degree of perceived "foreign accentedness" in utterances spoken by nonnative speakers.

Empirical rhythmic measurements are then taken of speech of the different nonnative speakers. By examining the correlation between perceived degree of foreign accent and the quantitative rhythmic measures, researchers hope to identify the perceptual cues listeners use in gauging speech rhythm patterns.

Using this approach, White and Mattys (2007) examined Spanish speakers of English and found that the greater their vowel duration variability within sentences, the more native-sounding they were rated by English speakers. This probably reflects vowel reduction: Spanish speakers who learn to reduce vowels in unstressed syllables (a characteristic of English, but not of Spanish; cf. section 3.3.1, subsection "Phonology and Typology") are likely to sound more like native speakers. A consequence of vowel reduction within sentences is that vowel duration variability increases, because some vowels become very short.

As noted in section 3.3.1 (subsection "Duration and Typology"), another empirical measure of rhythm influenced by vowel reduction is the nPVI, which measures the degree of durational contrast between adjacent vowels in a sentence rather than overall durational variability. White and Mattys found that the vowel nPVI of Spanish speakers of English was positively correlated with how native they sounded. Crucially, however, vowel duration variability was a better predictor of accent judgment than was nPVI. This suggests that vowel duration variability may be more perceptually relevant for speech rhythm than durational contrastiveness.

This is a very promising approach, because different rhythmic measures can be pitted against each other to see which best predicts perceptual data. However, the findings to date must be considered tentative because of an uncontrolled variable. This is variability in the degree to which nonnative speakers accurately produce the phonemes of the second language (i.e., the individual vowels and consonants). When judging a speaker's degree of foreign accent, listeners almost certainly base their judgments on some combination of segmental and suprasegmental cues. This is a problem because some nonnative speakers may produce native-sounding prosody but nonnative sounding segmental material, or vice versa. To compound the problem, different listeners may vary in the extent to which they weight segmental versus suprasegmental cues in judging how "foreign" a given nonnative speaker sounds. Thus to truly focus listeners' attention on rhythm, segmental cues must be made uniform. Resynthesis techniques, such as those used by Ramus and colleagues (cf. section 3.3.1, subsection "Perception and Typology") might provide one way to make sentences spoken by different nonnative speakers uniform in terms of phonemic material while preserving prosodic differences.

3.3.4 Final Comments on Speech Rhythm: Moving Beyond Isochrony

Although the history of speech rhythm research is tightly bound up with notions of periodicity (e.g., the isochrony of stresses or syllables), the evidence reviewed above suggests that the case for periodicity in speech is extremely weak. Thus progress in the study of speech rhythm requires conceptually decoupling "rhythm" and "periodicity," a point made in the introduction of this chapter. It is quite clear that speech has rhythm in the sense of systematic temporal, accentual, and grouping patterns of sound, and languages can be similar or different in terms of these patterns. However, the rhythms of language are not based on the periodic occurrence of any linguistic unit. Instead, the patterning is largely the *by-product* of phonological phenomena, such as the structure of syllables, vowel reduction, the location of lexical prominence, stress clash avoidance, and the prosodic phrasing of sentences. These phenomena lead to differences in the way utterances are organized in time.

The notion that rhythm in language is primarily *consequence* rather than *construct* stands in sharp contrast to rhythm in music, in which patterns of timing and accent are a focus of conscious design. Another salient difference between rhythm in speech and music, related to the lack of a periodic framework for speech rhythm, is the fact that speech rhythm conveys no sense of motion to a listener (cf. section 3.2.5). Do these differences mean that rhythm in language and music cannot be meaningfully compared? Absolutely not. As demonstrated in section 3.5 below, empirical comparisons are not only possible, they can be quite fruitful. They have had nothing to do with periodicity, however.

For those interested in cross-domain studies of rhythm, it is heartening to note that there is renewed interest in empirical studies of rhythm in speech production and speech perception (Ramus et al., 1999; Ramus & Mehler, 1999; Low et al., 2000; Grabe & Low, 2002; Lee & Todd, 2004, White & Mattys, 2007), and that there is much room for further work. For example, there is a need for more empirical data on listeners' judgments of how native-sounding a foreign speakers' utterances are, from the standpoint of rhythm. Such studies will need to employ creative ways of isolating the rhythm of speech from other phonetic dimensions of language, for example, using resynthesized speech in which phonetic content and pitch contours can be completely controlled (cf. Ramus & Mehler, 1999). There is also a need for studies that measure temporal patterning at multiple linguistic levels and that quantify relations between levels. It may be that important perceptual dimensions of speech rhythm are relational, such as having a high degree of contrast between adjacent syllable durations while simultaneously having a low degree of contrast between the duration of interstress intervals (some data pertinent to this idea are given at the end of this section). This is an area in which collaborations between linguists and music researchers would be especially useful.

In the remainder of this section, I would like to consider why periodicity has been (and continues to be) such an enduring concept in speech rhythm research. Below I offer several reasons for this historical phenomenon.

The simplest reason, of course, is the mistaken notion that rhythm *is* periodicity, or that rhythm *is* a regular alternation between strong and weak beats, rather than the broader notion of rhythm as systematic temporal, accentual, and phrasal patterning of sound, *whether or not this patterning is periodic*. Indeed, one need not look beyond music to see that a definition of rhythm *as* periodicity or *as* strong-weak beat alternation is overly simplistic: Many widespread musical forms lack one and/or the other of these features yet are rhythmically organized (cf. section 3.2).

The second reason that the notion of periodicity has endured may be the idea that it has a useful function in speech perception, such as making salient information predictable in time. There are psychological theories of auditory perception that propose that attention can be allocated more efficiently when events are temporally predictable, based on the idea that auditory attention employs internal oscillatory processes that synchronize with external rhythmic patterns (e.g., Jones, 1976; Large & Jones, 1999). Such theories provide a rationale for those interested in the idea that periodicity in speech is perceptually adaptive. Alternatively, those interested in periodicity might claim that it is useful because it creates a framework within which deviations are meaningful. This is the basis of Lehiste's idea that lengthening of interstress intervals in English can be used to mark phrase boundaries (cf. section 3.3.3, subsection "The Role of Rhythmic Predictability in Speech Perception"). The principal drawback of these perception-based arguments for periodicity is that the evidence for them is very weak. Although further research is needed, the current evidence suggests that periodicity does not have an important role to play in normal speech perception. This should not be surprising: The comprehension of speech *should be* robust to variation in the timing of salient events, because such variations can occur for a number of reasons. For example, a speaker may suddenly speed up or slow down for rhetorical reasons within a conversation. Under such conditions, relying on periodicity for comprehension seems a maladaptive strategy.

The third reason for periodicity's allure may be the belief that because various temporal patterns in human physiology (e.g., heartbeat, walking, chewing) exhibit periodic structure, speech is also likely to be periodic, perhaps even governed by rhythmic pattern generators. However, the use of rhythmic neural circuits for speech is not particularly plausible. The constant use of novel utterances in language means that articulators must be coordinated in different ways each time a new sentence is produced. Furthermore, the maneuvers that produce particular speech sounds depend on the local context in which they occur. Thus the motor patterns of speech cannot be predicted in advance with a high degree of precision. Without stereotyped movement patterns, evolution has no grounds for placing the control of speech in a rhythmic neural circuit. An analogy here is to multifingered touch-typing, a behavior involving the sequencing of overlapping movements of multiple articulators (the fingers). Although touch-typing is highly temporally organized, the resulting sequences are not based on periodic movements.

So far I have focused on negative reasons for the persistence of the concept of periodicity in speech. I will now briefly speculate on the positive reasons for the persistence of this concept, in other words, why periodicity in speech has been such an intuitively appealing notion to speech researchers, particularly those whose native language is English. (It is notable that the idea of periodicity in speech was promulgated by linguists who were English speakers, and that arguments for stress isochrony in English have been present since at least the 18th century; cf. Abercrombie, 1967:171; Kassler, 2005). First, it seems that English speakers find the interstress interval (ISI) to be a salient temporal unit in speech. For example, in a preliminary study, Cummins (2002) asked English listeners to repeat nonsense phrases such as "manning the middle" in time with an external pacing cue, so that the two stressed syllables were perceptually aligned with a periodically repeating two-tone pattern (for example, "man" would align with a high tone, and "mid" with a low tone; cf. Cummins & Port, 1998). This is equivalent to aligning the start and end of the ISI with two tones. Cummins also tested speakers of Italian and Spanish because these languages have lexical stress, permitting phrases to be constructed in a manner analogous to the English phrases (such as "BUSca al MOto" in Spanish, stress indicated by capitalization). Cummins observed that although English speakers learned the task quickly and performed accurately, Spanish and Italian speakers took much longer, were uncomfortable with the task, and produced a great deal of variability in their results. Cummins suggests that this difference is due to the fact that the ISI is not a salient perceptual unit for speakers of Italian and Spanish, despite the fact that there is lexical stress in these languages.

This intriguing finding raises the question of why ISI is salient to English listeners. Does it play some functional linguistic role? As discussed in section 3.3.3 (subsection "The Role of Rhythmic Predictability in Speech Perception"), ISI duration may play a role in signaling linguistic boundaries to English listeners, even if ISIs are not isochronous. Currently there is not enough evidence to say confidently what role the ISI plays, but let us assume for a moment that English speakers and listeners are sensitive to it as an entity. Given the empirical observations about the large variability in ISI duration in English (e.g., coefficients of variation around 33%; Dauer, 1983), why would listeners ever feel that ISIs were isochronous? One answer may concern the relative degree of variability in ISI durations compared to syllable durations. As suggested above, the impression of isochrony may be due in part to a lower degree of contrast between successive ISI durations relative to successive syllable durations. For example, consider Figure 3.12, which shows the same sentence as Figure 3.8a and Sound Example 3.5a ("the **last con**cert **gi**ven at the **o**pera was a tre**men**dous suc**cess**").

Syllable boundaries are marked with vertical lines (as in Figure 3.8), but now stressed syllables (indicated in by boldface above) have been marked with

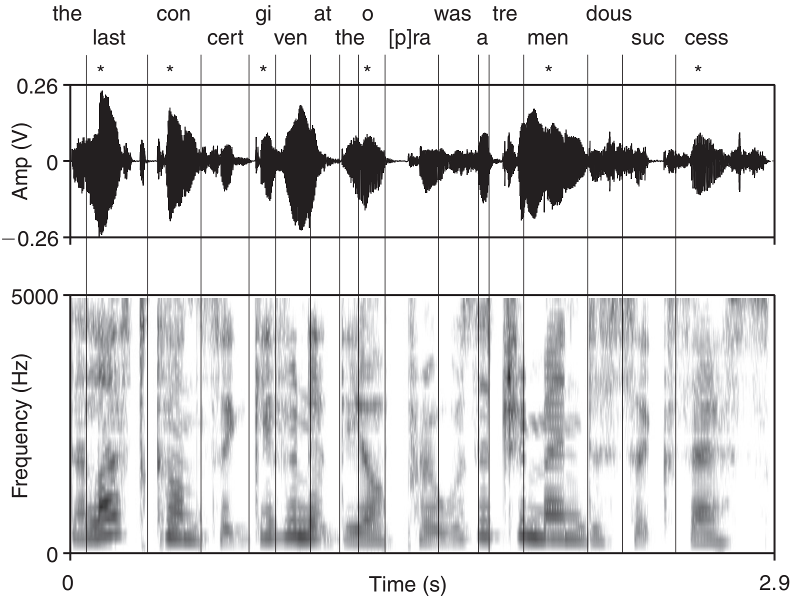

Figure 3.12 The English sentence of Figure 3.8a, with stresses marked by asterisks (*).

an asterisk. The asterisk was placed at the vowel onset of the stressed syllable, in other words, near its perceptual attack (its "P-center"; Morton et al., 1976; Patel et al., 1999). In this sentence, the nPVI of syllable durations is 59.3, and the nPVI of ISIs is 28.4, making the ratio $\text{nPVI}_{\text{ISI}}/\text{nPVI}_{\text{syll}}$ equal to 0.48. Thus in this particular case, the amount of durational contrast between adjacent ISIs is only about half of that between adjacent syllables.

I have computed similar ratios for each of the 20 sentences of British English from Ramus's database.[24] For 15 out of 20 sentences, this ratio was less than 1. The overall mean ratio across the 20 sentences was .83 (std = .45), and was significantly less than 1 ($p < .0001$ by a one-tailed t-test). An even stronger effect was observed when one computes the ratio of ISI duration variability to syllable duration variability (using the coefficient of variation, i.e., $\text{CV}_{\text{ISI}}/\text{CV}_{\text{syll}}$) Here 17 out of 20 sentences had a value less than 1, and the mean was .69 (std = .25), again significantly less than 1. Thus if the ear is sensitive to temporal patterning at the levels of both syllables and stresses, the low durational variability of ISIs *relative to the variability of syllable durations* might contribute to a sense that stresses are temporally regular. Of course, for this explanation to have any merit it must be shown that the ratio of ISI to syllable variability differentiates stress-timed from syllable-timed

[24] I am grateful to Laura Dilley for marking stressed syllables in these sentences.

languages. Languages such as Italian and Spanish would be good candidates for testing this hypothesis, because they are syllable-timed languages in which stress can be reliably marked.

Although the notion of isochrony in speech continues to beguile researchers, I suspect that it will have little or no role in the most fruitful research on speech rhythm in the coming years. Isochrony was important to the birth of speech rhythm studies, but it is a concept whose usefulness is exhausted. It is time to move on to a richer view of speech rhythm.

3.4 Interlude: Rhythm in Poetry and Song

As in the rest of this book, the focus of this chapter is on comparing ordinary speech to instrumental music. However, no comparison of rhythm in language and music is complete without a discussion of poetry and song. In these art forms, words are carefully chosen and consciously patterned for rhythmic effect. Of course, poetry and song are but two of numerous vocal genres with organized rhythms. In the United States certain styles of preaching in African American churches are notable for their rhythmic patterning, a taste of which can be heard in Martin Luther King Jr.'s famous "I Have a Dream" speech. In other cultures, it is possible to identify numerous genres of speech in which rhythmic design plays a role (see Agawu, 1995, for a fascinating case study of the range of rhythmically regulated forms of speech in an African society). The focus here is on poetry and song, however, because these have received the greatest amount of empirical research in terms of rhythm.

3.4.1 Rhythm in Poetry

The study of poetic rhythm has been the focus of a good deal of research by literary scholars (for introductions, see Gross, 1979; Fussell, 1979; Hollander, 2001). In this tradition, poetic "meter" refers to the abstract patterning scheme which governs the temporal structure of a poem, whereas "rhythm" refers to the actual patterning of durations and accents. For example, a great deal of English verse is written in iambic pentameter, a verse form consisting of 5 iambic feet, in which an iamb is a (weak + strong) syllable pattern. Naturally there are many exceptions to this pattern within iambic pentameter poetry: A particularly common one is the substitution of a trochaic foot, or (strong + weak) syllable pattern at the onset of a line. Thus the rhythm of a particular line may violate the overall meter of the poem.

Literary prosodists argue that listeners internalize the regularities of meter and perceive departures from this scheme as variation from a stable background (Richards, 1979:69; Adams, 1997:12). That is, meter is seen as having an intimate relationship with expectancy. This idea is related to the notion of musical meter as an abstract mental scheme, but differs from musical meter in an important way. Musical meter refers to *temporal* periodicity, whereas poetic meter involves *configurational* periodicity, in other words, the focus is on the repetition of some basic prosodic unit rather than on temporal periodicity per se. For example, in iambic pentameter it is the weak + strong configuration of the iambic foot that is the design focus, not the isochrony of stressed syllables. In various forms of French and Chinese verse, the number of syllables per line is strictly regulated, but there is no focus on making syllables periodic (equal in duration).

It is interesting to note that different languages tend to favor different kinds of poetic meters. For example, English verse has often tended toward purely stress-based forms in which regulation of the number of stresses per line is the focus, independent of the number of syllables (e.g., the meter of *Beowulf*, with four stresses per line). In contrast, English verse based on regulating the number of syllables per line without regard to stress is rare (Fussell, 1979:62–75). This likely reflects the powerful role of stress in ordinary English speech rhythm. Indeed, Fussell has argued that "a meter customary in a given language is customary just because it 'measures' the most characteristic quality of the language" (Fussell, 1974:498). Thus stress plays a dominant role in English poetry, but little role in French, in which the number of syllables per line is a more common concern. Japanese, in turn, often regulates the number of morae per line, as in the 5–7–5 mora structure of the haiku.

Lerdahl and Halle (1991) and Lerdahl (2003) have sought to unify the theoretical treatment of rhythm in poetry and music, using shared concepts such as hierarchical grouping structure and metrical grids. Our focus here, however, is on empirical research. Over the past few decades, poetic rhythm has attracted the interest of a number of speech scientists, who have made quantitative measurements of the temporal patterns of poetry. For example, the phonetician Gunnar Fant and colleagues have studied the acoustics of iambic versus trochaic lines of verse in Swedish (1991b). The researchers found that in iambic feet, the weak syllable is about 50% as long as the following strong syllable, whereas in trochaic feet, the weak syllable is about 80% of the duration of the preceding strong syllable (cf. Nord et al., 1990). This difference is likely due to preboundary lengthening, which acts to increase the duration of the final syllable in each foot (i.e., the strong syllable in an iamb and the weak syllable in a trochee). Thus iambic and trochaic feet are not simply mirror images of each other in terms of their temporal profiles: Iambic feet are much more temporally asymmetric.

These observations may be relevant to the study of the aesthetic effect of the two kinds of feet in poetic lines. For example, Adams (1997:55–57) notes that trochaic meters are often associated with awe and the suspension of reality, as in Blake's poem, "The Tyger," in which trochaic patterns dominate the first three lines of the first stanza:

(3.16)
Tyger! Tyger! burning bright
In the forests of the night
What immortal hand or eye
Could frame thy fearful symmetry?

This aesthetic property of trochaic meter may be partly due to its more uniform profile of syllabic durations, which goes against the grain of normal English speech rhythm and thus gives the resulting speech an incantatory feel.

Another prominent phonetician who has long conducted research on poetic rhythm is Ilse Lehiste (1991). In one set of studies, Lehiste examined the relationship between the timing of feet and of the lines in which feet are embedded. She found that the temporal variability of lines is lower than one would predict based on the variability of feet duration, suggesting that speakers make temporal compensations between feet in order to keep lines within a certain duration. That is, lines act as a unit of temporal programming in the recitation of poetry (Lehiste, 1990). Ross and Lehiste (1998, 2001) have also examined the interplay of linguistic and poetic rhythm in framing the temporal patterns of Estonian verse and folksongs.

3.4.2 Rhythm in Song

For languages with clearly defined stress, such as English, each phrase or sentence comes with a distinct pattern of stronger and weaker syllables. When words in these languages are set to metrical music, a relationship is established between the syllabic accent patterns and musical metrical accent patterns. Sensitivity to these relationships is part of the skill of writing music with words, and empirical research suggests that composers exploit this relationship for artistic ends.

Palmer and Kelly (1992) studied vocal lines in themes from Gilbert and Sullivan's 14 operettas, focusing on compound nouns (like the single word "blackbird") and adjective-noun pairs (like the two word phrase "black bird"). In English, compound nouns receive stress on the first syllable, whereas adjective-noun pairs receive stress on the second syllable. They studied how such words were aligned with the metrical structure of the music, and found that the stressed syllable tended to align with a metrically strong beat in the music. Given the complex texts of Gilbert and Sullivan's songs, this strategy of alignment may contribute a sense of precision and balance to the lyrics of these operettas.

Temperley (1999) looked at a different genre of vocal music, namely rock songs. In contrast to Palmer and Kelly, he found that verbal stress frequently *anticipated* metrical accent in rock songs by a fraction of a beat, as in the

Beatles' "Here Comes the Sun" (Figure 3.13). This systematic anticipation contributes a sense of syncopation and rhythmic energy to the song, and provides an example of how the systematic *misalignment* of verbal and musical stress adds dynamic energy to music.

The relationship between rhythm in speech and song is a fertile area that merits much more empirical investigation than it has received to date. In the remainder of this section, I outline three directions that this research could take. The first pertains to cultural differences in the prevalence of certain types of musical rhythms. For example, Yamomoto (1996) notes that children's songs based on triple rhythms (e.g., 6/8 time signature) are rare in Japan but common in Britain, and suggests that this might be due to differences in English versus Japanese speech rhythm. If Yamomoto is correct, one would predict that Japanese- versus English-speaking children would differ in how easily they can learn songs in these meters. Another language with which one could test a similar idea is Greek. Recall from section 3.3.2 (subsection "Questioning the Principle of Rhythmic Alternation in Speech") that Arvaniti (1994) showed that Greek, a language of the Balkan region, tolerates a more irregular alternation between stressed and unstressed syllables than does English. Also recall from section 3.2 that the Balkan region features music with irregularly spaced beats. Would Greek-speaking children find it easier to learn the irregular meters of Balkan songs than English-speaking children (cf. Hannon & Trehub, 2005)? In the studies outlined above, it would of course be essential that the two groups of children be matched for prior musical exposure to different musical meters. It may thus be best to work with immigrants who speak the native language at home but whose children are exposed to Western music. In such a case, if learning experiments reveal the predicted cultural differences, this would support the interesting hypothesis that a culture's speech rhythm predisposes it toward or away from certain musical rhythms.

A second direction for research in this area is to examine verbally improvised music that is accompanied by a rhythmic musical context, such as contemporary rap music. If the vocal and musical lines can be recorded on different audio tracks, and points of verbal and musical stress can be independently identified,

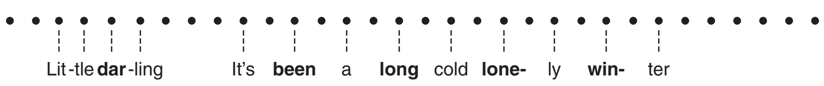

Figure 3.13 A musical metrical grid for a portion of the Beatles' "Here Comes the Sun." The lyrics are aligned below the grid, and linguistically stressed syllables are indicated in boldface: Note how most such syllables slightly precede strong metrical positions in the music. From Temperley, 1999.

then one could study temporal relations between verbal and musical accent points as a piece unfolds in time. It would be particularly interesting to study these relations in novice versus expert rap musicians, to see if part of being an expert in this genre is greater flexibility and/or precision in the manner in which the alignment of the two types of accents is handled. In a study such as this, identifying precise points in time for verbal and musical accent will be essential, and the issue of the perceptual attack time of syllables and of musical tones ("P-centers") comes to the fore (Morton et al., 1976; Gordon, 1987; Patel et al., 1999).

A final possible line of research is suggested by a correspondence between Richard Strauss and Romain Rolland in 1905 about musical text setting (Myers, 1968).[25] Strauss emphasizes the clear-cut relationship between syllabic accent in speech and metrical accent in music: "In German 'she' on the strong beat of a bar is absolutely impossible. For example, in a bar of 4/4, the first and third beat always have a necessary stress which can only be made on the radical [stressed] syllable of each word." He also expresses his frustration over variability in the alignment of word stress and musical stress in French opera: "Yesterday, I again read some of Debussy's Pélleas et Mélisande, and I am once more very uncertain about the principle of the declamation of French when sung. Thus on page 113, I found: 'Cheveúx, chéveux, dé cheveux.' For heaven's sake, I ask you, of these three ways there can all the same only be *one* which is right."

Rolland replies by emphasizing the mutability and subtlety of French word accent:

> The natural value of "cheveux" is chevéux. But a man in love will, when saying this word, put quite a special stress on it: "tes chéveux." . . . You see, the great difficulty with our language is that for a very large number of words, accentuation is variable,—never arbitrary, but in accordance with logical or psychological reasons. When you say to me: . . ." Of these 3 (cheveux) *only one can be right,* what you say is doubtless true of German, but not for French.

What this correspondence suggests is that German text setting is more rigid in its alignment of verbal and musical accent, whereas French is more permissive in terms of accent alignment between text and music. Indeed, Dell and Halle (in press; cf. Dell, 1989) report that French text-setting is quite tolerant of mismatches between verbal and musical accent, except at the ends of lines, where alignment tends to be enforced. They contrast this high degree of "mismatch tolerance" in French songs to a much lower degree found in English songs. The correspondence between Strauss and Rolland, and the work of Dell and Halle, suggest that languages have salient differences in the way they align music and text in terms of rhythmic properties, though quantitative work is

[25] I am grateful to Graeme Boone for bringing this correspondence to my attention.

needed to confirm this. These writings also lead to ideas for cross-cultural perceptual studies testing sensitivity to accent mismatches in songs. Specifically, listeners could be presented with different versions of a song that have text and tune aligned in different ways, with one version having many more accent mismatches. Listeners could then be asked to judge in which version the words and music go best together. One might predict that German listeners judging German songs (or English listeners judging English songs) would be more selective in terms of the pairings that sound acceptable than French listeners judging French songs.

3.5 Nonperiodic Aspects of Rhythm as a Key Link

One major theme of this chapter is that languages have rhythm (systematic temporal, accentual, and grouping patterns), but that this rhythm does not involve the periodic recurrence of stresses, syllables, or any other linguistic unit. Initially it may seem that "giving up on periodicity in speech" would mean that there is little basis for comparing rhythm in music and language. In fact, the opposite is true. By abandoning a fixation on periodicity one is freed to think more broadly about speech rhythm and its relationship to musical rhythm. As we shall see below, a focus on nonperiodic aspects of linguistic rhythm is proving fruitful in terms of comparing language and music at structural and neural levels.

3.5.1 Relations Between Musical Structure and Linguistic Rhythm

The notion that a nation's instrumental music reflects the prosody of its language has long intrigued music scholars, especially those interested in "national character" in music. Gerald Abraham explored this idea at length (1974, Ch. 4), noting as one example an observation of Ralph Kirkpatrick on French keyboard-music: "Both Couperin and Rameau, like Fauré and Debussy, are thoroughly conditioned by the nuances and inflections of spoken French. On no Western music has the influence of language been stronger" (p. 83). In a more succinct expression of a similar sentiment, Glinka (in *Theater Arts,* June 1958) wrote: "A nation creates music, the composer only arranges it" (cited in Giddings, 1984:91).

Until very recently, evidence for this idea has been largely anecdotal. For example, Garfias (1987) has noted that in Hungarian each word starts with a stressed syllable, and that Hungarian musical melodies typically start on strong beats (i.e., anacrusis, or upbeat, is rare). Although this is an interesting observation, it is possible that this is due to the fact that many such melodies come from folk songs. In this case, the linguistic influence on musical rhythm would be mediated by text. A more interesting issue, implied by Kirkpatrick, is whether linguistic rhythm influences the rhythm of instrumental music, in other words, music that is not vocally conceived.

One approach to this question was suggested by Wenk (1987). He proposed that cultures with rhythmically distinct languages should be examined to see if differences in musical rhythm reflect differences in speech rhythm. Wenk focused on English and French, prototypical examples of a stress-timed versus a syllable-timed language. Wenk and Wioland (1982) had previously argued that a salient rhythmic difference between the two languages was that English grouped syllables into units *beginning* with a stressed syllable, whereas French grouped syllables into units *ending* with a stressed syllable, as in:

x x x

(3.17a) / Phillip is / studying at the uni/versity

x x x

(3.17b) / Philippe / étudie / à l'université

Wenk and Wioland further argued that the stress at the ends of rhythmic groups in French was marked primarily by durational lengthening. Based on this idea, Wenk (1987) predicted that phrase-final lengthening would be more common in French versus English instrumental music. He tested this idea by having a professional musician mark phrase boundaries in English versus French classical music. The number of phrases in which the final note was the longest note in the phrase was then tallied for both cultures. Wenk found that it was indeed the case that more such phrases occurred in French than in English music.

Wenk's study was pioneering in its empirical orientation, but it also had limitations that make it difficult to accept these findings as a firm answer to the question of interest. Only one composer from each culture was examined (Francis Poulenc and Benjamin Britten), and from the oeuvre of each composer, only one movement from one piece was selected. Furthermore, no comparable empirical data for language rhythm were collected (e.g., the degree of phrase-final lengthening in English vs. French speech).

Despite its limitations, Wenk's study outlined a useful approach, namely to identify empirical rhythmic differences between two languages and then determine if these differences are reflected in the music of the two cultures. Pursuing this idea in a rigorous fashion entails three requirements. First, an empirical measure of speech rhythm is needed to quantify rhythmic differences between languages. Second, this same measure should be applicable to music so that language and music could be compared in a common framework. Third, both the linguistic and musical samples needed to be broad enough to insure that the findings are not idiosyncratic to a few speakers or composers.

Joseph Daniele and I conducted a study that set out to meet these criteria (Patel & Daniele, 2003a). Like Wenk, we focused on British English and French due to their distinct speech rhythms and because they have been the locus of strong intuitions about links between prosody and instrumental music (e.g., Hall, 1953; Abraham, 1974; Wenk 1987). Our work was inspired by recent phonetic research on empirical correlates of stress-timed versus syllable-timed speech rhythm (cf. section 3.3.1, subsection "Duration and Typology"). In particular, the work of Low, Grabe, and Nolan (2000) attracted our attention because it focused on something that could be measured in both speech and music, namely the durational contrast between successive elements in a sequence. Their measure, called the normalized pairwise variability index, or nPVI, had been applied to vowels in sentences from stress-timed and syllable-timed languages, and had been shown to be higher in stress-timed languages, likely due to the greater degree of vowel reduction in these languages (Grabe & Low, 2002; Ramus, 2002a; Lee & Todd, 2004; see the above-mentioned subsection of 3.3.1 for background on the nPVI).

Two aspects of this measure made it appealing for use with music. First, the nPVI a is purely relative measure of contrast. That is, the durational difference between each pair of intervals is measured *relative to* the average duration of the pair. This normalization, which was originally introduced to control for fluctuations in speech rate, makes the nPVI a dimensionless quantity that can be applied to both language and music. (For example, nPVI can be computed from speech durations measured in seconds and from musical durations measured in fractions of a beat.) Second, the nPVI has been applied to vowels. Vowels form the core of syllables, which can in turn be compared to musical tones (i.e., in setting words to music it is quite common for each note to be assigned to one syllable).[26] Our strategy, then, was to apply the nPVI to tone sequences from British and French instrumental music, to determine if differences emerged that reflected the rhythmic differences between British English and French speech.

Figure 3.14 shows the nPVI to British English versus continental French speech, based on measurements of vowel durations in sentences uttered by native speakers of each language. (The sentences are short, news-like utterances from the corpus of Nazzi et al., 1998.)[27]

[26] Although this is true for English and French, it should be noted that in Japanese it is the mora and not the syllable that gets mapped onto a musical note (Hayes, 1995a).

[27] The nPVI values for English and French speech shown in Figure 3.14 are taken from Patel et al. (2006), rather than from Patel and Daniele (2003a). Both studies show a significant difference between the two languages (English nPVI > French nPVI), but the 2006 study is based on more accurate measurements. See Patel et al. (2006) for measurement details, and for a list of all sentences analyzed.

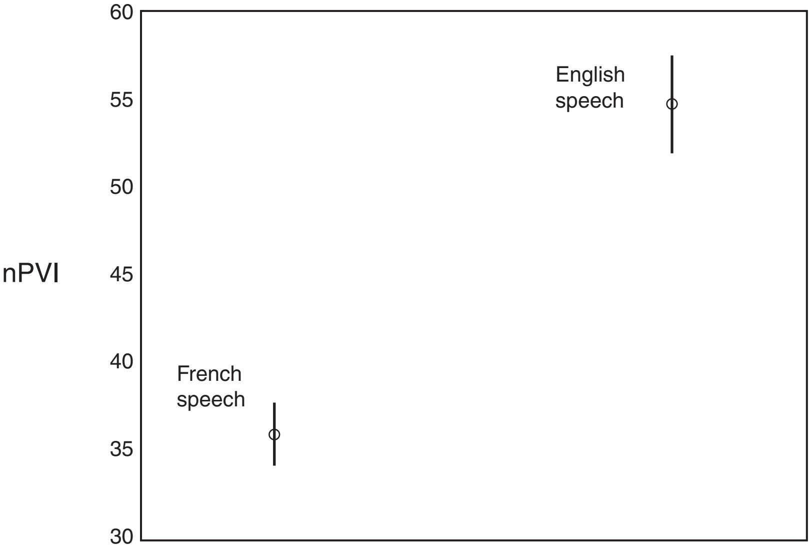

Figure 3.14 The nPVI of British English and French sentences. Error bars show +/- 1 standard error. Data from Patel, Iversen, & Rosenberg, 2006.

The nPVI is significantly higher for English than for French speech. Figure 3.15 gives an intuition for why this is the case by illustrating the pattern of vowel duration for one English and French sentence in this corpus (cf. Sound Examples 3.8a, b).

For example, in the top panel, the first two values (about 120 ms and 40 ms) are the durations of the first two vowels in the sentence (i.e., the vowels in "Finding"), and so on. Note how successive vowels tend to differ more in duration for the English sentence than for the French sentence. In the English sentence, some vowels are very short (often due to vowel reduction), whereas other vowels are quite long (often due to stress). This leads to a greater tendency for durational contrast between neighboring vowels, which is reflected in the nPVI score.

As mentioned above, an appealing aspect of the nPVI is that it can be applied to music in order to measure the durational contrast between successive notes. Western music notation indicates the relative duration of notes in an unambiguous fashion, as shown in Figure 3.16.

In the figure, the first note of each theme is arbitrarily assigned a duration of 1, and the durations of the remaining notes are expressed as a multiple or fraction of this value. (Any numerical coding scheme that preserves relative duration of notes would yield the same nPVI, because it is a normalized measure.) In this example, the nPVI of the Debussy theme is lower than that of the Elgar theme, even though the raw variability of note duration in the Debussy theme is

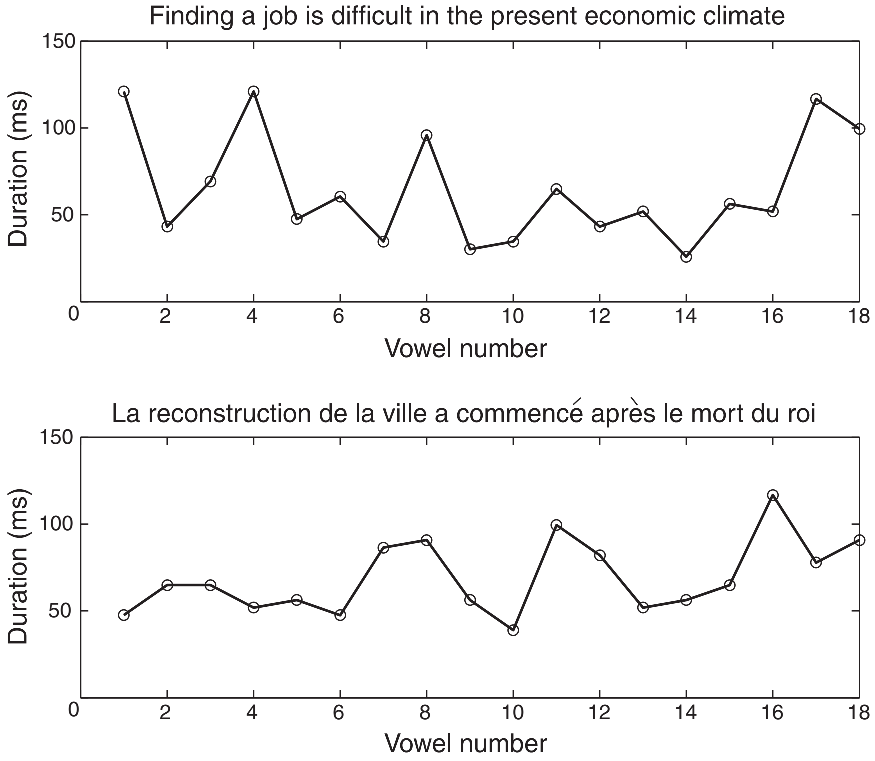

Figure 3.15 Vowel durations in an English and a French sentence. Note the greater degree of short-long contrast in the English sentence between adjacent vowel durations. The nPVI for the English sentence is 54.9, and for the French sentence is 30.0.

greater than that in the Elgar theme (as measured by the coefficient of variation, in other words, the standard deviation divided by the mean). This emphasizes the fact that the nPVI indexes the degree of contrast between successive elements in a sequence, not the overall variability of those elements.

Our source of musical material was a standard reference work in musicology, *A Dictionary of Musical Themes,* Second Edition (Barlow & Morgenstern, 1983), which focuses on the instrumental music of Western European composers. In choosing composers to include in our study we were guided by two factors. First, the composers had to be from a relatively recent musical era because measurements of speech prosody are based on contemporary speech, and languages are known to change over time in terms of sound structure. Second, the composers had to be native speakers of British English or French who lived and worked in England or France. Using these guidelines, we examined all English and French composers from Barlow and Morgenstern who were born in the 1800s and died in the 1900s, and who had at least five musical

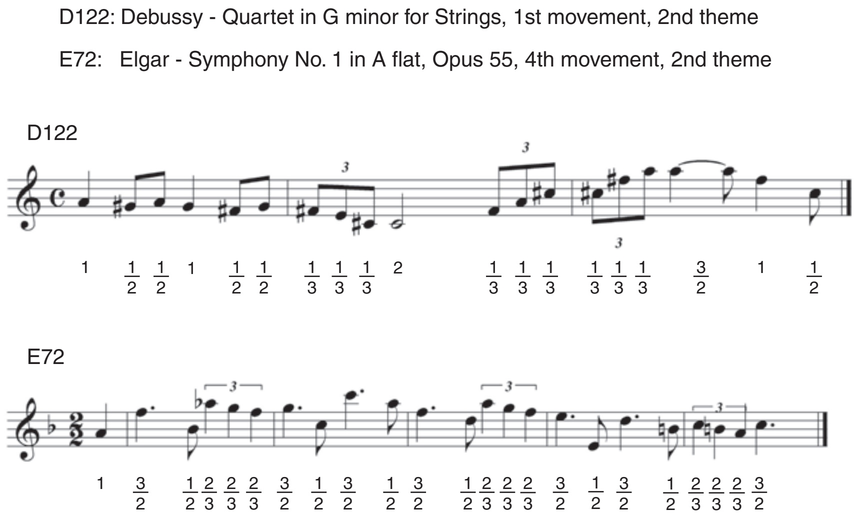

Figure 3.16 Two musical themes with the relative durations of each note marked. nPVI of D122 = 42.2, of E72 = 57.1. Themes are from Barlow & Morgenstern, 1983. From Patel & Daniele, 2003a.

themes in the dictionary that were eligible for inclusion in the study (see Patel & Daniele, 2003a, for inclusion criteria, and for the rationale of using music notation rather than recorded music for nPVI analysis). We chose composers who spanned the turn of the century because this era is noted by musicologists as a time of “musical nationalism” in Europe.

Based on our criteria, 16 composers were included in the study, including English composers such as Elgar, Delius, and Vaughan Williams, and French composers such as Debussy, Poulenc, and Saint-Saëns. About 300 musical themes were represented, and one musical nPVI value was computed for each theme. The results of our analysis of musical nPVI are shown in Figure 3.17, along with the speech nPVI values. Remarkably, the two cultures have significantly different musical nPVI values, with the difference being in the same direction as the linguistic nPVI difference (see Patel & Daniele, 2003a, and Patel et al., 2006, for further details).

Thus there is empirical evidence that speech rhythm is reflected in musical rhythm, at least in turn-of-the century classical music from England and France. How is this connection between language and music mediated? Some musicologists have proposed that national character arises from composers adapting folk melodies into their compositions. Because such melodies are typically from songs, it may be that the rhythm of words influences the rhythm of these melodies, thus giving the melodies a language-like rhythmic pattern. However, we believe that this may not be the best explanation for our finding, because our

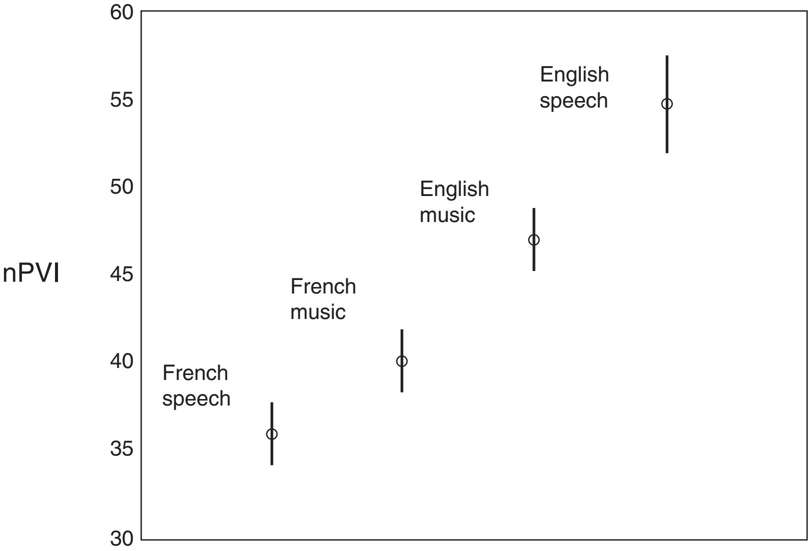

Figure 3.17 The nPVI of British English and French musical themes. Error bars show +/- 1 standard error. Data from Patel & Daniele, 2003a, and Patel, Iversen, & Rosenberg, 2006.

study included numerous composers who are not thought to be strongly influenced by folk music, such as Elgar and Debussy (Grout & Palisca, 2000). Instead, we feel there may be a more direct route from language to music. It is known from studies of language acquisition that the perceptual system is sensitive to the rhythmic patterns of language from a very early age (Nazzi et al., 1998; Ramus, 2002b). Composers, like other members of their culture, internalize these patterns as part of learning to speak their native language. (One mechanism for this internalization is a process called statistical learning, which is discussed in more detail in the next chapter.) We suggest that when composers write music, linguistic rhythms are "in their ears," and they can consciously or unconsciously draw on these patterns in weaving the sonic fabric of their music. This does not imply that the connection between linguistic and musical rhythm is obligatory. Rather, this link is likely to be greater in historical epochs where composers seek a national character for their music.

Our findings for English and French speech and music immediately raised two questions. Would the musical nPVI difference be observed if a broader sample of English and French themes and composers were studied? Perhaps more importantly, would our result generalize to other cultures in which stress- versus syllable-timed languages are spoken? Fortunately, Huron and Ollen (2003) provided answers to these questions. Using an electronic version of *A Dictionary of Musical Themes* created by Huron, they computed the nPVI of a much larger sample of English and French musical themes (about 2000 themes, composed between the mid-1500s and mid-1900s). They confirmed that the nPVI of English music was significantly higher than that of French music, though the difference was smaller than that found by Patel and Daniele (likely due to less stringent sampling criteria). They also computed the musical nPVI for a range of other nations, analyzing almost 8,000 themes from 12 nationalities over more than 3 centuries. Of the nationalities they examined, five can be assigned to stress-timed languages and three to syllable timed languages (Fant et al., 1991a; Grabe & Low, 2002; Ramus, 2002b). These are listed in Table 3.1 along with their musical nPVI values. (The data in table 3.1 represent corrected values of the original table in Huron & Ollen, 2003, kindly provided by David Huron. See this chapter's appendix 2 for data from more cultures.)

Four out of the five nations with stress-timed languages (American, Austrian, English, and Swedish) do indeed have higher musical nPVI values than the three nations with syllable-timed languages, providing support for the idea that stress-timed and syllable-timed languages are associated with distinctive musical rhythms. However, German music is a notable exception: It has a low musical nPVI value despite the fact that German is a stress-timed language with a high nPVI value for speech (Grabe & Low, 2002; Dellwo, 2004).

However, there may be a historical reason why German music has a low nPVI, namely the well-known influence of Italian music on German music (Kmetz et al., 2001). Because Italian music has a low nPVI, stylistic imitation of this music might outweigh any linguistic influence of the German language on the nPVI of German music. One way to test this idea is to examine the nPVI

Table 3.1 Musical nPVI Values for Eight Different Nationalities

	Musical nPVI	
	Mean	S.E.
Nationalities With Stress-Timed Languages		
American	46.7	1.0
Austrian	45.1	0.6
English	45.6	0.9
German	43.2	0.6
Swedish	50.0	2.4
Nationalities With Syllable-Timed Languages		
French	43.4	0.7
Italian	41.4	1.0
Spanish	42.5	1.9

in historical perspective, for example, as a function of each composer's birth year. When themes from 14 German composers were examined in this fashion a striking trend emerged, as shown in Figure 3.18 (Patel & Daniele, 2003b; Daniele & Patel, 2004).

Over the course of 250 years, nPVI almost doubled, a trend that is highly statistically significant. (Interestingly, this trend is also evident for the six Austrian composers we included in our study.) Given what is known about the history of the German language, this is unlikely to reflect a change in the rhythm of German from syllable-timed to stress-timed during this period (C. Heeschen, personal communication). Instead, it most likely reflects historical changes in musical style, perhaps including a waning influence of Italian music on German music over this period. In fact, the finding would be consistent with the idea that Italian music had a strong influence on German music during the Baroque era (1600–1750), less influence during the Classical era (1750–1825), and the least influence during the Romantic era (1825–1900). More generally, it suggests

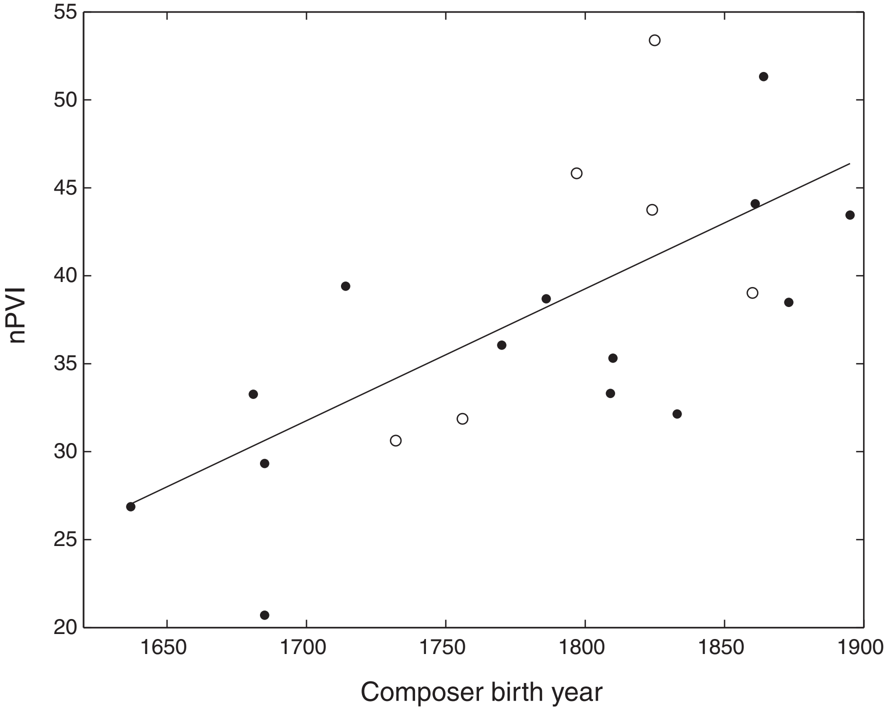

Figure 3.18 nPVI as a function of composer birth year for 20 composers. (Solid dots = German composers; open dots = Austrian composers.) The best-fitting linear regression line is shown. From Patel & Daniele, 2003b.

that in studying linguistic influences on speech rhythm, it is important to keep in mind historical influences that can run counter to linguistic influences.[28]

Taking a step back, musical nPVI research demonstrates that the rhythmic structure of speech and music can be fruitfully compared without any resort to notions of periodicity. It is worth noting that the research done so far hardly exhausts what can be done using this measure. For example, one could apply the nPVI to recordings of performed music rather than to music notation. One could also examine the nPVI of performances of the same piece of instrumental music by musicians who speak stress-timed versus syllable-timed languages, to see if the native language influences temporal patterns in music performance (cf. Ohgushi, 2002). Finally, one could study improvised music, for example, by studying jazz musicians who speak different dialects with different rhythmic qualities (e.g., in the United States, perhaps a northeast dialect versus a southern dialect). In this case, the nPVI could be used to investigate whether the temporal pattern of speech is reflected in the rhythm of improvised music.

3.5.2 Relations Between Nonlinguistic Rhythm Perception and Speech Rhythm

The idea that nonlinguistic rhythm perception can be influenced by one's native language has been articulated by both linguists and music researchers. Over 50 years ago, Jakobson, Fant, and Halle (1952:10–11) made the following claim:

> Interference by the language pattern affects even our responses to nonspeech sounds. Knocks produced at even intervals, with every third louder, are perceived as groups of three separated by a pause. The pause is usually claimed by a Czech to fall before the louder knock, by a Frenchman to fall after the louder; while a Pole hears the pause one knock after the louder. The different perceptions correspond exactly to the position of word stress in the languages involved: in Czech the stress is on the initial syllable, in French, on the final and in Polish, on the penult.

The groupings suggested by Jakobson et al. can be schematically represented as follows, in which each x represent a knock and the upper case X's are louder:

(3.18) X x x X x x X . . . = (X x x) Czech
= (x x X) French
= (x X x) Polish

[28] Unlike German music, English and French music do not show a significant increase in nPVI over the equivalent time period, based on themes in Barlow and Morgenstern's dictionary (Greig, 2003). This raises an interesting musicological puzzle: Why do German and Austrian music show such a strong historical change in this measure of rhythm, whereas English and French music do not?

The claim of Jakobson et al. is certainly provocative, but there has been no empirical evidence to support it. Nevertheless, the idea of a link between native language and nonlinguistic rhythm perception persists. For example, Stobart and Cross (2000) have documented a form of music from the Viacha people of the Bolivian highlands in which the local manner of marking the beat is different from what most English-speaking listeners perceive. Sound Example 3.9 illustrates this music with an Easter song played on a small guitar (*charango*). The position at which the Viacha clap or tap their foot to the beat can be heard at the end of the excerpt. This tendency to mark the shorter event in each group of two notes as the beat is contrary to the tendency of English speakers to hear the pattern iambically, that is, with the beat on the second event of each pair. Stobart and Cross speculate that the tendency to mark the beat trochaically (on the first member of each group) is related to stress patterns in words of the local language, Quechua.

The two examples above differ in that the former concerns segmentation (rhythmic grouping), whereas the latter concerns beat perception. Note, however, that neither refers to notions of periodicity in speech. Instead, they both refer to patterns of lexical stress and how this influences nonlinguistic auditory perception. Thus once again we see that interesting claims about rhythmic relations between music and language can be made without any reference to periodicity in speech.

How can one assess whether the native language influences the perception of nonlinguistic rhythm? As a first step, it is necessary to demonstrate that there are cultural differences in nonlinguistic rhythm perception. Rhythmic segmentation or grouping is of particular interest in this regard, as intimated by Jakobson et al. (1952). This is because psycholinguistic research indicates that that the rhythm of one's native language leads to segmentation strategies that are applied even when listening to a foreign language. The idea that the native language can influence nonlinguistic rhythmic segmentation is thus just one step away from this idea (cf. section 3.3.3, subsection "The Role of Rhythm in Segmenting Connected Speech").

Yet at the current time, it is widely believed that elementary grouping operations reflect general auditory biases not influenced by culture. This belief stems from a century-old line of research in which researchers have investigated rhythmic grouping using simple tone sequences (Bolton, 1894; Woodrow, 1909). For example, listeners are presented with tones that alternate in loudness (. . . loud-soft-loud-soft . . .) or duration (. . . long-short-long-short . . .) and are asked to indicate their perceived grouping. Two principles established a century ago, and confirmed in numerous studies since, are widely accepted:

1. A louder sound tends to mark the beginning of a group.
2. A lengthened sound tends to mark the end of a group.

These principles have come to be viewed as universal laws of perception, underlying the rhythms of both speech and music (Hayes, 1995b; Hay & Diehl, 2007). However, the cross-cultural data have come from a limited range of cultures (American, Dutch, and French). Are the principles truly universal? A study by Kusumoto and Moreton (1997) suggested otherwise, finding that American versus Japanese listeners differed with regard to Principle 2 above. This study motivated a replication and extension of this work by Iversen, Patel, and Ohgushi (2008), described below.

Iversen et al. had native speakers of Japanese and native speakers of American English listen to sequences of tones. The tones alternated in loudness ("amplitude" sequences, Sound Example 3.10a) or in duration ("duration" sequences, Sound Example 3.10b), as shown schematically in Figure 3.19.

Listeners told the experimenters how they perceived the grouping. The results revealed that Japanese and English speakers agreed with principle 1): both reported that they heard repeating loud-soft groups. However, the listeners showed a sharp difference when it came to principle 2.) Although English speakers perceived the "universal" short-long grouping, many Japanese listeners strongly perceived the opposite pattern, in other words, repeating long-short groups. (cf. Figure 3.19). Because this finding was surprising and contradicted a "law" of perception, Iversen et al. replicated it with listeners from different parts of Japan. The finding is robust and calls for an explanation. Why would native English and Japanese speakers differ in this way?

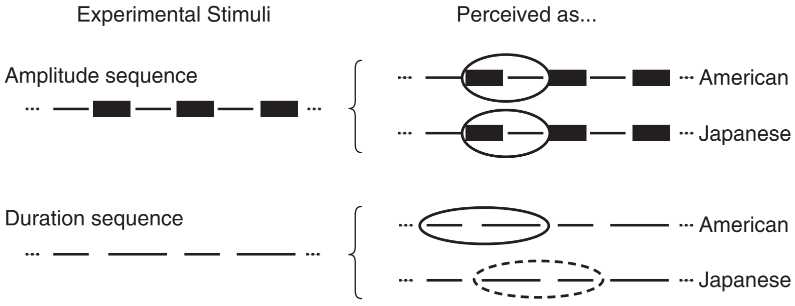

Figure 3.19 *Left side:* Schematic of sound sequences used in the perception experiment. These sequences consist of tones alternating in loudness ("amplitude sequence," top), or duration ("duration sequence," bottom). In the amplitude sequence, thin bars correspond to softer sounds and thick bars correspond to louder sounds. In the duration sequence, short bars correspond to briefer sounds and long bars correspond to longer sounds. The dots before and after the sequences indicate that only an excerpt of a longer sequence of alternating tones is shown. *Right side:* Perceived rhythmic grouping by American and Japanese listeners, indicated by ovals. Solid black ovals indicate preferences that follow "universal" principles of perception, while the dashed black oval indicates a preference that violates the purported universals.

Assuming that that these different perceptual biases are not innate, they key question is what aspect of auditory experience might be responsible for this difference. Two obvious candidates are music and speech, because these sound patterns surround humans throughout their life. Both patterns present the ear with sequences of sound that must be broken into smaller coherent chunks, such as phrases in music, or phrases and words in speech. Might the temporal rhythm of these chunks differ for music or speech in the two cultures? That is, might short-long patterns be more common in American music or speech, and long-short be more common in Japanese music or speech? If so, then learning these patterns might influence auditory segmentation generally, and explain the differences we observe.

Focusing first on music, one relevant issue concerns the rhythm of how musical phrases begin in the two cultures. For example, if most phrases in American music start with a short-long pattern (e.g., a "pick-up note"), and most phrases in Japanese music start with a long-short pattern, then listeners might learn to use these patterns as segmentation cues. To test this idea, we examined phrases in American and Japanese children's songs (because we believe these perceptual biases are probably laid down early in life). We examined 50 songs per culture, and for each phrase we computed the duration ratio of the first to the second note and then counted how often phrases started with a short-long pattern versus other possible patterns (e.g., long-short, or equal duration). We found that American songs show no bias to start phrases with a short-long pattern. Interestingly, Japanese songs show a bias to start phrases with a long-short pattern, consistent with our perceptual findings. However, the musical data alone cannot explain the cultural differences we observe, because this data cannot explain the short-long grouping bias of American listeners.

Turning to language, one basic difference between English and Japanese concerns word order (Baker, 2001). For example, in English, short grammatical (or "function") words such as "the," "a," "to," and so forth, come at the beginning of phrases and combine with longer meaningful (or "content") words (such as a noun or verb). Function words are typically "reduced," having short duration and low stress. This creates frequent linguistic chunks that start with a short element and end with a long one, such as "the dog," "to eat," "a big desk," and so forth. This fact about English has long been exploited by poets in creating the English language's most common verse form, iambic pentameter.

Japanese, in contrast, places function words at the ends of phrases. Common function words in Japanese include "case markers," short sounds that can indicate whether a noun is a subject, direct object, indirect object, and so forth. For example, in the sentence "John-san-ga Mari-san-ni hon-wo age-mashita," ("John gave a book to Mari") the suffixes "ga," "ni," and "wo" are case markers indicating that John is the subject, Mari is the indirect object and "hon" (book) is the direct object. Placing function words at the ends of phrases creates frequent chunks that start with a long element and end with a short one, which is just the opposite of the rhythm of short phrases in English (cf. Morgan et al., 1987).

Apart from short phrases, the other short meaningful chunks in language are words. Because our perception experiment focused on two-element groups, we examined the temporal shape of common disyllabic words in English and Japanese. English disyllabic words tend to be stressed on the first syllable (e.g., MO-ney, MAY-be; Cutler & Carter, 1987), which might lead one to think that they would have a long-short rhythmic pattern of syllable duration. To test this, we examined syllable duration patterns for the 50 most common disyllabic words in the language (from a corpus of spontaneous speech), and measured the relative duration of the two syllables. Surprisingly, common words with stress on the first syllable did not have a strong bias toward a long-short duration pattern. In contrast, common words with stress on the second syllable, such as "a-BOUT," "be-CAUSE," and "be-FORE," had a very strong short-long duration pattern. Thus the average duration pattern for common two-syllable words in English was short-long (Figure 3.20).

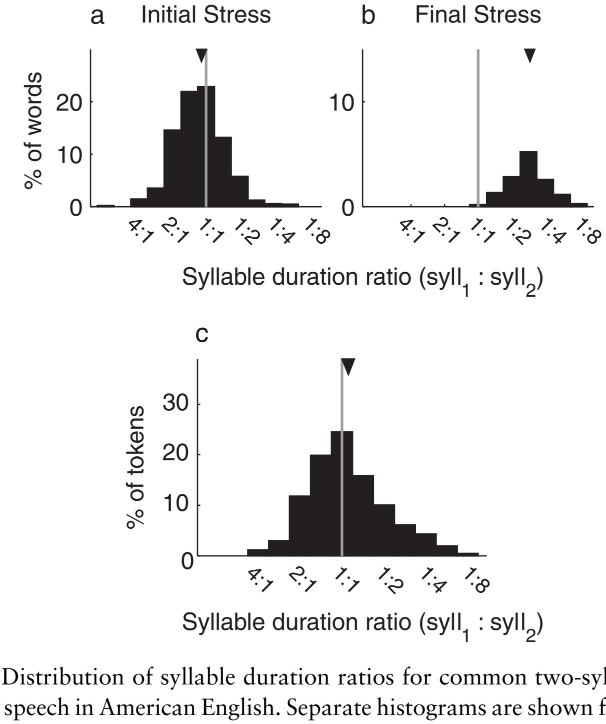

Figure 3.20 Distribution of syllable duration ratios for common two-syllable words in spontaneous speech in American English. Separate histograms are shown for initial-stress versus final-stress words in (a) and (b), and combined data are shown in (c), weighted by word frequency. Averages indicated by arrowheads. The overall distribution in (c) has a significant short-long bias (average ratio = 1 : 1.11).

This means that a short-long rhythm pattern is reflected at both the level of small phrases and common disyllabic words in English. We also examined syllable duration patterns in the 50 most common disyllabic words in Japanese. In contrast to English, the average duration pattern for such words was long-short. Thus once again, linguistic rhythm mirrored the results of the perception experiment.

Taking a step back, our results show that the perception of rhythmic grouping, long thought to follow universal principles, actually varies by culture. Our explanation for this difference is based on the rhythms of speech. Specifically, we suspect that learning the typical rhythmic shape of phrases and words in the native language has a deep effect on rhythm perception in general. If our idea is correct, then rhythmic grouping preferences should be predictable from the temporal structure of small linguistic chunks (phrases and words) in a language.

These findings highlight the need for cross-cultural work when it comes to testing general principles of auditory perception. Much of the original work on rhythmic grouping of tones was done with speakers of Western European languages (e.g., English, Dutch, and French). Although these languages do indeed have important differences, they all follow the pattern of putting short function words at the onset of small linguistic phrases, which may account for the similarity of perceptual grouping in these cultures. A more global perspective reveals that languages with phrase-final short function words are widespread, but exist largely outside of Europe, for example, in India and East Asia (Haspelmath et al., 2005). We predict that native speakers of these languages will group tones of alternating duration like Japanese listeners do (long-short).

An important future direction for this work concerns the development of rhythmic grouping preferences in childhood. Do infants have an innate bias for a particular grouping pattern (e.g., short-long), which is then modified by experience (cf. Trainor & Adams, 2000)? Or are they rhythmic "blank slates"? Regarding the perception of rhythm by adults, if speakers of different languages perceive nonlinguistic rhythm differently, this could help explain reports of differences between Westerners and Japanese in the performance of simple musical rhythms (Ohgushi, 2002; Sadakata et al., 2004). That is, simple rhythms may be performed differently in different cultures because they are perceived differently during learning. This would indicate that experience with speech shapes nonlinguistic rhythm cognition at a very basic level.

3.5.3 Neural Relationships Between Rhythm in Speech and Music

In this chapter, I have claimed that certain aspects of speech rhythm and musical rhythm show a striking similarity, such as the grouping of events into phrases, whereas other aspects are fundamentally different, such as the role of temporal periodicity. To what extent do neural data support this claim? Is there evidence that some aspects of rhythm in speech and music are handled by similar brain systems, whereas other aspects show little neural overlap?

Focusing first on grouping, there is evidence for overlap in brain processing of phrase boundaries in both domains. This evidence comes from electrical brain responses (event-related potentials, ERPs) in normal individuals. Steinhauer et al. (1999) demonstrated that the perception of phrase boundaries in language is associated with a particular ERP component termed the "closure positive shift" (CPS), a centro-parietal positivity of a few hundred milliseconds that starts soon after the end of an intonational phrase. Further studies using filtered or hummed speech (to remove lexical cues and leave prosodic cues) showed that the CPS is sensitive to prosodic rather than syntactic cues to phrase boundaries (Steinhauer & Friederici, 2001; Pannekamp et al., 2005). Inspired by this work, Knösche et al. (2005) examined the ERPs in musicians to the ends of musical phrases, and found a component similar to the CPS reported by Steinhauer et al. Using MEG, they also identified brain areas that were likely to be involved in the generation of the CPS in music. These areas included the anterior and posterior cingulate cortex and the posterior hippocampus. Based on the roles these areas play in attention and memory, the researchers argue that the musical CPS does not reflect the detection of a phrase boundary per se, but memory and attention processes associated with shifting focus from one phrase to the next.

The studies of Steinhauer et al. and Knösche et al. point the way to comparative neural studies of grouping in language and music. There is much room for further work, however. For example, in the Knösche et al. study the sequences with phrase boundaries have internal pauses, whereas the sequences without phrase boundaries do not. It would be preferable to compare sequences with and without phrase boundaries but with identical temporal structure, for example, using harmonic structure to indicate phrasing (cf. Tan et al., 1981). This way, ERPs associated with phrase boundaries cannot be attributed to simple temporal differences in the stimuli. It would also be desirable to conduct a within-subjects study of brain responses to phrases in language and music. Such comparative work should attend to the absolute duration of musical versus linguistic phrases, as the neural processes involved in grouping may be influenced by the size of the temporal unit over which information is integrated (Elbert et al., 1991; von Steinbüchel, 1998).

Turning to the question of periodicity, if speech rhythms and periodic musical rhythms are served by different neural mechanisms, then one would predict neural dissociations between linguistic rhythmic ability and the ability to keep or follow a beat in music. The neuropsychological literature contains descriptions of individuals with musical rhythmic disturbance after brain damage, or "acquired arrhythmia" (e.g., Mavlov, 1980; Fries & Swihart, 1990; Peretz, 1990; Liégeois-Chauvel et al., 1998; Schuppert et al., 2000; Wilson et al., 2002; Di Pietro et al., 2003). Two notable findings from this literature are that rhythmic abilities can be selectively disrupted, leaving pitch processing skills relatively intact, and that there are dissociations between rhythmic tasks requiring simple discrimination of temporal patterns and those requiring the evaluation or production of periodic patterns (e.g., Peretz, 1990). For example, Liégeois-Chauvel et al. (1998) found that patients with lesions in the anterior part of the left or right superior temporal gyrus were much more impaired on a metrical task than on a temporal discrimination task. The metrical task involved identifying a passage as a waltz or a march, whereas the temporal discrimination task involved a same different judgment on short melodic sequences that differed only in terms of their duration pattern. In the metrical task, patients were encouraged to tap along with the perceived beat of the music to help them in their decision. Wilson et al. (2002) describe a case study of a musician with a right temporoparietal stroke who could discriminate nonmetrical rhythms but who could not discriminate metrical patterns or produce a steady pulse.

Unfortunately, none of these studies explicitly set out to compare rhythmic abilities in speech and music. Thus the field is wide open for comparative studies that employ quantitative measures of both speech and musical rhythm after brain damage. It would be particularly interesting to study individuals who were known to have good musical rhythmic abilities and normal speech before brain damage, and to examine whether disruptions of speech rhythm are associated with impaired temporal pattern discrimination, impaired metrical abilities, or both.

Another population of individuals who would be interesting to study with regard to speech and musical rhythm are individuals with "foreign accent syndrome" (Takayama et al., 1993). In this rare disorder, brain damage results in changes in speech prosody that give the impression that the speaker has acquired a foreign accent. It remains to be determined if this disorder is associated with systematic changes in speech rhythm, but if so, one could examine if such individuals have any abnormalities in their musical rhythmic skills.

Of course, a difficulty in studying acquired arrhythmia and foreign accent syndrome is that such cases are quite rare. Thus it would be preferable to find larger populations in which either speech rhythm or musical rhythmic abilities were impaired, in order to conduct comparative research. One population that holds promise for comparative studies are tone-deaf or "congenital amusic" individuals who have severe difficulties with music perception and production which cannot be attributed to hearing loss, lack of exposure to music, or any obvious nonmusical social/cognitive impairments (Ayotte et al., 2002). One advantage of working with such individuals is that they can easily be found in any large community through a process of advertising and careful screening (Ayotte et al., 2002; Foxton et al., 2004). Such individuals appear to have problems with basic aspects of pitch processing, such as discriminating small pitch changes or determining the direction of small pitch changes (i.e., whether pitch goes up or down) (Peretz & Hyde, 2003; Foxton et al., 2004). Interestingly, they do not seem to be impaired in discriminating simple temporal patterns and can synchronize successfully to a simple metronome. However, they do have difficulty synchronizing to the beat of music (Dalla Bella & Peretz, 2003). Of course, it could be that the difficulty in synchronizing with music is simply due to the distraction caused by a stimulus with pitch variation, due to deficits in pitch processing (cf. Foxton et al., 2006). Thus future studies of beat perception in congenital amusia should use complex rhythmic sequences with no pitch variation, such as those used in the study of Patel, Iversen, et al. (2005) described in section 3.2.1 above (cf. Sound Examples 3.3. and 3.4). If musically tone-deaf individuals cannot synchronize to the beat of such sequences, this would suggest that the mechanisms involved in keeping a beat in music have nothing to do with speech rhythm (because the speech of musically tone-deaf individuals sounds perfectly normal).[29]

I suspect that future research will reveal little relationship between speech rhythm abilities in either production or perception and musical rhythm abilities involving periodicity (such as metrical discrimination or beat perception and synchronization). This would support the point that periodicity does not play a role in speech rhythm.

3.6 Conclusion

Speech and music involve the systematic temporal, accentual, and phrasal patterning of sound. That is, both are rhythmic, and their rhythms show both important similarities and differences. One similarity is grouping structure: In both domains, elements (such as tones and words) are grouped into higher

[29] Of course, before any firm conclusions can be drawn, the speech rhythm of tone-deaf individuals would need to be quantitatively measured to show that it did not differ from normal controls (for example, using the nPVI). Also, it is possible that tone-deaf individuals cannot keep a beat because their pitch perception problem has caused an aversion to music, so that they have not had enough exposure to music to learn how to keep a beat. Thus it would be preferable to work with individuals who have normal pitch perception and who enjoy music, but who cannot keep a beat. The existence of such "rhythm-deaf" individuals is intuitively plausible, as there are certainly people who like music but claim to have "two left feet" when it comes to dancing, and/or who cannot clap along with a beat. It should be possible to find a population of such people through a process of advertising and screening, akin to the procedures used to find congenital amusics.

level units such as phrases. A key difference is temporal periodicity, which is widespread in musical rhythm but lacking in speech rhythm. Ironically, the idea that speech has periodic temporal structure drove much of the early research on speech rhythm, and was the basis for a rhythmic typology of languages which persists today (stress-timed vs. syllable-timed languages). It is quite evident, however, that the notion of isochrony in speech is not empirically supported. Fortunately, much recent empirical research on speech rhythm has abandoned the notion of isochrony, and is moving toward a richer notion of speech rhythm based on how languages differ in the temporal patterning of vowels, consonants, and syllables. A key idea that motivates this research is that linguistic rhythm is the product of a variety of interacting phonological phenomena, and not an organizing principle, unlike the case of music.

It may seem that breaking the "periodicity link" between speech and music would diminish the chance of finding interesting rhythmic relations between the domains. In fact, the converse is true. Changing the focus of comparative work from periodic to nonperiodic aspects of rhythm reveals numerous interesting connections between the domains, such as the reflection of speech timing patterns in music, and the influence of speech rhythms on nonlinguistic rhythmic grouping preferences. Although many more connections await exploration, it seems clear that some of the key processes that extract rhythmic structure from complex acoustic signals are shared by music and language.

Appendix 1: The nPVI Equation

This is an appendix for section 3.3.1, subsection "Duration and Typology."

The nPVI equation is:

$$\mathrm{nPVI} = 100/(m-1) \times \sum_{k=1}^{m-1} \left| (d_k - d_{k+1}) / ((d_k + d_{k+1})/2) \right|$$

In this equation, m is the number of durations in the sequence (e.g., vowel durations in a sentence) and d_k is the duration of the kth element. The nPVI computes the absolute value of the difference between each successive pair of durations in a sequence, normalized by the mean of these two durations (this normalization was originally introduced to control for fluctuations in speech rate). This converts a sequence of m durations to a sequence of $m - 1$ contrastiveness scores. Each of these scores ranges between 0 (when the two durations are identical) and 2 (for maximum durational contrast, i.e., when one of the durations approaches zero). The mean of these scores, multiplied by 100, yields the nPVI of the sequence. The nPVI value for a sequence is thus bounded by lower and upper limits of 0 and 200, with higher numbers indicating a greater degree of durational contrast between neighboring elements.

Appendix 2: Musical nPVI Values of Different Nations

This is an appendix for Chapter 3, section 3.5.1, Table 3.1. Data kindly provided by David Huron.

In the tables below, ♯ C = number of composers, *sd* = standard deviation.

Nationality	Mean	♯ Themes	♯ C	*sd*
American	46.7	478	32	22.4
Armenian	43.1	33	1	22.2
Austrian	45.1	1,636	22	23.7
Austro-Hung	45.9	14	1	18.7
Belgian	48.7	41	6	19.4
Bohemian	44.4	30	1	22.9
Brazilian	41.5	5	1	18.3
Catalan	48.9	12	1	16.8
Cuban	36.0	17	2	17.9
Czech	46.9	266	6	24.3
Danish	51.0	7	1	24.6
English	45.6	741	27	24.4
Finnish	44.9	169	2	25.3
Flemish	25.2	3	1	3.8
French	43.4	1,343	52	25.2
German	43.2	2,379	39	26.8
Hungarian	45.4	244	8	25.4
Irish	44.1	16	3	25.7
Italian	41.4	572	46	23.9
Mexican	28.4	13	2	19.3
Norwegian	45.2	122	2	20.3
Polish	50.2	60	8	18.7
Romanian	42.4	26	2	21.3
Russian	41.3	853	25	23.4
Spanish	42.5	108	8	19.3
Swedish	50.0	12	3	29.3

Data from the above table regrouped by language: English = American, English, Irish; French = French, Belgian; German = German, Austrian, Austro-Hungarian; Slavic = Russian, Czech, Polish, Bohemian; Spanish = Spanish, Catalan, Cuban, Mexican; Scandinavian = Danish, Norwegian, Swedish (not Finnish)

Language	**Mean**	**♯ Themes**	**♯ C**	***sd***
English	46.0	1,235	62	23.6
French	43.6	1,384	58	25.0
German	44.0	4,029	62	25.6
Slavic	43.1	1,209	40	23.5
Spanish	41.0	150	13	19.4
Scandinavian	45.9	141	6	21.3

This page intentionally left blank

Chapter 4

Melody