Perception of rhythmic grouping depends on auditory experience[a)]

John R. Iversen[b)] and Aniruddh D. Patel
The Neurosciences Institute, 10640 John Jay Hopkins Drive, San Diego, California 92121

Kengo Ohgushi
Kyoto City University of Arts, Kyoto, 610–1197 Japan

(Received 22 June 2007; revised 11 July 2008; accepted 17 July 2008)

Many aspects of perception are known to be shaped by experience, but others are thought to be innate universal properties of the brain. A specific example comes from rhythm perception, where one of the fundamental perceptual operations is the grouping of successive events into higher-level patterns, an operation critical to the perception of language and music. Grouping has long been thought to be governed by innate perceptual principles established a century ago. The current work demonstrates instead that grouping can be strongly dependent on culture. Native English and Japanese speakers were tested for their perception of grouping of simple rhythmic sequences of tones. Members of the two cultures showed different patterns of perceptual grouping, demonstrating that these basic auditory processes are not universal but are shaped by experience. It is suggested that the observed perceptual differences reflect the rhythms of the two languages, and that native language can exert an influence on general auditory perception at a basic level.
 [DOI: 10.1121/1.2973189]

PACS number(s): 43.66.Mk, 43.70.Kv, 43.75.Cd [MSS] Pages: 2263–2271

I. INTRODUCTION

A. Aims

The perception of rhythm is central to how we find structure and meaning in speech and music (Lerdahl and Jackendoff, 1983). In perceiving rhythm, we naturally perceive events as *grouped* into higher-level patterns. Such grouping is an essential step in the interpretation of complex sound sequences. In language, for example, listeners must segment words and phrases from the ongoing speech stream in order to make sense of the incoming signal. In music, comprehending melodic structure involves segmenting tone sequences into smaller coherent chunks (e.g., motives and phrases) in order to discern larger patterns. The human proclivity for auditory grouping is so strong that it is even applied to sequences of physically identical sounds, as, for example, when an electronic metronome is heard as "tick tock" when, in fact, each sound is the same (Bolton, 1894). The fundamental question behind the present work is whether grouping is an innate building block of perception or if instead it is learned from the environment.

The weight of evidence to date has supported the former view, placing grouping together with other basic auditory perceptual mechanisms thought to operate in a bottom-up fashion, reflecting, for example, innate mechanisms that evolved to help animals sort sounds into distinct sources (Bregman, 1990). Two principles, established over a century ago (Bolton, 1894; Woodrow, 1909), are widely accepted and thought to universally govern the subjective grouping of simple sound sequence:

(1) A louder sound tends to mark the beginning of a group;
(2) a lengthened sound (or interval between sounds) tends to mark the end of a group.

These principles have been confirmed in numerous studies since their first description (Vos, 1977; Trainor and Adams, 2000; Hay and Diehl, 2007). Principle 2 is observed also in infants (Trainor and Adams, 2000). Similarities of these principles to Gestalt principles underlying the perception of visual patterns are often noted as additional support for their universality (Wertheimer, 1938; Fraisse, 1982).

The importance of these principles extends beyond the perception of artificial tone sequences and has bearing on language perception and learning. These innate grouping preferences have been proposed to affect aspects of language learning, such as the learning of language segmentation (e.g., Trehub and Trainor, 1993), and to account for the minimal set of rhythmic units in the world's languages (the "iambic-trochaic law;" Hayes, 1995). The principles have recently been shown to apply to speechlike sequences of phonemes in native listeners of English and French, languages with distinct speech rhythms, suggesting universality and relevance to language processing (Hay and Diehl, 2007; Dauer, 1983; Ramus *et al.*, 1999).

The question of whether such grouping principles might instead be determined by experience, which was raised long ago (Jakobson *et al.*, 1952), has only recently been raised again (Drake and Bertrand, 2001) and has received little ex-

[a)]Portions of this work were presented in "Perception of Nonlinguistic Rhythmic Stimuli by American and Japanese Listeners," Proceedings of the International Congress of Acoustics, Kyoto, Japan, April 2004 and "Nonlinguistic rhythm perception depends on culture and reflects the rhythms of speech: Evidence from English and Japanese," Fourth ASA/ASJ Joint Meeting, Honolulu, November, 2006.

[b)]Electronic mail: iversen@nsi.edu

perimental examination. Interestingly, the earliest proposal that nonlinguistic grouping might be shaped by experience had to do with language. In their foundational text Preliminaries to Speech Analysis (1952), Jakobson *et al.* made the following claim: "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.* (1952) can be schematically represented as follows, where each x represents a knock and the upper case X's are louder

$$\begin{aligned} \ldots \text{X x x X x x X} \ldots &= (\text{X x x}) \quad \text{Czech} \\ &= (\text{x x X}) \quad \text{French} \\ &= (\text{x X x}) \quad \text{Polish}. \end{aligned}$$

While this claim is provocative, there has been no empirical evidence to support it. Bell (1977) attempted to test the claim by examining rhythmic grouping perception in tone sequences with every third tone differing in intensity, pitch, or duration. He found no significant differences in grouping perception among native speakers of English, Bengali, French, Persian, and Polish. However, the sample sizes in this study were small, and there were confounds in the stimuli, which made the results hard to interpret. In contrast to the proposal of Jakobson *et al.* (1952), further studies have since confirmed the universality of grouping principles for speakers of English, French, and Dutch (Hay and Diehl, 2007; Vos, 1997).

There are several reasons, however, to suggest that experience could play a role in shaping the perception of rhythmic grouping. The first, and most direct, is an empirical discovery made by Kusumoto and Moreton (1997), who found a difference between American and Japanese listeners with regard to rhythmic grouping in simple tone sequences. In a study conducted in the United States, they presented listeners with sequences of tones alternating in amplitude or duration. Members of both cultures perceived sequences with alternating loud and soft tones as repeating "loud-soft" groups of two elements. In contrast, a cultural difference was found for sequences with tones that alternated in duration: Americans showed a strong bias for hearing a repeating "short-long" group of two elements, while Japanese speakers did not. In fact, many Japanese reported strongly hearing the opposite "long-short" pattern. This challenged the idea that grouping perception follows universal principles and motivated the current study's focus on rhythm perception by English versus Japanese speakers.

Several additional findings provide reasons to suspect that language experience could shape nonlinguistic rhythm perception. The first concerns psycholinguistic research on segmenting words from connected speech. Such research has indicated that experience with a language's rhythm leaves a permanent influence on a listener in terms of rhythmic segmentation strategies. That is, segmentation strategies (such as the "metrical segmentation strategy" in English, whereby a native listener posits a word boundary at every stressed syllable) are a perceptual habit of a listener and are even applied (inappropriately) to languages with other rhythmic patterns (see Cutler, 2000 for a review). This raises the possibility that strategies developed for segmenting meaningful linguistic units from connected speech could carry over and also influence how one segments nonlinguistic rhythmic patterns.

The second concerns evidence for connections between linguistic and nonlinguistic rhythmic patterning. Patel and Daniele (2003) found that timing patterns in British English and continental French speech are reflected in the classical instrumental music of these two cultures (see also Huron and Ollen, 2003). While that finding concerned rhythmic production rather than perception (and timing rather than grouping), the relevant point concerns the mechanism proposed to account for this cross-domain similarity. This mechanism was statistical learning of the rhythms of the native language, i.e., tracking rhythmic patterns in language and acquiring implicit knowledge of their statistical properties, without any direct feedback (see Patel *et al.*, 2006). It was proposed that this implicit knowledge could influence the creation of rhythmic patterns in a nonlinguistic domain (music). The current study broadens this idea to consider whether statistical learning of the rhythms of speech might be responsible for shaping basic auditory grouping biases, which reveal themselves in the perception of tone sequences that are simpler than either speech or music. Thus the issue at hand is whether learning the characteristic rhythms of meaningful units in the auditory environment (which is dominated by speech for humans) can shape low-level rhythm perception mechanisms.

Finally, infants show a remarkable sensitivity to the rhythms of their native language and music (Mehler *et al.*, 1996; Nazzi *et al.*, 1998; Hannon and Trehub, 2005), suggesting that mechanisms for learning rhythmic regularities from language and music exist from birth (Mehler *et al.*, 1996; Nazzi *et al.*, 1998).

Thus, there are several reasons to revisit the question of how experience shapes low-level perception of rhythm. Recently, the issue of top-down influences on basic auditory processing has received renewed attention (see Davis and Johnsrude, 2007). To date, however, only a few studies have addressed how language might influence basic auditory processes, and these studies have focused on the perception of pitch (i.e., Bent *et al.* 2006; see Deutsch, 1991). The current study complements these studies by its focus on rhythm.

B. Overview

In the present work, perceptual grouping is studied in Japanese and English speakers using simple tone sequences in which sounds alternate in a single parameter (amplitude or duration). The use of simple tone sequences follows the tradition in auditory research that first established the grouping principles under study (Bolton, 1894; Woodrow, 1909). In contrast to previous work (which did not test non-Western listeners), the present study finds that even simple grouping

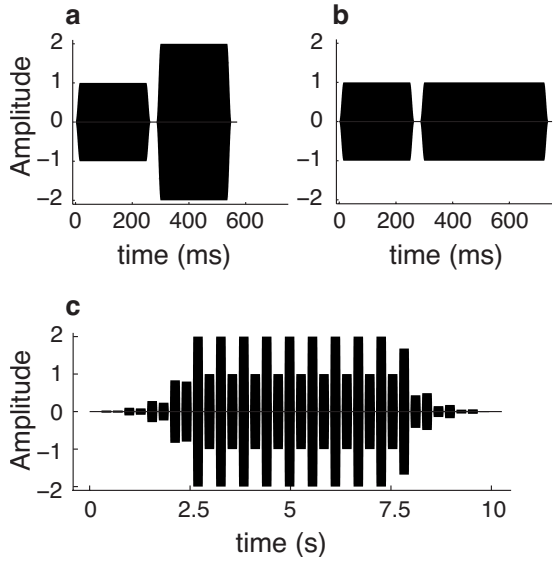

FIG. 1. Amplitude and duration stimulus sequences. Every second tone had a different amplitude or duration. Stimuli consisted of a basic tone (500 Hz, complex; 150 or 250 ms duration) alternating with a second tone of increased amplitude (1.5 or 2 times greater) or increased duration (1.25, 1.5, or 3 times greater). (a) Example of tone pair from an amplitude sequence (basic duration 250 ms, amplitude ratio=2). (b) Example of a tone pair from a duration sequence (basic duration 250 ms, duration ratio 1.5). (c) Example of a complete amplitude sequence. See text for additional stimulus details.

perception depends on culture: Americans and Japanese hear different grouping patterns in identical sequences of sound.

The finding of a cultural difference motivates the question of what the source of the difference might be. Under the assumption that the difference in grouping perception reflects auditory experience (rather than innate factors), this paper proposes the hypothesis that experience with the most common auditory input, speech, underlies this more general perceptual difference.

II. EXPERIMENT: PERCEPTION OF RHYTHMIC GROUPING

A. Introduction

To test the universality of grouping preferences, native speakers of English or Japanese were presented with repeating sequences of two alternating tones in which the second tone had either increased amplitude or duration relative to the first tone, at one of several fixed ratios (Fig. 1). Listeners were asked to report their perceived grouping of the stimulus (e.g., long-short or short-long), and also to indicate how confident they were of their judgment. If grouping principles (1) and (2) above (Sec. I A) are indeed universal, one would predict that all listeners would perceive tones of alternating amplitude as repeating loud-soft groups, with the higher amplitude tone beginning the group, and would perceive tones of alternating duration as short-long, with the longer duration tone ending the group.

As noted in Sec. I, similar research in the past has focused on listeners from Western European cultures and has found cross-cultural agreement in terms of grouping preferences. Kusumoto and Moreton (1997) were the first to test individuals from an East Asian culture and found a cultural difference, motivating the current research. The present research extends these findings by using native Japanese speakers living in Japan and with no foreign living experience, by having larger sample sizes and by collecting confidence ratings on listeners' perceptual judgments. These changes are intended to increase the reliability of the results and enable meaningful individual-listener analysis in addition to group-level analyses.

B. Methods

Stimuli were 10 s sequences in which tones alternated in either amplitude (loud-soft) or duration (long-short) (henceforth "amplitude sequences" and "duration sequences;" see Fig. 1). The basic tone was a 500 Hz complex tone (15 ms rise/fall) with a duration of either 150 or 250 ms. The complex tone was a low-pass square wave consisting of the fundamental and first three odd harmonics and was constructed to match previous work on grouping (Kusumoto and Moreton, 1997; Trainor and Adams, 2000). The alternating tone was constructed by multiplying either the amplitude or duration of the basic tone by one of several small ratios (amplitude: 1.5 or 2; duration: 1.25, 1.75, or 3). The two tones alternated with a gap of 20 ms between them. To mask possible effects of starting tone order, stimuli were slowly faded in (and out) over 2.5 s (double-logarithmic ramp), and each sequence was presented both forward and reversed, yielding a total of 20 sequences (2 base durations $\times$ 2 orders of presentation [forward and reversed] $\times$ 5 ratio parameters [2 amplitude ratios + 3 duration ratios], hence four different sequences per ratio). Stimuli were generated in MATLAB, at CD quality (44.1 kHz sample rate) and were presented in free field in a classroom setting. Sound levels were not measured, but were verified to be easily audible to all participants (see EPAPS).

Listeners were familiarized with the experiment by hearing one example sequence of the two types (but with different parameters than those used in the experiment). The experimental sequences were then presented in random order, and listeners were asked to indicate their perceived grouping by circling pairs of tones on diagrams schematically depicting the stimuli, in answer booklets with one page per sequence. The starting tone (e.g., long or short) of diagrams in the booklets was counterbalanced to avoid any possible visual bias due to the diagrams. Listeners also rated the certainty of each judgment, ranging from 3 (completely certain) to 1 (guessing). The experiment was conducted in a classroom setting, enabling data to be collected in parallel for multiple participants: 43 native American English speakers) (in San Diego; henceforth referred as "English speakers") and 46 native speakers of Japanese (in Kyoto; all were speakers of the Kansai dialect of Japanese; three participants were excluded because they had lived abroad for more than six months). Participants were college aged 17–25 years. Musical training (years studied) was similar in the two groups (English: 5.0 ± 4.7 years, Japanese: 7.0 ± 5.8 years; n.s., $p=0.09$, unpaired t-test). Although study of English is part of compulsory education in Japan, none of the native Japanese speakers rated themselves as "highly proficient"

speakers of English. None of the native speakers of English rated themselves as proficient in Japanese. All instructions, both verbal and written on the answer booklets, were given in the native language. To test the robustness of our findings for Japanese listeners, replication studies were conducted in an identical manner in different regions of Japan (Niigata, $n=38$; Tokyo, $n=54$).

To analyze the behavioral data, each participant's grouping response and confidence rating for each of the 20 stimuli was tabulated. Only the completely certain (level 3) responses (55% of all responses) were analyzed further. [Responses with intermediate certainty (31% of all responses) showed the same pattern of results as the high-certainty responses, but with slightly weaker preference, while those marked as guesses (14%) showed no preference.] Confidence ratings showed no evidence of differences by culture: 91% of Japanese listeners had at least one confident response to the duration task, compared with 87% of English listeners. Listeners varied in the number of high-certainty responses, but the mean number of high-certainty responses was not significantly different between cultures for either type of sequence [amplitude: 3.7 (s.d. 2.6) for English speakers and 3.8 (s.d. 2.6) for Japanese speakers (out of a possible 8); duration: 7.3 (s.d. 3.3) for English speakers and 5.7 (s.d. 3.6) for Japanese (out of a possible 12)].

For both amplitude and duration sequences, individual listener responses were pooled across the four stimulus sequences for each ratio, as neither basic tone duration nor sequence order had a significant effect on responses. For amplitude sequences, the percentage of each participant's high-certainty responses in which a soft-loud grouping was reported was calculated for each of the two amplitude ratios. Similarly, for duration sequences, the percentage of high-certainty short-long responses was found, for each of the three duration ratios. With individual average responses for each participant in hand, group means for each of the stimulus conditions were computed by averaging across participants.

In the analysis, all participants are weighted equally, independent of the number of high-confidence responses they made. To examine any potential bias due to weighting individuals with few high-confidence responses equally with those with many high-confidence responses, all analyses of the original studies were repeated including only participants who gave high-confidence responses to at least half of the stimuli (i.e., high-confidence responses to at least 4 of the 8 amplitude sequences or 6 of the 12 duration sequences). This excluded an average of 42% of participants from each language group (35% of English; 48% of Japanese). Results from this restricted analysis were statistically indistinguishable from those computed on the entire set of participants (all $p>0.38$, Mann-Whitney U), suggesting that including participants with fewer responses does not bias the results. Consequently, we used data from all participants.

C. Results

The perception of amplitude sequences was similar for all listeners, consistent with the universality of the first grouping principle: both Japanese and English listeners preferred a loud-soft grouping [Fig. 2(a)]. For both groups, the preference was highly significant (compared to no preference; Wilcoxon signed-rank test, pooled across ratio: English $p<0.001$; Japanese $p<0.0001$). As a group, the Japanese listeners had a significantly stronger preference for loud-soft grouping (91% of all responses; mean across ratios) than the English listeners [68% of all responses; two-way analysis of variance (ANOVA) on language and ratio; $F(1,139)=17.8$, $p<0.0001$]. There was no significant effect of amplitude ratio [$F(1,139)=0.69$, $p=0.41$). English and Japanese listeners had nearly identical distributions of confidence ratings, with 50% and 51% of responses receiving the highest confidence.

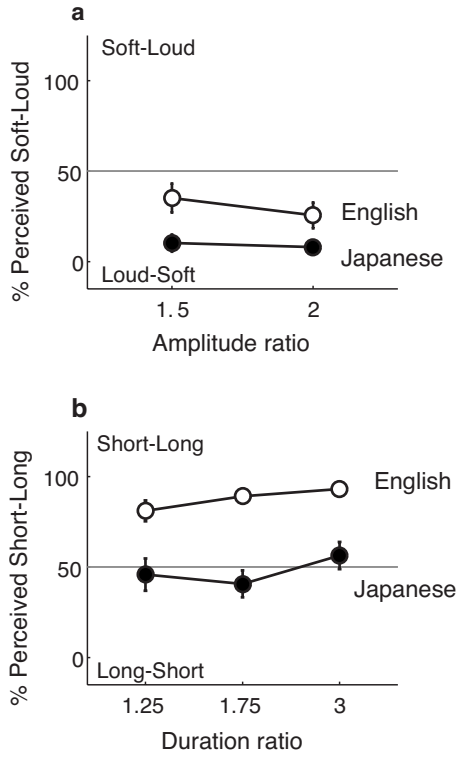

FIG. 2. Grouping preferences as a function of native language. (a) Amplitude sequence grouping preferences of English (open symbols, $n=43$) and Japanese (filled symbols, $n=43$) listeners. The across-subject mean ($\pm$ s.e.) percentage of sequences heard grouped as soft-loud is shown for two amplitude ratios. Both language groups strongly preferred a loud-soft grouping. (b) Duration sequence grouping preferences. There is a large difference between language groups. English listeners strongly preferred a short-long grouping preference, while Japanese listeners on average did not show a preference.

In contrast, a striking difference between English and Japanese listeners was found in the perception of duration sequences [Fig. 2(b)]. English listeners showed a strong preference for short-long grouping (89% of all responses; $p<0.0001$, Wilcoxon signed-rank test), as predicted by the grouping principles, but Japanese listeners as a group showed no preference ($p=0.39$). The differences between the language groups were highly significant [$F(1,219)=85.8$, $p<0.0001$]. On the strength of this result, then, the second grouping principle appears not to be universal.

Additional insight into the patterns of response may be gained by examining the distribution of individual participants' mean responses for the amplitude and duration stimuli

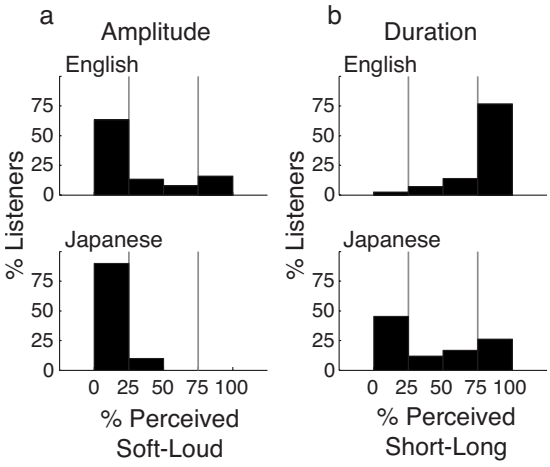

FIG. 3. Distribution of individual listeners' grouping preferences. (a) Distribution of amplitude sequence grouping preferences of English (top) and Japanese (bottom) listeners. The distribution shows each listener's percentage of all amplitude sequences heard as soft-loud. For both English and Japanese listeners, the majority preferred a loud-soft grouping (63% of English listeners; 90% of Japanese listeners), with most participants choosing this preference consistently for all stimuli. (b) Distribution of duration grouping preference for English (top) and Japanese (bottom) listeners. The distribution shows each listener's percentage of all duration sequences heard as short-long. The large majority of English listeners chose a short-long grouping (77%). In contrast, nearly half (45%) of the Japanese listeners chose the opposite long-short grouping, which English speakers almost never chose. Japanese responses were more varied, and 26% of Japanese listeners chose the short-long grouping. Thin vertical lines delineate regions defined as strong preference (0%–25% and 75%–100% preference).

pooled across ratio, base duration, and order of presentation [Figs. 3(a) and 3(b), respectively]. When data are pooled in this way, each participant simply responded to 8 amplitude sequences and 12 duration sequences. [Thus a participant who had high-confidence responses to 10 out of 12 duration sequences, and responded short-long in all such cases, would contribute a point to the 100% column of the histogram in Fig. 3(b)].

For amplitude sequences, the distributions for English [Fig. 3(a), top] and Japanese listeners [Fig. 3(a), bottom] are highly skewed toward a loud-soft grouping, as would have been expected from the group means shown in Fig. 2(a). The majority of participants in both groups had strong preferences for the loud-soft grouping, where strong preference is defined as a preference score in the extreme 1/4 of the response scale (e.g., strong preference for loud-soft is any response score $<=25\%$).

The distribution of individual participant responses to duration sequences is highly illuminating, because it shows that the majority of individual Japanese listeners do have strong grouping preferences [Fig. 3(b)]. Thus the Japanese group result of no preference seen in Fig. 2(b) is not representative of individual perception. Specifically, while the distribution for English listeners shows that a majority (77%) have strong preference for short-long grouping [Fig. 3(b), top] in accord with the group results, Japanese individual listeners, in contrast, had more widely distributed grouping preferences [Fig. 3(b), bottom]. The distribution of Japanese responses is highly significantly different from a uniform distribution (which would be the case if all Japanese had no preference and were responding randomly; Kolmogorov–Smirnov test, $p<0.005$, $df=41$). Instead, most Japanese listeners (71%) had strong preferences for either short-long or long-short grouping. Japanese preferences were not equally distributed, however, but were significantly biased toward long-short grouping ($p<0.001$, $df=41$, Wilcoxon signed-rank, null hypothesis of equal preferences for both groupings). Indeed, 45% of Japanese listeners had a strong preference for the long-short grouping, versus 26% for short-long grouping. The substantial proportion of listeners with a long-short grouping preference is striking given that only a single English listener (<2% of the population) expressed any perception of long-short grouping [Fig. 3(b), top panel].

Two replication studies on an additional 92 Japanese listeners living in different parts of Japan (Niigata and Tokyo) were conducted to test the consistency of the Japanese grouping results. This brings the number of Japanese participants to 135. The results of these studies were indistinguishable from those seen for Kyoto listeners (ANOVA on individual participants' mean response with test location as factor. [amplitude sequences: $F(2,115)=0.094$, $p=0.91$; duration sequences: $F(2,121)=1.42$, $p=0.25$]. This replication suggests that the Japanese result is reliable.

D. Discussion

With respect to the purported universality of grouping principles (Sec. I A), both Japanese and English listeners followed the first principle (grouping based on amplitude), but only English listeners followed the second principle (grouping based on duration). Japanese listeners showed variation across individuals in their duration-based grouping preferences, but the majority of individuals expressed a strong preference for one or the other grouping. A salient finding was that the most common Japanese preference was for long-short grouping, a choice made by virtually no English listeners. This result is incompatible with the universality of rhythmic grouping principles. The current findings confirm and extend those of Kusumoto and Moreton (1997). As they run counter to long-held claims for universality, two additional experiments in Japan were conducted, which confirmed the findings of the initial study. Hence the finding appears robust, leading to the question of how this cultural difference in perception arises.

Before turning to this question in the general discussion, it is worth addressing a possible methodological concern about the task we used to assess grouping. The current study used a metalinguistic task to probe grouping preferences (circling of images in answer booklets), raising the question of how the results might compare to tasks that use other perceptual methods to probe grouping. This question can be answered for English listeners because their duration-based grouping preferences have been explored by Trainor and Adams (2000) using a purely perceptual gap-detection paradigm. The task required listeners to detect a short gap placed within a repeating sequence of short-short-long tones. It was found that the gap was more detectable when following one of the short tones than it was following the long tone. It was inferred that the perceived grouping boundary is after the long tone (assuming that a gap that violates grouping struc-

ture would be more noticeable). The results of that study converge with the current findings in showing a preference for short-long grouping among English listeners. Hence, it is reasonable to expect that perceptual measures of grouping would replicate the findings reported here for English versus Japanese listeners, though further work is needed to test this empirically.

III. GENERAL DISCUSSION

The current study focused on the perception of grouping, a basic aspect of rhythm perception by which sequences are perceptually segmented into higher-order units. In contrast to much previous research, a cultural difference in the perception of grouping in simple tone patterns was demonstrated: Native English speakers consistently perceived sequences of alternating-duration tones as repeating short-long groups; native Japanese speakers most often perceived the opposite long-short grouping (a grouping preference that almost never occurred among English speakers) and were as a group more variable in their preferences. The observed difference in perception between English and Japanese speakers is notable because the principles underlying auditory grouping have long been thought to be innate and universal.

How can the difference be explained? We assume that the key factor is the different auditory experiences of listeners living in America versus Japan (rather than, say, a different genetic background). The most obvious source of cultural differences in auditory experience is the dominant language of the culture. Two questions must be addressed: What is it in the experience of English speakers that causes them to prefer short-long grouping of tones? How does the experience of Japanese speakers create many listeners with the opposite long-short preference? Below, we propose the hypothesis that language experience is the source of differences in grouping perception. The theoretical perspective adopted is that statistical learning of the duration patterns of common rhythmic units in speech is responsible for shaping low-level grouping biases. Such learning presumably occurs automatically, via exposure to the rhythmic patterns of speech (see Patel *et al.*, 2006), where gaining implicit knowledge of these patterns may serve a useful role in learning word order, or in segmentation of meaningful units from the speech stream.

If speech rhythm is the key factor in shaping the duration-based grouping biases, then linguistic rhythms in English and Japanese should predict these biases. In particular, short-long patterns should be prevalent in English, while Japanese would be expected to be biased toward long-short patterns (though perhaps not as strongly, in light of the lower conformity in Japanese responses). This approach follows the style of proposals made by Jakobson *et al.* (1952) and Kusumoto and Moreton (1997) in that language is proposed to be responsible for shaping basic rhythmic biases. However, while these researchers argued that the rhythmic shape of words was the key factor in driving nonlinguistic grouping preferences, the current study focuses on a different hypothesis, namely, that the most important linguistic unit influencing grouping perception is not the word, but rather the phrase. Specifically, it is proposed that a syntactic parameter influencing the word ordering of language ("head direction") ultimately influences grouping perception, because of the way it influences the rhythm of phrases in Japanese and English.

A. Rhythmic differences between English and Japanese

Broadly speaking, head direction refers to the order of various structural constituents in sentences, such as verbs and their objects and function-content word ordering (Baker, 2001; Nespor and Vogel, 1986; Nespor *et al.*, 1996). English, for example, is a head-complement language, in which verbs precede objects; Japanese is a complement-head language in which objects precede verbs. We focus here on the consequence of head direction on function-content word order and suggest that this ordering could drive the perceptual grouping of nonlinguistic sounds.

The distinction between function words and content words is a universal feature of language (Selkirk, 1984; Shi *et al.*, 1998). Function words (functors), such as articles, prepositions, conjunctions, and pronouns, are high-frequency linguistic elements from a small closed set of words that provide a structural framework for the semantically meaningful content words of a sentence. In English, functors typically *precede* their syntactically related content word, as in "the dog," "to eat," etc., Japanese, in contrast, typically places functors *after* the related content word (Baker, 2001). (Examples of function morphemes in Japanese include case marking particles such as "ga" and "wo" or "ni," which indicate whether a noun is a subject, a direct object, or an indirect object). Notably, the functor/content distinction is available to infants at an early age, well before individual words are recognized (Shi, *et al.*, 2006), and sensitivity to the relative order of functors and content words emerges within the first year of life (Gervain *et al.*, 2008).

A key element of the proposal is the suggestion that functor/content ordering will have acoustic consequences that predict grouping preferences. Functors tend to be monosyllabic and phonologically reduced (Selkirk, 1996; Shi *et al.*, 1998; Shi *et al.*, 2006). It seems likely then that the syntactic difference in functor location will result in an acoustic difference in the typical rhythmic shape of functor + content word phrases (Nespor *et al.*, 1996; Christophe *et al.*, 2003). If the functor has a shorter duration than its associated content word, which is plausible, the syntactic difference in placement of functors would lead to characteristic temporal rhythms in the two languages that match perceptual grouping preferences: short-long in English and long-short in Japanese. To illustrate this, consider short phrases corresponding most closely to the two-element rhythms used in the perceptual study, namely, two-syllable phrases consisting of a monosyllabic function word plus a monosyllabic content word, e.g., "the book" in English versus "hon ga" in Japanese (book+subject marker). Given the difference in placement of functors between English and Japanese, one can make a strong prediction that such short two-syllable phrases will have a short-long duration pattern in English but a long-short duration pattern in Japanese.[1] Thus, the proposal is that

the acoustic-level ramifications of the syntactic organization of word ordering shape nonlinguistic grouping. Such an argument extends more generally to larger phrases because phrase and utterance edges are typically marked by functors (e.g., Morgan *et al.* 1987; Gervain *et al.*, 2008).

While we have suggested the rhythm of linguistic phrases as the driving factor in shaping nonlinguistic segmentation biases, we briefly consider the original proposal of Jakobson *et al.* (1952) and of Kusumoto and Moreton (1997) that the rhythm of words contributes to these biases. Initially it seems unlikely that common disyllabic words in English would be associated with a short-long duration pattern, since English is well known to have a bias for word-initial stress, and since duration is an important acoustic correlate of syllable stress in this language (Cutler and Carter, 1987; Delattre, 1966). However, the tendency for word-initial stress applies to nouns, not to verbs (such as "begin") or multisyllable function words (such as "about"), which are final stress and are highly represented (e.g., comprising 28% of all tokens of the 50 most common disyllabic words in the Brown corpus) (Francis and Kučera, 1982). The large representation of final-stress words would be expected to contribute many short-long elements to English. Similarly, there are reasons to expect that disyllabic words in Japanese would have a long-short temporal pattern. Kubozono (personal communication) has commented "Disyllabic words consisting of two moras (Light+Light) tend to turn into three mora words (Heavy+Light) by the lengthening of the first syllable. Moreover, disyllabic words consisting of four moras (H+H) tend to be shortened to H+L sequences as the final syllable becomes monomoraic. These two tendencies are observed in a wide range of linguistic phenomena in Japanese, from historical sound changes to baby words and word games" (see Kubozono, 2003, 2004). To establish the plausibility of word rhythms being a source of the perceptual biases, measurements of syllable duration patterns in infant-directed speech would need to be made.

Finally, this study has focused on language as a force in shaping basic rhythmic processing, but another important form of auditory experience is music. Might differences in the musics of the two cultures underlie the perceptual grouping differences observed in the current work? For example, if it were found that phrases in Western music typically start with a short-long pattern (e.g., a "pick-up note," as in the opening of the song "Greensleeves") and that phrases in Japanese music typically start with a long-short pattern, this could account for the perceptual differences in grouping preference. It has been noted that while pick-up notes often exist in Western music, they traditionally do not occur in Japanese music (Koizumi, 1984). A preliminary analysis of children's songs by our laboratory suggest that most musical phrases in both cultures start with notes of equal duration, suggesting that musical patterns cannot explain the biases seen in study 1 (see EPAPS). Further works analyzing larger corpora of music or quantifying the actual musical experience of infants are clearly called for to more fully test this possibility. An additional consideration is whether familiarity with western music notation might affect the grouping preferences expressed by adults.

B. Language development

The preceding discussion suggested a correlation between the perceptual grouping preferences of listeners and rhythmic patterns in their languages, which is consistent with an experience-based explanation of grouping preferences. However, the question of whether language experience actually causes the perceptual grouping differences awaits more direct study. Developmental studies could illuminate the timing of development of perceptual grouping, and whether it corresponds with language development milestones.

One important question is whether infants start life with inborn rhythmic grouping preferences, or whether these emerge purely as a function of linguistic experience. For example, do all infants begin with a short-long grouping preference, which is then reinforced by some languages and partially overridden by others? Or, do infants begin life with no rhythmic grouping bias and learn the bias while learning language? An answer to this question was recently provided by research on 5–8 month old English-immersed infants (living in Canada) by Yoshida *et al.* (2008). Using nonlinguistic tone sequences based on those in the current study, these researchers found that, like adults, 7–8 month olds prefer short long rhythmic groups (replicating Trainor and Adams, 2000). Interestingly, however, 5–6 month olds do not show a grouping preference. These results suggest that there is no innate bias, but that grouping preferences are learned sometime between 6 and 8 months. This is a period in which many language skills also emerge (including sensitivity to the order of function and content words, Gervain *et al.*, 2008), a clear suggestion that learning of linguistic factors (such as word order and word segmentation) and perceptual grouping interact. Further research will be needed to determine the precise timing and direction of causality, namely, whether grouping perception precedes and informs language processing, or may be a secondary consequence of it.

C. Linguistic and nonlinguistic rhythm in cross-cultural perspective

The syntactic proposal outlined above also unifies what may have been a puzzling result that supported claims for the universality of grouping perception: why native speakers of three distinct European languages (English, Dutch, and French) all show a similar preference in duration-based grouping of tone sequences, namely, a preference for short-long groups, despite these languages being rhythmically heterogeneous (Vos, 1977; Bell 1977; Hay and Diehl, 2007). Notably, these three languages span two traditional linguistic rhythm classes, with English and Dutch being "stress-timed" languages and French being a "syllable-timed" language. Related empirical research has revealed systematic differences in the temporal structure of English and Dutch on the one hand and French on the other, based on the temporal patterning of vowels and consonants in sentences (Ramus *et al.*, 1999; Grabe and Low, 2002; Lee and Todd, 2004). The key point for the current purposes, however, is that all of these languages have a level of linguistic structure at which they are similar, namely, in placing functors before content words

(e.g., the book, "le livre," and "het boek"). According to the current view, it is this syntactic similarity (ultimately expressed in terms of durational contrasts) that is responsible for the common preference for short-long groups across these cultures, via mechanisms suggested earlier.

If this view is correct, it should be possible to predict nonlinguistic grouping preferences in other cultures based on the syntactic structure of the native language. Since most European languages place functors before content words (Dryer, 2005), one can predict that native speakers of these languages will exhibit a preference for short-long grouping. Some European and many non-European languages, however (e.g., in Asia and India), resemble Japanese in placing functors after content words. Accordingly, one can predict that native speakers of these languages (e.g., Turkish, Korean, and Marathi) will prefer long-short grouping.

In making predictions about grouping perception based on native language, two additional factors must be kept in mind. First, the temporal structure of common words may also be important in shaping nonlinguistic grouping biases, as in the original proposal of Jakobson *et al.* (1952). The second factor that should be kept in mind concerns multilingualism. If implicit learning of linguistic duration patterns shapes nonlinguistic grouping, then exposure to or facility in a second language may have an influence on grouping preferences. In terms of the current study, the variability in grouping preferences seen among Japanese listeners may have reflected variation in the degree of proficiency or experience with English. Future cross-cultural work should quantify proficiency in non-native languages, to see if this can account for within-culture variability seen in grouping preferences. It also remains to be studied if grouping preferences are "locked in" during language development (as word segmentation strategies may be; Cutler, 2000) or can be modified by experience later in life.

D. Conclusion

The present work demonstrated that perception of basic grouping in nonlinguistic auditory sequences varies by culture, in contrast to long-held views about universal principles governing such grouping. The source of cultural variation in grouping perception is hypothesized to lie in experience of the rhythms of language, suggesting that learning of language may have consequences for the low-level rhythmic perception of sound.

ACKNOWLEDGMENTS

We thank N. Azechi, J. Fry, B. Hayes, H. Kubozono, E. Moreton, D. Roland, S. Roland, M. Sadakata, S. Shattuck-Huffnagel, N. Warner, and two anonymous reviewers. This work was supported by Neurosciences Research Foundation as part of its program on music and the brain at The Neurosciences Institute, where J.R.I. is the Karp Foundation Fellow and A.D.P. is the Esther J. Burnham Fellow. This work was additionally supported by Grants No. 14101001 and No. 19103003 from the Japan Society for the Promotion of Science.

[1]Indeed, the short-long bias of such phrases groups in English is a likely source of the prevalence of "iambic" (weak-strong) meters in English poetic verse (Gall, 1987).

Baker, M. C. (**2001**). *The Atoms of Language* (Basic Books, New York).
Bell, A. (**1977**). "Accent placement and perception of prominence in rhythmic structures," in *Studies in Stress and Accent*, edited by L. Hyman (UCLA Department of Linguistics, Los Angeles), pp. 1–13.
Bent, T., Bradlow, A. R., and Wright, B. A. (**2006**). "The influence of linguistic experience on the cognitive processing of pitch in speech and non-speech sounds," J. Exp. Psychol. Hum. Percept. Perform. **32**, 97–103.
Bolton, T. (**1894**). "Rhythm," Am. J. Psychol. **6**, 145–238.
Bregman, A. S. (**1990**). *Auditory Scene Analysis* (MIT, Cambridge, MA).
Christophe, A., Nespor, M., Guasti, M. T., and Van Ooyen, B. (**2003**). "Prosodic structure and syntactic acquisition: The case of the head-direction parameter," Dev. Sci. **6**, 211–220.
Cutler, A., and Carter, D. (**1987**). "The predominance of strong initial syllables in the English vocabulary," Comput. Speech Lang. **2**, 133–142.
Cutler, A. (**2000**). "Listening to a second language through the ears of a first," Interpreting **5**, 1–23.
Dauer, R. M. (**1983**). "Stress-timing and syllable-timing reanalyzed," J. Phonetics **11**, 51–62.
Davis, M. H., and Johnsrude, I. S. (**2007**). "Hearing speech sounds: Top-down influences on the interface between audition and speech perception," Hear. Res. **229**, 132–147.
Delattre, P. (**1966**). "A comparison of syllable length conditioning among languages," IRAL **4**, 183–198.
Deutsch, D. (**1991**). "The tritone paradox: An influence of language on music perception," Music Percept. **8**, 335–347.
Drake, C., and Bertrand, D. (**2001**). "The quest for universals in temporal processing in music," Ann. N.Y. Acad. Sci. **930**, 17–27.
Dryer, M. S. (**2005**). "Relationship between the order of object and verb and the order of adposition and noun phrase," in *The World Atlas of Language Structures*, edited by M. Haspelmath, M. S. Dryer, D. Gil, and B. Comrie (Oxford University Press, New York), pp. 386–389.
See EPAPS Document No. E-JASMAN-124-045810 for sound examples and music analysis. This document can be reached via a direct link in the online article's HTML reference section or via the EPAPS homepage (http://www.aip.org/pubservs/epaps.html).
Fraisse, P. (**1982**). "Rhythm and tempo," in *The Psychology of Music*, edited by D. Deutsch (Academic, London), pp. 149–180.
Francis, W. N., and Kučera, H. (**1982**). *Frequency Analysis of English Usage: Lexicon and Grammar* (Houghton Mifflin, Boston).
Gall, S. M. (**1987**). "Versification," in *Poetry in English: An Anthology*, edited by M. L. Rosenthal (Oxford University Press, Toronto).
Gervain, J., Nespor, M., Mazuka, R., Horie, R., and Mehler, J. (**2008**). "Bootstrapping word order in prelexical infants: a Japanese-Italian cross-linguistic study," Cogn. Psychol. **57**, 56-74.
Grabe, E., and Low, E. L. (**2002**). "Durational variability in speech and the rhythm class hypothesis," in *Laboratory Phonology 7*, edited by C. Gussenhoven and N. Warner (Mouton de Gruyter, Berlin), pp. 515–546.
Hannon, E. E., and Trehub, S. E. (**2005**). "Tuning in to musical rhythms: Infants learn more readily than adults," Proc. Natl. Acad. Sci. U.S.A. **102**, 12639–12643.
Hay, J. S. F., and Diehl, R. L. (**2007**). "Perception of rhythmic grouping: Testing the iambic/trochaic law," Percept. Psychophys. **69**, 113–122.
Hayes, B. (**1995**). *Metrical Stress Theory: Principles and Case Studies* (University of Chicago Press, Chicago).
Huron, D., and Ollen, J. (**2003**). "Agogic contrast in French and English themes: Further support for Patel and Daniele (2003)," Music Percept. **21**, 267–271.
Jakobson, R., Fant, G., and Halle, M. (**1952**). "Preliminaries to Speech Analysis: The Distinctive Features and Their Correlates," Acoustics Laboratory, MIT, Technical Report No. 13.
Koizumi, F. (**1984**). *Research on Japanese Traditional Music 2 (Rhythm). [Nihon dento ongaku no kenkyu 2 (Rizumu)]*. (Ongaku no tomo sha, Tokyo).
Kubozono, H. (**2003**). "The syllable as a unit of prosodic organization in Japanese," in *The Syllable in Optimality Theory*, edited by C. Fery and R. van der Vijver (Cambridge University Press, Cambridge), pp. 99–122.
Kubozono, H. (**2004**). "Weight neutralization in Japanese," J. Japanese Ling. **20**, 51–70.
Kusumoto, K., Moreton, E. (**1997**). "Native language determines parsing of nonlinguistic rhythmic stimuli," J. Acoust. Soc. Am. **105**, 3204.

Lee, C. S., and Todd, N. P. McA., (**2004**). "Toward an auditory account of speech rhythm: Application of a model of the auditory 'primal sketch' to two multi-language corpora," Cognition **93**, 225–254.
Lerdahl, F. and Jackendoff, R. (**1983**). "A generative theory of tonal music" (MIT Press, Cambridge, MA).
Mehler, J., Dupuox, E., Nazzi, T., and Dehaene-Lambertz, D. (**1996**). "Coping with linguistic diversity: The infant's viewpoint," in *Signal to Syntax*, edited by J. L. Morgan and D. Demuth (Lawrence Erlbaum, Mahwah), pp. 101–116.
Morgan, J. L., Meier, R. P., and Newport, E. L. (**1987**). "Structural packaging in the input to language learning: Contributions of prosodic and morphological marking of phrases in the acquisition of language," Cogn. Psychol. **19**, 498–550.
Nazzi, T., Bertoncini, J., and Mehler, J. (**1998**). "Language discrimination in newborns: Toward an understanding of the role of rhythm," J. Exp. Psychol. Hum. Percept. Perform. **24**, 756–777.
Nespor, M., and Vogel, I. (**1986**). *Prosodic Phonology*. (Foris, Dordrecht).
Nespor, M., Guasti, M. T., and Christophe, A. (**1996**). "Selecting word order: The rhythmic activation principle," in *Interfaces in Phonology*, edited by U. Kleinhenz (Akademie, Berlin), pp. 1–26.
Patel, A. D., and Daniele, J. R. (**2003**). "An empirical comparison of rhythm in language and music," Cognition **87**, B35–B45.
Patel, A. D., Iversen, J. R., and Rosenberg, J. C. (**2006**). "Comparing the rhythm and melody of speech and music: The case of British English and French," J. Acoust. Soc. Am. **119**, 3034–3047.
Ramus, F., Nespor, M., and Mehler, J. (**1999**). "Correlates of linguistic rhythm in the speech signal," Cognition **73**, 265–292.
Selkirk, E. O. (**1984**). *Phonology and Syntax: The Relation Between Sound and Structure*. (MIT, Cambridge, MA).
Selkirk, E. (**1996**). "The prosodic structure of function words," in *Signal to syntax: Bootstrapping from Speech to Grammar in Early Acquisition*, edited by J. L. Morgan and K. Demuth (Lawrence Erlbaum, Mahwah, NJ), pp. 187–213.
Shi, R., Morgan, J., and Allopenna, P. (**1998**). "Phonological and acoustic bases for earliest grammatical category assignment: A cross-linguistic perspective," J. Child Lang **25**, 169–201.
Shi, R., Werker, J. F., and Cutler, A. (**2006**). "Recognition and representation of function words in English-learning infants," Infancy **10**, 187–198.
Trainor, L. J., and Adams, B. (**2000**). "Infants″ and adults″ use of duration and intensity cues in the segmentation of tone patterns," Percept. Psychophys. **62**, 333–340.
Trehub, S. E., and Trainor, L. J. (**1993**). "Listening strategies in infancy: The roots of music and language development," in *Thinking in Sound: The Cognitive Psychology of Human Audition*, edited by S. McAdams and E. Bigand (Oxford University Press, Oxford) pp. 278–327.
Vos, P. (**1977**). "Temporal duration factors in the perception of auditory rhythmic patterns," Scientific Aesthetics/Sciences de l'Art **1**, 183–199.
Wertheimer, M. (**1938**). in *A Source Book of Gestalt Psychology*, edited by W. Ellis (Routledge and Kegan Paul, London), pp. 71–88.
Woodrow, H. (**1909**). "A quantitative study of rhythm: The effect of variations in intensity, rate and duration," Arch. Psychol. (Frankf) **14**, 1–66.
Yoshida, K. A., Iversen, J. R., Patel, A. D., and Werker, J. F. (**2008**). "Development of Abstract Grammatical Representation," *26th Biennial International Conference on Infant Studies*, Vancouver, BC, Canada.