Hierarchical Structure of Musical and Visual Meter in Cross-modal “Fit” Judgments

Stephen E. Palmer & Joshua Peterson

Abstract—The metrical hierarchy of musical rhythm is defined by the structure of emphasis on beats in measures and has been studied in several ways [1]. Here we investigated the perceived structure of 3/4 and 4/4 time signatures in auditory and visual meter using cross-modal goodness-of-fit ratings for visual and auditory probes, respectively. In the *auditory context conditions*, four measures in 3/4 or 4/4 time were defined by a louder beat followed a series of 2 or 3 softer, equally-timed beats, respectively. A visual probe circle was introduced into the next four measures at one of 12 phase-angles relative to the auditory downbeat: 0, 45, 60, 90, 120, 135, 180, 225, 240, 270, 300, or 315 degrees. In the *visual context conditions*, context and probe modalities were reversed, with analogous visual rhythms being defined by a larger downbeat circle followed by a series of 2 or 3 smaller circles, with an auditory probe in the last four measures at one of the same 12 phase-angles. Participants rated how well the probe stimulus “fit” the rhythmic context in the other modality. For the visual context conditions, the probe’s effect on fit-ratings revealed the expected beat-defined metrical hierarchy. In 4/4 time, fit ratings were highest for beat 1, next highest for probes at (or near) beats 2, 3, and 4, and lowest for probes at non-beats. In 3/4 time, they decreased similarly from beat 1 to beats 2 and 3, and from them to non-beats. The auditory context conditions produced unexpected results, however, with a single broad peak at and following the downbeat, and little evidence of elevated fit ratings for other beats over non-beats. Similar results were obtained when participants made explicit ratings of cross-modal synchrony using the same stimuli. Various factors relevant to explaining the asymmetry between these cross-modal conditions are discussed.

***Keywords*—Goodness-of-fit, Metrical hierarchy, rhythm, time signatures.**

I. Introduction

The metrical hierarchy of rhythm in music is defined by the structure of emphases related to beats in measures at multiple levels. This hierarchy is an important component of music theory because it structures the perceptual organization of music in time (e.g., Cooper & Meyer 1960; Lehrdal & Jackendoff, 1983) [2, 3]. Its empirical basis was explored most notably by C. Palmer and Krumhansl (1990) [1], who investigated three different measures of metrical structure: frequency distributions of note onsets in the scores of musical compositions, memory confusions in discrimination tasks, and goodness-of-fit judgments of temporal patterns in metrical contexts. In each case they found evidence of a multi-level metrical hierarchy of accent strength, with the strongest emphasis on the primary beat (downbeat) of each measure, and the next-strongest emphasis on the secondary beats of the time signatures they investigated: 2/4, 3/4, 4/4, and 6/8.

Most relevant to the issue of how listeners perceive temporal organization and structure in music, C. Palmer and Krumhansl (1990) [1] measured goodness-of-fit judgments in the auditory modality for higher-pitched probe beats relative to lower-pitched context beats. The context beats marked the downbeats of a series of several measures, and participants were asked to think of each context beat as the first of N beats, where N = 2, 3, 4, or 6. They were then asked to rate "how well the high-pitched (probe) beats fit with the (low-pitched) context beats" using a 7-point Likert scale, in which 7 indicated the best fit. The metrical hierarchy was evident in the pattern of their results: probe beats received higher ratings when they coincided with the primary and secondary beats of the implied metrical hierarchy than when they did not. These effects were stronger for people with more extensive musical training than for those with less training. Such experiments have restricted the probe and context modalities to auditory events, presumably because the temporal structure of music is primarily perceived in the auditory modality.

Nevertheless, the metrical structure of music and music-related events can also be evident in the visual, tactile, and kinesthetic modalities. Perhaps most obviously, the conductor of an orchestra or chorus moves her hands rhythmically to indicate metrical structure, typically using downward hand motions to indicate the primary beat of a measure (the downbeat) and sideways or upward motions to indicate secondary beats. These movements communicate metrical structure in the visual modality to musicians so that they can coordinate and synchronize their performances. Musicians themselves also move their hands and bodies rhythmically in time with the metrical hierarchy, as do novice and expert dancers alike. Further examples of aesthetic domains in which the timing of visual events provides information about the metrical structure of simultaneously played music include music videos and music visualizers.

In this report we describe our initial attempts to study the

Stephen E. Palmer is with the Department of Psychology, University of California, Berkeley CA 94720-1650 (phone: 510-684-4447; fax: 510-642-5293; e-mail: sepalmer@gmail.com).

Joshua Peterson is with the Department of Psychology, University of California, Berkeley CA 94720-1650 (e-mail: jpeterson@berkeley.edu).

cross-modal perception of musical meter in the visual and auditory modalities. We used an extension of C. Palmer and Krumhansl's (1990) [1] goodness-of-fit paradigm that we adapted for cross-modal stimuli by presenting larger and smaller white circles visually and playing louder and softer beats auditorily. Half of the participants rated the fit of a visual probe flash relative to an auditory context of loud and soft beats in 3/4 or 4/4 time, and the other half rated the fit of an auditory probe beat relative to a visual context of larger and smaller circles in 3/4 or 4/4 time. We expected that both cross-modal context conditions would essentially replicate the results of C. Palmer and Krumhansl (1990) [1], with the caveat that the auditory context might provide stronger evidence of metrical structure than the visual context because of its stronger associations with music perception. This did not turn out to be the case, however.

II. Methods

A. Design. The experimental design consisted of three orthogonal factors: *context condition* (visual or auditory), *duration condition* (long, medium, or short), and *probe phase-angle condition* relative to the downbeat (0°, 45°, 60°, 90°, 120°, 135°, 180°, 225°, 240°, 270°, 300°, or 315°). Context condition and duration condition were between-subjects variables, whereas probe phase-angle was a within-subject variable.

B. Participants. 24 undergraduates at the University of California, Berkeley, participated in the study. The first 8 were run in the long duration condition, the next 8 in the medium duration condition, and the last 8 in the short duration condition. In each duration condition, 4 participants ran in the visual context conditions and 4 in the auditory context conditions, randomly assigned across participants. Musical training varied unsystematically from 2 to 10 years, but we do not yet have enough participants in the sample to meaningfully examine training effects in comparable stimulus conditions.

C. Stimuli. In the auditory context conditions, four measures in 3/4 or 4/4 time were defined by a louder beat followed a series of 2 or 3 softer, equally-timed beats, respectively. A visual probe of a solid white circle was introduced into each of the next four measures at one of 12 phase-angles relative to the auditory downbeat: 0°, 45°, 60°, 90°, 120°, 135°, 180°, 225°, 240°, 270°, 300°, and 315°(see Fig. 1). In the visual context conditions, context and probe modalities were reversed, with analogous visual rhythms being defined by presenting a larger (downbeat) white circle followed by a series of 2 or 3 smaller white circles, with an auditory probe in the last four measures at one of the same 12 phase-angles.

The downbeats occurred at the 0° probe phase-angle in both the 3/4 and 4/4 time signatures, with each measure being 2 seconds long. The secondary beats occurred at 120° and 240° in the 3/4 time condition, and at 90°, 180°, and 270° in the 4/4 time condition as indicated in Fig. 1. The 60°, 180°, and 300° phase-angles are more closely related to the 3/4 condition because they are the "back-beats" or "half-beats" intermediate between the 3/4 beats at 0°, 120°, and 240°. In contrast, the 45°, 135°, 225°, and 315° phase-angles are more closely related to the 4/4 condition because they are likewise intermediate between the 4/4 beats at 0°, 90°, 180°, and 270°.

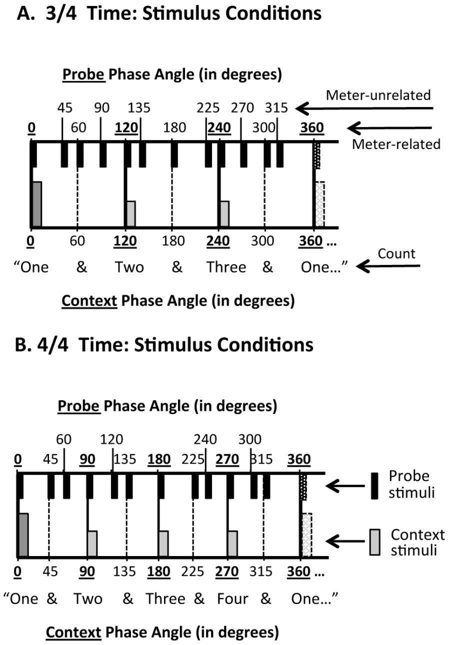
A. 3/4 Time: Stimulus Conditions
Probe Phase Angle (in degrees)
45 90 135 225 270 315
Meter-unrelated
0 60 120 180 240 300 360
Meter-related
0 60 120 180 240 300 360 …
"One & Two & Three & One…"
Count
Context Phase Angle (in degrees)
B. 4/4 Time: Stimulus Conditions
Probe Phase Angle (in degrees)
60 120 240 300
0 45 90 135 180 225 270 315 360
Probe stimuli
Context stimuli
0 45 90 135 180 225 270 315 360 …
"One & Two & Three & Four & One…"
Context Phase Angle (in degrees)

Fig. 1. Temporal structure of cross-modal stimulus events in 3/4 time (A) and 4/4 time (B). Context stimuli in one modality are represented in gray at the bottom, and probe stimuli in the other modality are represented at the top. Only one probe was presented per trial. Taller, darker context rectangles represent louder or brighter downbeat stimuli.

There were three event duration conditions, which will be referred to as long, medium, and short. In the long condition, the auditory stimuli were loud and soft snare drumbeats that lasted 250 ms including natural reverberation, and the visual stimuli were large and small circles that lasted 32 ms. In the medium condition, the snare drumbeat was trimmed to 100 ms (no reverberation and 50ms fade-out) and the visual stimuli were again 32ms. In the short condition, the auditory stimuli were loud and soft sinusoids at 523.25 Hz (C5) that lasted 16 ms (to match one 60 Hz monitor frame refresh cycle) and the visual stimuli were large and small circles that also lasted 16 ms (one frame). The long duration condition was studied initially to provide natural sounding auditory stimuli and natural looking visual stimuli. The medium and short duration conditions were included to test stimulus conditions with increasingly more precise timing to avoid cross-modal contamination in the synchronization of stimuli.

D. Procedure. Participants were asked to rate how well the probe stimulus (one event per measure) "fit" the rhythmic context of the events in the other modality (three or four beats per measure) using a 400-pixel horizontal line-mark scale. The ends were labeled "Good fit" and "Bad fit" with the left/right positions of the labels counterbalanced across subjects. Participants slid a restricted mouse cursor horizontally over the scale and clicked at the point they felt best represented how well the probe events fit the context events. Ratings ranged from -200 (worst fit) to +200 (best fit).

III. Results

The probe's position affected the fit-ratings quite differently for the visual and auditory context conditions (Fig. 2). When the context consisted of a temporal pattern of visual events and the probe was auditory, the results are reasonably similar to those of C. Palmer and Krumhansl (1990) [1], clearly showing the effects of the expected metrical hierarchy (Figs. 2A and 2B). The evidence for this claim comes from examining the differences between the fit ratings for the 3/4 and 4/4 time.

When the time signature of the visual contextual events was 3/4 (i.e., circles that were *large-small-small*, etc.), fit ratings were highest when the auditory probe coincided with the visual downbeat at 0° and next highest when it coincided with beats 2 and 3 (at 120° and 240°). Rating were also relatively high when the auditory probe was presented slightly after synchrony with visual beats 2 and 3 (at 135° and 270°), presumably because they were relatively "close enough" to be perceived as coinciding with the visual event. There was essentially no evidence that fit ratings for the probes that coincided with the "back-beats" halfway between context beats (i.e., the gray circles in Fig. 2A at 60°, 180°, and 300°) were rated as fitting better than the probes that did not fit into the metrical hierarchy of the 3/4 time signature at all (i.e., the white circles in Fig. 2A at 45°, 90°, 225°, and 315°).

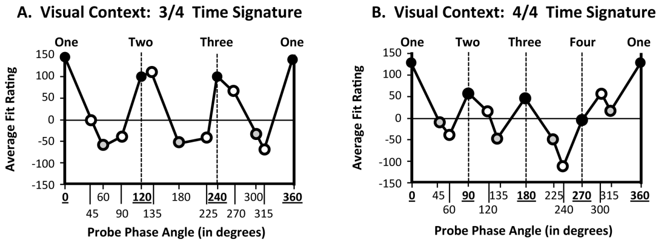

Fig. 2. Average fit ratings for auditory probes within visual contexts in 3/4 time (A) and 4/4 time (B).

When the temporal time signature of the visual events was 4/4 (i.e., circles that were large-small-small-small, etc.), fit ratings were notably higher when the auditory probe coincided with the visual beats at 0°, 90°, 180°, and 270° than when they did not. Again, there was no indication that fit ratings for the probes that coincided with the "back-beats" midway between the beats (i.e., the gray circles in Figure 2A at 45°, 135°, 225°, and 315°) were rated as fitting better than the probes that did not fit into the metrical hierarchy of the 4/4 time signature (i.e., the white circles in Figure 2A at 60°, 120°, 240°, and 300°).

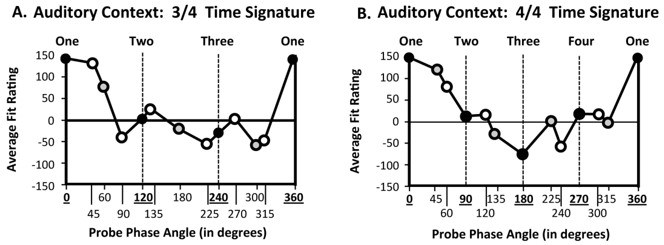

Fig. 3. Average fit ratings for visual probes within auditory contexts in 3/4 time (A) and 4/4 time (B).

To our surprise, the pattern of fit ratings was distinctly different when the context beats were auditory and the probe was visual (Fig. 3) than when the context beats were visual and the probe auditory (Fig. 2). In both the 3/4 (Fig. 3A) and 4/4 time signatures (Fig. 3B), there was a single broad peak around 0° that lasted through 45° and even 60°, dropping to near-zero only when the phase-angle reached 90°. This is in stark contrast with the results in the corresponding visual context conditions, where the auditory probes' fit ratings dropped to nearly zero by a 45° phase lag. We also note that the auditory contexts produced no clear differentiation between the structure of the fit ratings at later time lags in the 3/4 and 4/4 conditions, as were so obviously present for the visual contexts.

However, it is interesting to note that there are two quite modest peaks in the 3/4 condition (Fig. 3A) that occurred when the visual probes occurred just *after* the expected peaks at the 120° and 240° phase-angles, as if the visual probe was experienced a bit *after* the temporally synchronized auditory context events. The timing of these "late" peaks is also potentially consistent with the width of the peak for the downbeat that begins at 0°. That is, both effects are consistent with the possibility that the visual probes took longer to process and were temporally more variable when their fit was being evaluated relative to the auditory context events (i.e., when participants were attending to the visual probe as the to-be-judged event). The same tendency is not apparent in the 4/4 condition, however (Fig. 3B), perhaps because the more rapid series of auditory context events caused more extensive overlap.

Even so, it is interesting that any delay and/or variability in processing visual events does *not* seem to disrupt the perception of meter in the visual context condition where the probes are auditory. That is, peaks are clearly evident at the expected phase-angles in Fig. 2.

We decided to find out whether the asymmetries between the visual context and auditory context conditions in the cross-modal fit ratings might be due to corresponding asymmetries in people's ability to detect cross-modal simultaneity in these stimuli. We ran 8 additional participants asking them to rate "the alignment of the beep and flash in time" using the same rating scale. The results (Fig. 4) show the same pattern

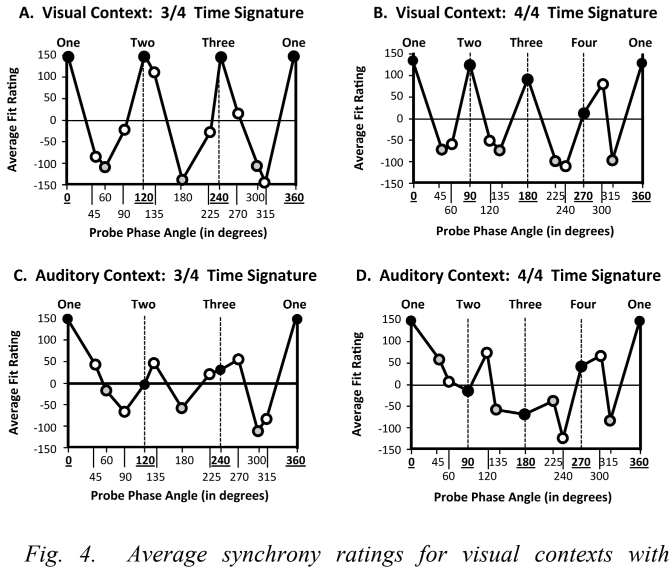

Fig. 4. Average synchrony ratings for visual contexts with auditory probes in 3/4 time (A) and 4/4 time (B) and for auditory contexts with visual probes in 3/4 time (C) and 4/4 time (D).

as the fit ratings. The synchrony ratings for the visual context conditions peak sharply at the synchronous cross-modal beats (Figs. 4A and 4B), and they are highly correlated with the corresponding fit ratings in Fig. 2 (r = .93, .84 for the 3/4 and 4/4 time signatures, respectively, p < .001). In contrast, the synchrony ratings for the auditory context conditions have sharp peaks only around the downbeat (near 0°), with reduced peaks just after the synchronous beats. They too are highly correlated with the corresponding fit ratings (r = .70, .80 for the 3/4 and 4/4 time signatures, respectively, p < .012).

IV. Discussion

Why might this asymmetrical pattern of cross-modal effects arise in both people's fit ratings and their simultaneity ratings? Several factors seem likely to be relevant.

First, there is a great deal of evidence from the sensory motor literature that visual processing is both slower and less accurate temporally than auditory processing (Repp, 2003 [4]; see Repp, 2005 [5], and Repp & Su, 2013 [6], for reviews). For example, people are able to synchronize finger-tapping at much faster beat-rates with auditory metronomes than visual ones. Slower and more variable processing of visual probes could explain several features of the present results: the broad peaks evident around the downbeat in the auditory context conditions (Figs. 3A and 3B), the greatly reduced secondary peaks around the on-beats in 3/4 time (Fig. 3A), and perhaps even the lack of secondary peaks around the on-beats in 4/4 time (Fig. 3B). Slower, more variable processing of visual events alone, however, fails to predict why the results from the visual context conditions are so much cleaner and more precise than those from the auditory context conditions (compare Figs. 2 and 3). If visual events cause problems only when they constitute probes and not when constitute the metrical context, some other factor(s) must be at work.

A second consideration is the differential role of temporal certainty/uncertainty in the context and probe events of the task itself. Participants initially perceive four measures of the context meter alone. This presumably allows them to create a precise temporal template for anticipating upcoming context beats. In measure 5, the first appearance of the probe is maximally uncertain, because it can occur anywhere in the measure. Its additional occurrences in measures 6-8 presumably reduce its temporal uncertainty, but not to the same degree that the context beats in measures 1-4 do. This consideration suggests that by the 5th measure, the visual context condition may have produced a very stable and accurate template of expectations for context beats, even if processing is slower and more variable in vision than audition.

Third, the two cross-modal conditions may not have been equally cross-modal, in that participants may well have converted the metrical visual context into quasi-auditory form by inwardly, or even outwardly, counting the meter throughout the trial ("***one***, two, three, ***one***, two, three," etc. for 3/4 time). This transformation would have the desirable effect of making the ostensibly cross-modal "visual context" condition much more nearly uni-modal, since both the probes and context would end up being represented auditorily.

Note that the same is unlikely to be true in the auditory context condition with the visual probe. If the optimal strategy is to convert the visual stimulus into an auditory one, perceivers would have to inwardly subvocalize to the visual probe. The problem is that the probe occurs at a very uncertain time.

It seems likely that any or all of three of these factors may be relevant to understanding both the fit ratings and the synchrony ratings. Further experiments manipulating variables related to these factors should help determine the extent to which they influenced the present results.

V. Conclusions

Cross-modal fit ratings visual contexts with auditory probes replicate and extend prior results on the metrical hierarchy for auditory probes in auditory contexts by C. Palmer and Krumhansl (1990). Corresponding ratings for visual probes relative to auditory contexts are quite different, however, showing elevated fit ratings primarily on and shortly after the louder auditory downbeat. Similar results were obtained when participants were explicitly asked to rate the degree of simultaneity of a probe within a context using the same stimuli. This correspondence strongly implies that people tend to find that synchronous cross-modal stimuli "fit well" together, and that difficulties in detecting cross-modal synchrony can lead to corresponding perceptions of poor fit. In future research we hope to track down the reason for these differences in different cross-modal combinations. C. Palmer and Krumhansl's (1990) prior results were for unimodal auditory-to-auditory stimuli – i.e., auditory probes to auditory contexts – using only downbeats as context. By systematically varying the probe and context modalities and the temporal nature of the contextual stimuli, we hope to be able to isolate the conditions under which this asymmetrical pattern of fit ratings is obtained and those under which it can be eliminated.

Acknowledgment

We thank Nori Jacoby, Caroline Palmer, and Carol Krumhansl for their comments on a previous draft of this paper. This research was sponsored in part by NSF Grant BCS-1059088 from the NSF to S.E.P.

References

[1] Palmer, C., & Krumhansl, C. L. (1990). Mental representations for musical meter. *Journal of Experimental Psychology: Human Perception and Performance*, *16*(4), 728.

[2] Cooper, G., & Meyer, L. B. (1960). The rhythmic structure of music. University of Chicago Press, Chicago.

[3] Lerdahl, F., & Jackendoff, R. (1983). An overview of hierarchical structure in music. *Music Perception: An Interdisciplinary Journal*, *1*(2), 229-252.

[4] Repp, B. H. (2003) Rate Limits in Sensorimotor Synchronization With Auditory and Visual Sequences: The Synchronization Threshold and the Benefits and Costs of Interval Subdivision, Journal of Motor Behavior, 35:4, 355-370.

[5] Repp, B. H. (2005). Sensorimotor synchronization: A review of the tapping literature. *Psychonomic bulletin & review,* 12(6), 969-992.

[6] Repp, B. H., & Su, Y. (2013). Sensorimotor synchronization: a review of recent research (2006–2012). *Psychonomic Bulletin & Review,* 20(3), 403-452.