Neuroscience 2018 *389* 141-151

MULTISENSORY INTEGRATION IN SHORT-TERM MEMORY: MUSICIANS DO ROCK ☆

AVIGAEL M. AIZENMAN,[a] JASON M. GOLD[b] AND ROBERT SEKULER[c]*

[a] *University of California, Berkeley, CA, USA*

[b] *Indiana University, Bloomington, IN, USA*

[c] *Brandeis University, Waltham, MA, USA*

Abstract—Demonstrated interactions between seeing and hearing led us to assess the link between music training and short-term memory for auditory, visual and audiovisual sequences of rapidly presented, quasi-random components. Visual sequences' components varied in luminance; auditory sequences' components varied in frequency. Concurrent components in audiovisual sequences were either congruent (the frequency of an auditory item increased monotonically with the luminance of the visual item it accompanied), or incongruent (an item's frequency was uncorrelated with luminance of the item it accompanied). Subjects judged whether the last four items in a sequence replicated its first four items. With audiovisual sequences, subjects were instructed to ignore the sequence's auditory components, basing their judgments solely on the visual input. Subjects with prior instrumental training significantly outperformed their untrained counterparts, with both auditory and visual sequences, and with sequences of correlated auditory and visual items. Reverse correlation showed that the presence of a correlated, concurrent auditory stream altered subjects' reliance on particular visual items in a sequence. Moreover, congruence between auditory and visual items produced performance above what would be predicted from simple summation of information from the two modalities, a result that might reflect a contribution from special-purpose, multimodal neural mechanisms.

This article is part of a Special Issue entitled: Sequence Processing

Key words: categorization, multisensory, short-term memory, audiovisual integration.

INTRODUCTION

Unsurprisingly, practice playing an instrument enhances music-related skills (Hyde et al., 2009; Kraus and Chandrasekaran, 2010). Surprisingly, though, such training has also been linked to superior performance on tasks with little or no obvious connection to music (Chan et al., 1998; Oxenham et al., 2003; Francois and Schön, 2011; Bergstrom et al., 2012; Strait et al., 2012). Demonstrated cross-talk between auditory and visual processing Sekuler et al., 1997; Guttman et al., 2005; Berger and Ehrsson, 2013) might explain these generalized effects of musical training. However, evidence for cross-modal impact of music training is weak or even negative (e.g., Cohen et al., 2011). The "modality-appropriateness" hypothesis (Welch and Warren, 1980) suggests one way to understand failures to find cross-modal effects of music training. Specifically, the modality-appropriateness hypothesis asserts that when visual and auditory processing are compared, the advantage goes to vision when *spatial* attributes must be processed, but the advantage shifts to audition when *temporal* attributes are critical (Welch, 1999; Guttman et al., 2005). Moreover, this hypothesis, which was based on behavioral results alone, received additional support from a recent functional magnetic resonance imaging (fMRI) study. Michalka et al. (2015) showed that task demands can dynamically recruit different modality-related frontal lobe regions: a visual task with rapid stimulus presentation activates cortical regions normally implicated in auditory attention, while an auditory task that demands spatial judgements activates regions normally implicated in visual attention. Therefore, to examine music training's possible impact on vision we devised a task in which subjects judged visual sequences that were presented at a rate rapid enough to advantage processing by the auditory attentional system.

Our study built on a visual task that Gold et al. (2014) adapted from one introduced by Agus et al. (2010) to study auditory memory. Gold et al. (2014) presented subjects with rapidly presented sequences of quasi-random luminance levels, and asked them to judge whether the second four luminance levels in a sequence identically repeated the first four. Some of their results showed that performance in their task required both successful perceptual segmentation and good short-term memory. Stimuli in their experiments entailed sequences of rapid variation along "elemental" or "low-level" sensory dimensions (Magnussen, 2000; Pasternak and Greenlee, 2005). Sequences of low-level sensory attributes are

☆ Supported by CELEST, an NSF Science of Learning Center (SBE-035478), National Institutes of Health grant EY-019265, and by AFOSR grant FA9550-10-1-0420. We thank Trevor Agus, Barbara Shinn-Cunningham and Randolph Blake for insightful comments on an earlier version of this paper, and Abigail Noyce and Arielle Keller for their assistance. A portion of this research was presented at a meeting of the Vision Sciences Society.

*Corresponding author.

E-mail address: vision@brandeis.edu (R Sekuler).

http://dx.doi.org/10.1016/j.neuroscience.2017.04.031

useful experimental probes, in part because they reduce the likelihood that subjects' performance would be mediated by verbal labels (Miller and Gazzaniga, 1998; Kahana and Sekuler, 2002). For our purposes, such sequences offer another potential advantage. Although subjects' self-reports are not dispositive (Nisbett and Wilson, 1977), some of Gold et al.'s (2014) subjects volunteered that as they were observing the visual sequences, they generated subvocal tunes, in the form of melodic contours. In other words, they claimed to have recruited auditory imagery for what nominally was a purely visual task (Berger and Ehrsson, 2013), suggesting a form of cross-talk between modalities that Guttman et al. (2005) described as "hearing what the eyes see".

Prior demonstrations of cross-talk between seeing and hearing led us to ask whether musicianship would enhance processing of rapidly presented visual stimuli. For an answer, we adapted Gold et al.'s (2014) paradigm, comparing music-trained and non-trained subjects performance with rapidly presented stimulus sequences of varying luminance levels or auditory frequencies (Rammsayer and Altenmüller, 2006). Additionally, as many ordinary events generate multisensory signals, and the confluence of signals from multiple senses can powerfully influence perception (e.g., Thomas, 1941; Chen and Spence, 2010), the paradigm allowed us to test music-trained and non-trained subjects with multisensory sequences, that is, sequences whose co-occurring audio and visual components were perceptually congruent or perceptually incongruent.

Method

In all of our test conditions, subjects had to judge whether the first four items in a rapidly presented stimulus sequences of eight items were or were not repeated by the last four items. Fig. 1 shows schematic examples of our unimodal stimuli, with Visual stimuli in Panel A and Auditory stimuli in Panel B. Items of each type were drawn from a homogeneous pool and were devoid of semantic content.

As in Gold et al. (2014), visual stimuli were presented against a uniform background of average luminance 19.03 cd/m^2 on a 17″ CRT monitor (Sony Trinitron UltraScan P780) with a resolution of 1024 × 768 pixels and a refresh rate of 75 Hz. Display luminances were linearized by means of a calibration-based lookup table. Stimulus sequences in which luminance and/or auditory frequency were generated and presented by an Apple iMac computer, using Matlab (version 7.7) and extensions from the Psychophysics Toolbox (Brainard, 1997). Each visual sequence comprised eight luminance levels presented in rapid succession to the same 4.1° × 4.1° (128 × 128 pixels) region at the display's center. Each luminance level in an eight-item sequence was presented for 10 complete refreshes of the CRT screen (∼133 ms), which meant that a complete eight-item sequence played out in ∼1067 ms. A viewing distance of 57 cm was enforced by means of a chin rest.

Auditory stimuli were streams of eight equal-duration pure tones, each ∼133 ms in duration. These tones were sampled at 44.1 kHz and presented at 70–72 db

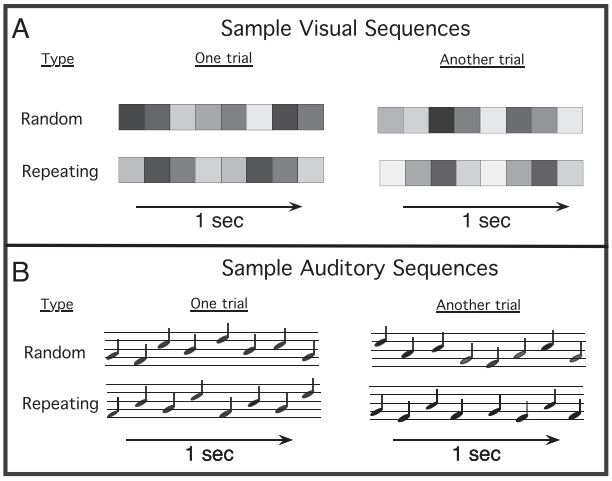

Fig. 1. Schematic examples of auditory and visual unimodal stimuli. Panel A: exemplars of auditory stimuli; Panel B: exemplars of visual stimuli. In each panel, two examples are shown for the Random condition (the last four items in an eight-item sequence are uncorrelated with the first four) and for the Repeating condition (the last four items in an eight-item sequence repeat identically the first four).

(A) through Sennheiser HD280 supra-aural earphones. To eliminate audible transients that would arise from abrupt changes in frequency from one tone to another, the leading and trailing edges of each tone were tapered with a raised cosine (∼1.13 ms rise or fall time).

For a multimodal stimulus sequence, auditory and visual components of the sequence were presented synchronously. The synchronization of auditory and visual sequences was assessed using photodiode and microphone inputs to a dual-trace oscilloscope. Measurements showed that the two streams were synchronized to ±7 ms.

To determine what luminances would be presented, the eight samples drawn for the trial were translated into equivalent luminance contrasts. Contrast was defined as $(L_{pix} - L_{bg})/L_{bg}$, where L_{pix} is the luminance of a stimulus pixel, and L_{bg} is the display's background luminance, which was held constant at 19.03 cd/m^2. The resulting samples ranged from 2 cd/m^2 to 42 cd/m^2. When a stimulus sequence included an auditory component, the eight luminances in the sequence were translated into equivalent pure tones whose frequencies were a linear function of luminance (see Fig. 2).

The stimulus-generation algorithm started by drawing eight random samples from a normal distribution, $\mathcal{N}(0, 0.2)$. Samples more than ±2 standard deviations from the mean of the normal distribution were discarded and replaced. Together with the distribution's relatively small standard deviation, censoring extreme values homogenized items presented in any stimulus sequence. This made it difficult for subjects base judgments on any highly-distinctive, "oddball" item or items. For example, successive samples in visual sequences differed by no more than 29.83 cd/m^2, with 10% of successive samples differing by 1.33 cd/m^2 or less, and 50% of successive values differing by 7.03 cd/m^2 or less.

Please cite this article in press as: Aizenman AM et al. Multisensory integration in short-term memory: Musicians do rock. Neuroscience (2017), http://dx.doi.org/10.1016/j.neuroscience.2017.04.031

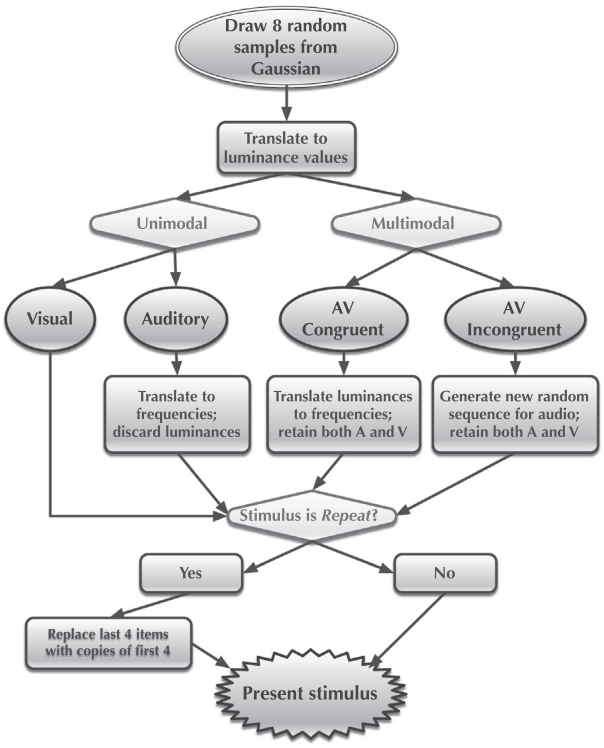

Fig. 2. Flowchart for stimulus generation. The steps in the stimulus generation algorithm are explained in the text.

EXPERIMENTAL PROCEDURES

Valid comparisons between unimodal and multimodal conditions demand that baseline performance with Auditory and Visual sequences be equated. Were the separate unimodal contributions to a multimodal sequence substantially unequal, performance with a multimodal sequence would be dominated by the more potent of the two unimodal drivers. As a first step toward equating performance with Visual and Auditory sequences, we drew on existing brightness-pitch cross-modal matching results reported by Marks (1974). In that study, subjects adjusted the pitch of a tone to match various achromatic Munsell patches. As we were committed to using the same luminance range that Gold et al. (2014) had used, the cross-modal matching results implied that we should use a set of frequencies spanning between two and three octaves, 100–555 Hz (~A2♭ to ~C5♯ on an equal-tempered musical scale).

A preliminary experiment tested 12 Non-musicians with Visual stimuli generated and presented just as Gold et al. (2014) did, and also with Auditory sequences that were drawn from the frequency range implied by Marks's (1974) cross-modal matching result, that is, 100–555 Hz. For each sensory modality, Visual or Auditory, we randomly interleaved trials on which the last four items in a sequence replicated the first four, and trials on which items in the two half-sequences were independent of one another. Subjects had to judge whether the items were replicated or not, which allowed us to define performance in terms of d'. With Visual stimuli, the result nicely replicated the d' value reported by Gold et al. (2014), but Auditory sequences produced considerably higher d' values. In particular, subjects' ability to detect a within-sequence repetition of items was far better with Auditory sequences than with Visual sequences, mean d' values of 2.41 (SeM = 0.15) and 1.23 (SeM = 0.14), respectively ($t(11) = 4.24$, $p < .01$). This substantial difference between Auditory and Visual performance could have come from differences between the conditions that had been used to establish cross-modal matches and the conditions presented to our subjects. In particular, Marks (1974) asked subjects to match individual Auditory and Visual items with self-paced viewing and listening times, and under conditions that put no burden on subjects' memory. In contrast, our task not only imposed a considerable burden on subjects' short-term memory, but, more importantly, presented items in succession at a high rate (8 Hz). Welch and Warren's (1980) modality-appropriateness hypothesis suggests that our task's emphasis on temporal attributes of a sequence would advantage auditory processing over visual processing, as we found. Moreover, the large difference between d' values for Auditory and Visual sequences is consistent with the idea that sensory encoding of pitch sequences, in particular, is aided by special-purpose neural mechanisms responsive to frequency shifts (Cousineau et al., 2009).

Whatever its cause, the approximately twofold difference in d' values in our preliminary experiment suggests that if the unimodal stimuli from that experiment were combined in multisensory sequences, performance would be dominated by the sequences' Auditory components, rendering valid assessment of multisensory integration difficult. To avoid that likelihood while retaining the luminance range that Gold et al. (2014) used, we narrowed the range of auditory frequencies that would be used in the experiment proper. Specifically, the tones comprising auditory sequences were drawn from a range of 344–400 Hz. In musical terms, this reduced range of tones went from slightly below F4 to slightly above G4.

Within each block, stimulus sequences comprised two different structural categories. In some sequences, hereafter termed "Repeat" sequences, the last four items in the sequence repeated the first four items identically and in order; all items were reconstituted anew for each trial. In other sequences, hereafter called "Random" sequences, each item of the eight was the product of an independent sample (see Fig. 2); these sequences, too, were reconstituted anew for each trial. In each block of trials, Repeat and Random sequences were randomly intermingled, with both trial types occurring equally often.

Subjects attempted to identify whether halves of an eight-item stimulus sequence repeated or not, that is, whether a stimulus was a Repeat or Noise. Unimodal stimuli, Visual or Auditory, were presented in separate blocks of 75 trials each.

Fig. 3 presents schematic examples of both classes of multimodal stimuli, AVcongruent and AVincongruent. With multimodal sequences, subjects were instructed to

Please cite this article in press as: Aizenman AM et al. Multisensory integration in short-term memory: Musicians do rock. Neuroscience (2017), http://dx.doi.org/10.1016/j.neuroscience.2017.04.031

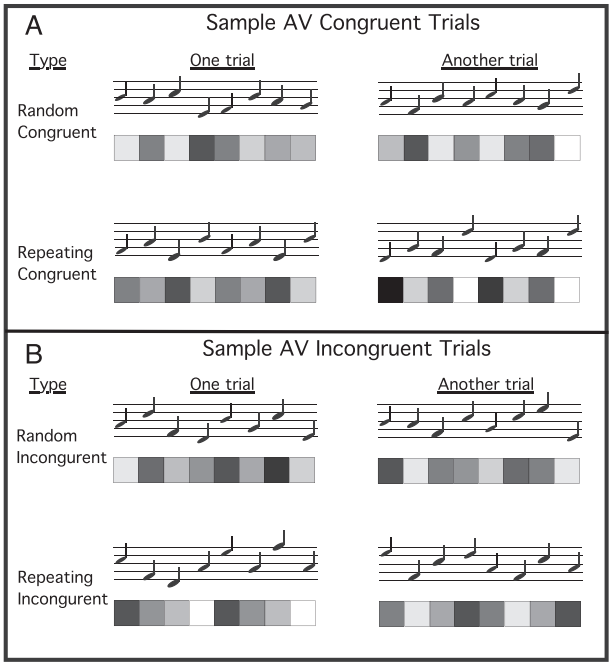

Fig. 3. Schematic examples of multimodal stimuli. Panel A: exemplars of stimuli whose audio and visual components were congruent; Panel B: exemplars of stimuli whose audio and visual components were incongruent, that is uncorrelated. In each panel, two examples are shown for the Random condition (the last four items in the eight-item visual sequence are uncorrelated with the first four) and for the Repeat condition (the last four items in the eight-item visual sequence repeat identically the first four).

ignore a sequence's auditory dimension, and to base judgments solely a sequence's visual attributes. In order to probe the limits of ability to ignore concurrent audio signals, we devised two classes of multimodal sequences: Congruent sequences in which variation in frequency was a monotone function of the accompanying luminance, and Incongruent sequences in which variation in frequency was uncorrelated with variation in luminance. All sequences comprised eight items presented in rapid succession, at 8 Hz. Finally, to assess whether musical training impacted performance with rapidly presented sequences, we tested equal numbers of subjects who had had music training and subjects who had not. Previous results on the processing of temporal sequences led us to hypothesize that music training would advantage subjects not only on auditory sequences, but with rapidly presented visual sequences as well (Deliége, 1996; Deliége et al., 1996). As it has long been known that concurrent co-modulation promotes integration or binding of auditory and visual signals (Thomas, 1941), we hypothesized that co-modulation would help subjects to recognize when items in a sequence were repeated.

The stimulus-generating algorithm required a second step in order to produce multisensory, Audiovisual stimuli whose Visual and Auditory components were incongruent. This step is represented in Fig. 2. Note that a second, "dummy" set of eight luminance samples was drawn from the zero-mean Gaussian distribution. The tonal equivalents to members of this new set were derived and substituted for the tonal equivalents to the actual luminances already set for that trial. The result was a set of frequencies that were uncorrelated with the set of luminances. To produce Repeat sequences, stimuli in which the last four items duplicated the first four, we simply discarded the sequence's last four items – whether unimodal or multimodal – and substituted for them exact copies of the first four items. With this last step, the algorithm could generate any of the stimulus types that the experiment required.

Audiovisual stimuli were presented in blocks of 200 trials. Within a block,AVcongruent and AVincongruent sequences of both Repeat and Random types occurred roughly equally often, in randomly intermixed fashion. The two unimodal conditions, Visual and Auditory, were presented in separate blocks of 100 trials each. Those one hundred trials were randomly intermingled Random and Repeat sequences. Each subject was tested in a total of six blocks of trials: two blocks devoted to Audiovisual trials, and two blocks devoted to each unimodal condition (Auditory and Visual). The order of the six blocks was counterbalanced over subjects.

Three hundred milliseconds after any stimulus sequence ended, a message on the screen prompted the subject for a key press that signaled whether elements in the sequence repeated. Feedback, in the form of a text message, followed. Subjects were encouraged to rest after every 50 trials, but were asked to remain seated throughout the experiment.

The order in which Auditory, Visual and Audiovisual trial blocks were presented was counterbalanced across subjects. Before beginning the experiment, subjects practiced with 20 trials of each stimulus type – Auditory, Visual, AVcongruent and AVincongruent.

Answers to a questionnaire about music training was used to constitute two groups, one comprising people who had musical training, and another comprising people with little or no training. The questionnaire was administered after testing, which kept subjects from knowing that we would be segregating data into groups. Following Skoe and Kraus (2012), a subject qualified as a "musician" if he or she had played one or more musical instruments for six or more years, and was continuing to play/practice an instrument up to the time of the experiment. A "non-musician" was defined as someone who either had never played a musical instrument or had played a musical instrument for three or fewer years, more than six years before study participation.[1] The musicians on average had 10.93 years of musical training.

Fourteen musicians and fourteen Non-musicians, all between the ages of 18 and 22 years of age, participated in this experiment. Each was compensated $10 (U.S.) per experimental session. Nine subjects in each group were female. Table 1 summarizes the history of musical training reported by subjects who qualified as Musicians. All subjects had normal acuity

[1] We recognize that merely having played an instrument for some time does not truly make someone a musician, at least as that term is usually used. However, the terms "musician" and "non-musician" are convenient, if imperfect surrogates.

Please cite this article in press as: Aizenman AM et al. Multisensory integration in short-term memory: Musicians do rock. Neuroscience (2017), http://dx.doi.org/10.1016/j.neuroscience.2017.04.031

Table 1. Gender and age at which musical training began, years of musical training and instrument(s) played by musically trained subjects. For subjects who reported playing multiple instruments, instruments are listed in order of earliest learned.

Subject	Gender	Starting Age (years)	Musical Training (years)	Instrument
1	F	10	12	Violin, Piano
2	M	4	15	Cello, Violin, Guitar
3	F	5	15	Violin
4	F	7	12	Piano, Clarinet, Bass Clarinet
5	F	4.5	15	Piano, Drums
6	M	6	15	Piano
7	F	7	12	Piano,Flute, Saxophone
8	F	7	9	Piano
9	M	5	9	Piano, Flute, Guitar
10	M	10	8	Saxophone, Bass, and Guitar
11	M	12	6	Guitar, Piano
12	F	9	6	Flute, Piano
13	F	7	8	Alto Saxophone, Violin
14	F	9	11	Piano, Flute, Guitar

and hearing, and had best-corrected visual acuity, measured using a Snellen chart, of at least 20/40. Hearing was indexed by a subject's pure tone average (*PTA*; the average threshold in each participant's better ear for 1, 2, and 4 kHz). All subjects' *PTA*s, as measured with a Beltone 120 audiometer, were $\leqslant$25 dB (HL), which qualifies as clinically normal hearing (Mueller and Hall, 1998). The experimental protocol was submitted to, and approved by Brandeis University's Committee for the Protection of Human Subjects.

RESULTS

Performance with various stimulus types was measured by the d' values produced by each subject. These values were defined as the difference between the standard scores associated with the proportions of "hits" (pr["repeat"—Repeat]) and "false positives" (pr["repeat"—Noise]). Fig. 4 shows musically trained and non-trained subjects' mean d' values for Auditory, Visual, AVcongruent and AVincongruent trials. These results were analyzed with separate ANOVAs on results from unimodal and multimodal stimuli. ANOVAs were followed by *t*-tests.

First, results from two types of unimodal stimuli, Auditory and Visual, did not reliably differ ($F_{1,26} = .009$, $p = 0.923$, $\eta^2_G = 0.0001$). This suggests that we achieved our goal of equating the two types of unimodal stimuli. Second, the two groups of subjects, Musicians differed in distinguishing unimodal random from unimodal repeating sequences, $F_{1,26} = 8.243$, $p = 0.008$, $\eta^2_G = 0.15$. This reflected the fact that Musicians outperformed Non-musicians, both with Auditory ($t(26) = 2.47$; $p = .02$) and with Visual ($t(26) = 2.03$; $p = .05$) sequences. Finally, the interaction between subject group and stimulus type failed to reach statistical significance $F_{1,26} = 2.169$, $p = 0.153$, $\eta^2_G = 0.034$.

Turning to Audiovisual stimuli, the ANOVA showed no significant overall difference between Musicians and Non-musicians ($F_{1,26} = 2.667$, $p = 0.115$, $\eta^2_G = 0.075$), although the congruency between Auditory and Visual components did matter: $F_{1,26} = 182.989$, $p < .00001$, $\eta^2_G = 0.594$). The interaction between subject group and stimulus congruency approached, but did not attain statistical significance: $F_{1,26} = 3.952$, $p < .057$, $\eta^2_G = 0.031$. This near-significant interaction term was explained by *t*-tests that compared the performance of Musicians and Non-musicians for AVcongruent and AVincongruent trials: the two groups differed on AVcongruent trials ($t(26) = 2.214$; $p = .036$), but not on AVincongruent trials ($t(26) = .587$; $p = .562$).

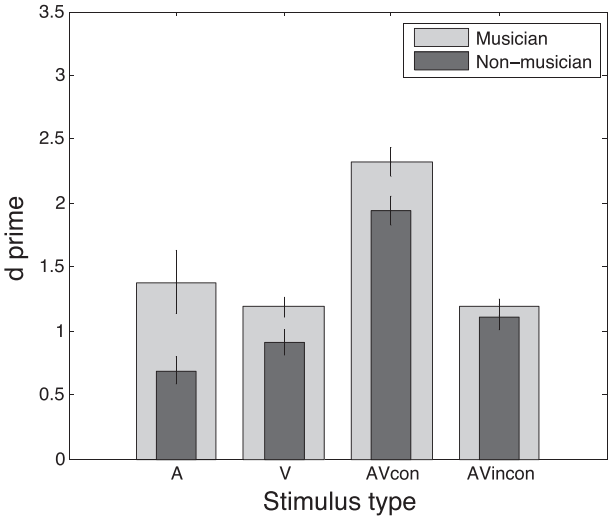

Fig. 4. Mean d' values. For each condition, the mean d' value for Musicians (lighter bar) is stacked atop the corresponding value for Non-musicians (darker bar). Error bars represent ±1 within-subject standard error.

The pattern of results shown in Fig. 4 led us to ask whether Musicians' advantage over Non-musicians with Visual stimuli was enhanced further when subjects were tested with stimuli that had Auditory components. For an answer, we computed two sets of difference scores by subtracting each subject's d' value for Visual sequences from that subject's d' value for Auditory sequences, and, separately, from that subject's d' value for AVcongruent sequences. Independent-samples' *t*-tests compared Musicians and Non-musicians on each set of difference scores. The advantage that Musicians showed with

Please cite this article in press as: Aizenman AM et al. Multisensory integration in short-term memory: Musicians do rock. Neuroscience (2017), http://dx.doi.org/10.1016/j.neuroscience.2017.04.031

Visual stimuli did not significantly increase, either with Auditory or AVcongruent stimuli ($p = 0.08$ and $p = 0.36$, respectively; $df = 26$, one-tailed t-tests). The corresponding test with results from AVincongruent sequences showed that Musicians and Non-musicians did not differ significantly ($p = .56$).

The panels in Fig. 5 show how subjects' performance with various kinds of stimulus sequences relates to years of musical training. In each panel, two best fit lines are shown. Each dashed line represents the best least squares linear fit based on subjects from both groups; each darker, unbroken line represents the best fit based on Musician subjects only. Here, we will summarize results for all subjects (corresponding values for only Musician subjects are given in Table 2). First, years of music training and d' were significantly correlated for both Auditory and Visual unimodal sequences, although for Auditory sequences there is considerable scatter, particularly among subjects who had the most years of musical training. For multimodal stimuli, number of years of music training was significantly correlated with performance on AVcongruent sequences, but not with AVincongruent sequences. Thus, with some, but not all kinds of stimulus sequences, performance is significantly related to years of music training.

Reverse correlation

To determine whether performance differences between Musicians and Non-musicians were associated with differences in subjects' processing strategies, we turned to reverse correlation analysis (also called "response classification"; Ahumada et al., 1975; Ahumada and Beard, 1997; Murray et al., 2002). This analytic technique computes the correlation between subjects' responses across trials and the contrast of each item in a sequence, where contrast is defined by the relationship between an item's luminance and the background luminance. The result is a set of weights that shows the relative influence exerted by each item on subjects' decisions. Recall that subjects were instructed to judge whether the last four items in a sequence did or did not identically repeat the first four items. Previously, when this analysis was applied to data from subjects who performed this same task, but only with visual stimuli, Gold et al. (2014) found that subjects disproportionately weighted items in particular ordinal positions within a stimulus sequence. Specifically, reverse correlation revealed that subjects gave particular weight to the final items in each half of an eight-item Visual sequence. To see what effect the addition of correlated and uncorrelated auditory sequences might have on this strategy, we performed the same analysis on Visual, Auditory, AVcongruent and AVincongruent data. Specifically, vectors containing the eight contrast values shown in the visual sequence on each trial were sorted according to the four possible stimulus–response combinations. The vectors were then averaged for each stimulus–response combination, and Eq. 1 was used to produce the mean kernel $\vec{c}$

$$\vec{c} = (\overline{rR} + \overline{rN}) - (\overline{nN} + \overline{nR}), \quad (1)$$

where xY denotes the combination of response x (either "repeating" or "not repeating") and stimulus Y (either Repeating or Random).

The result, $\vec{c}$, is an eight-element vector whose values are the relative weights assigned to items comprising the sequence that the subject tried to categorize. These values are best understood in terms of the signed contrasts of items in a sequence. An item is said to have positive contrast if its luminance is greater than the background; an item is said to have negative contrast if its luminance is less than the background. With those definitions in mind, a positive weight at position i in $\vec{c}$ indicates that positive contrast values at position i promotes a "repeat" response, while a negative contrast value promotes "non repeat" responses. Correspondingly, a negative weight at position i in $\vec{c}$ indicates that positive contrast values promote "non repeat" responses, while negative contrast values promotes "repeat" responses. Finally, if an observer's classification of a stimulus were not correlated with the stimulus value occupying a particular ordinal position in a sequence, the resulting mean kernel for that ordinal position would not significantly differ from zero.

Fig. 6A–C shows reverse correlations based on the luminance variation within sequences. Panel A shows results for unimodal Visual sequences; Panel B shows results for AVcongruent sequences; and Panel C shows results for AVincongruent sequences. Panel D shows reverse correlations for unimodal Auditory sequences, that is, sequences within which pitch varied, but had no accompanying luminance variation. In each panel, filled symbols (●) represent results for Musicians; open symbols (○) represent results for Non-musicians. The horizontal gray band in each plot shows the 95% confidence region centered on a reverse correlation of zero. The height of each region was computed from 2000 simulated experiments in which the sequence actually presented on each trial was replaced by a new random sample (Efron and Tibshirani, 1991; Gold et al., 2014).

Consider first the Visual-only sequences. The reverse correlation functions in Fig. 6A strongly resemble ones reported previously for similar Visual sequences (Gold et al., 2014). The fourth and eighth items in an eight-item sequence exert the strongest influence on subjects' judgments. Moreover, both Musicians and Non-musicians exhibit this pattern. Thus, the differences in d' values between Musicians and Non-musicians with Visual-only stimuli probably result from some other aspect of the way each group processes stimulus information, for example, differences in their internal noise levels (Burgess et al., 1981) or in their temporal uncertainty (Pelli, 1985).

Next, consider results with Audiovisual-Congruent stimuli. Unlike what was seen with Visual-only stimuli, here there is a marked difference between Musicians' and Non-musicians' strategies. In particular, Musicians appear to have maintained the same strategy that they used when no Auditory stream was present. In contrast, Non-musicians show no consistent preferential

Please cite this article in press as: Aizenman AM et al. Multisensory integration in short-term memory: Musicians do rock. Neuroscience (2017), http://dx.doi.org/10.1016/j.neuroscience.2017.04.031

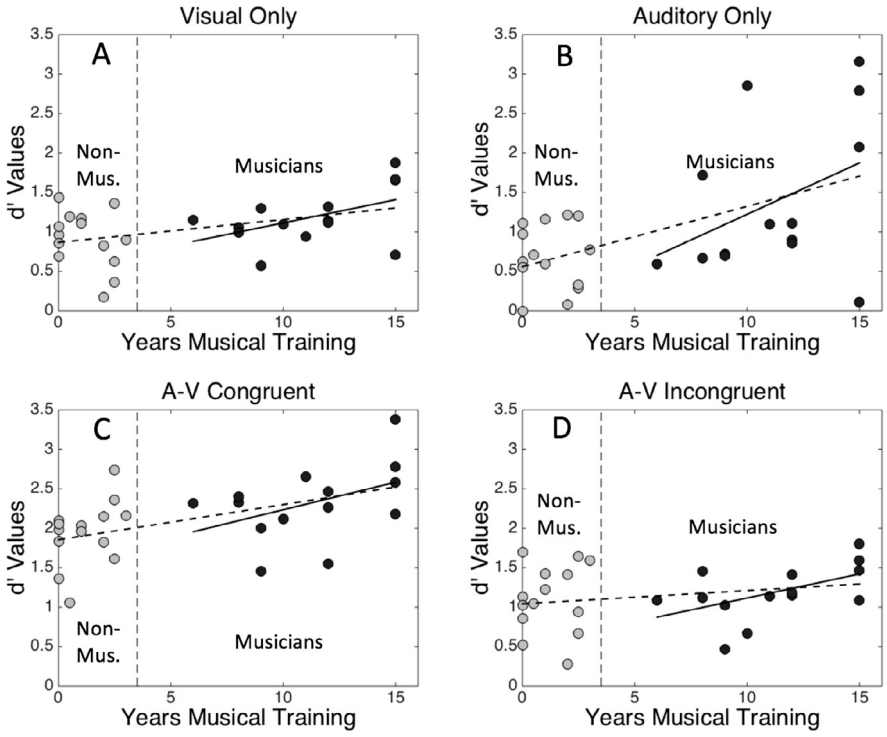

Fig. 5. *d'* values for individual subjects as a function of years of musical training. Results for unimodal sequences are shown in Panels A and B; results for multimodal sequences are shown in Panels C and D. In each panel, two best fit linear functions are show: one (dashed lines) is based on all subjects; the other (continuous lines) is based on data only from Musician subjects.

Table 2. Pearson *R* (and *p*) values between *d'* values for various stimulus types and years of music training. Note that *p* values are one-tailed.

Stimulus type	All subjects	Musicians only
Visual	0.423 (0.013)	0.486 (0.039)
Auditory	0.524 (0.002)	0.316 (0.136)
AVcongruent	0.522 (0.002)	0.440 (0.058)
AVincongruent	0.240 (0.109)	0.581 (0.015)

weighting for any item or items within the eight-item sequence. Gold et al. (2014) argued that their subjects gave extra weight to a sequence's fourth and eighth items in order to deal with intrinsic uncertainty about the temporal boundaries of the visual sequences they were seeing. They reasoned that knowledge of these temporal boundaries would facilitate comparisons between the two halves of a sequence. That interpretation predicts that subjects' performance would be reduced if they did not use such a strategy, which is exactly what we found. Thus, it appears that the presence of a correlated Auditory stream interferes with the ability of Non-musicians to maintain the strategy that they normally use to overcome the limiting effects of temporal uncertainty. Musicians, on the other hand, appear to be much less affected by the concurrent, correlated Auditory stream.

Finally, consider results with AVincongruent stimuli. Here, both Musicians and Non-musicians seem to change the strategy they adopted with Visual-only stimuli. Further, recall that with AVincongruent stimuli there was no significant difference between Musicians' and Non-musicians' performance. Apparently, the inclusion of an uncorrelated Auditory stream keeps subjects from differentially weighting particular visual items in a rapidly presented sequence.

DISCUSSION

Our results demonstrate that musical training is associated with enhanced ability to detect repetitions of items within rapidly presented sequences of several varieties. In particular, music training has some association with ability to detect repetition of items within Auditory sequences (judging tonal items), as well as within Visual sequences (judging luminance items). Music training is associated also with performance with visual items embedded in multimodal sequences, so long as the sequence's Auditory and Visual items are correlated. Interestingly, this effect does not extend to stimulus sequences whose concurrent items are uncorrelated. So we can add these advantages to others already known to result from music training.

Lest non-musician readers of this paper rush out to acquire music training, we are obliged to reflect on the distinction between correlation and causation. The mere fact that Musicians outperform Non-musicians on some task does not mean that music training *per se* was responsible for that difference. After all, it could be that people who, absent training, would show greater facility on the task, and would be more inclined to initiate and continue training. In our study, a more dispositive result is presented in Fig. 5B, which suggests that Musicians' advantage with Auditory sequences is significantly related to years of training. Although the correlation between years of training and performance is suggestive, it is not dispositive. For example, a pre-existing talent for processing Auditory sequences might encourage someone not only to initiate, but also to persist in learning to play an instrument. Given the definition of "Musician" that we and others have adopted (Kraus and Chandrasekaran, 2010; Bergstrom et al., 2012), a proper experimental test of a potential causal link between music training, on one hand, and performance on a task like ours, on the other, would require taking subjects who had never had music training, randomly assigning some subjects to undergo music training for an extended period, while control subjects received equivalent non-music training for the same period (Barrett et al., 2013). Ideally, the effect of differential training would be assessed not only via differential behavioral

Please cite this article in press as: Aizenman AM et al. Multisensory integration in short-term memory: Musicians do rock. Neuroscience (2017), http://dx.doi.org/10.1016/j.neuroscience.2017.04.031

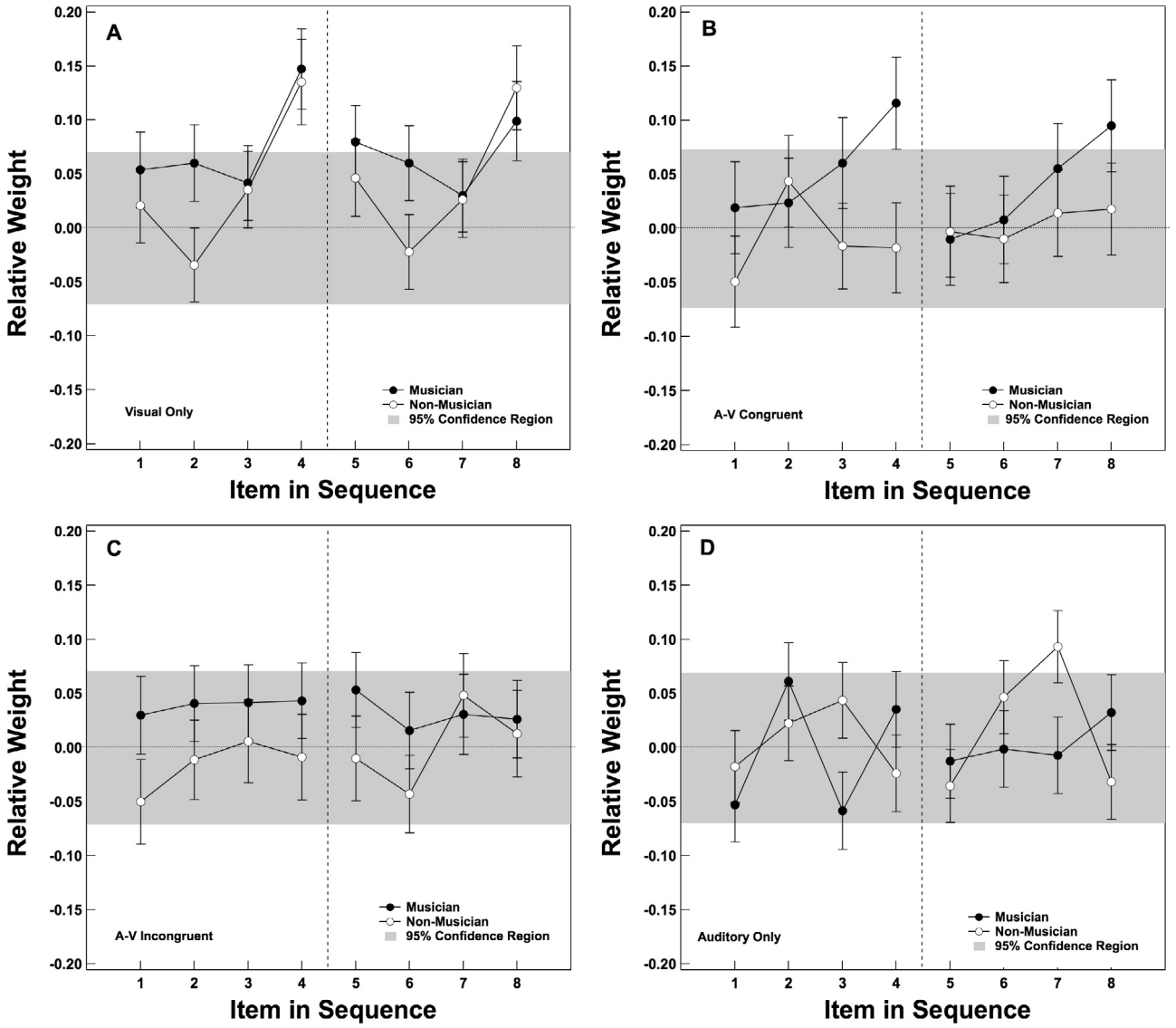

Fig. 6. Panels A–C: Reverse correlations based on a sequence's Visual attributes. In each panel, separate curves are shown for Musician (●) and Non-musician (○) subjects. Panel D: reverse correlations for unimodal Auditory sequences. Panel A shows results for Visual unimodal sequences; Panel B shows results for AVcongruent sequences; Panel C shows results for AVincongruent sequences, and Panel D shows results for Auditory unimodal sequences. Error bars around each data point are ±1 standard deviations of 2000 bootstrap samples. These empirical standard deviations of bootstrapped replications approximate ±1 standard errors of the sample statistic itself. For each bootstrap sample, the complete data set represented in a panel was resampled with replacement. The horizontal gray bands in each panel represent 95% confidence regions whose derivation is explained in the text. Results have not been corrected for multiple comparisons.

changes, but also in terms of correlated changes in the brain. To our knowledge, only one study meets these stringent criteria. In that study, Hyde et al. (2009) formed two groups of children who were a bit over six years old at the start of the 15-month study. One group received weekly, 30-min private keyboard lessons during the study; a second group received no instrumental music training, but instead participated in weekly 40-min group music classes, during which they sang and played with drums and bells. Comparisons of pre- and post-study measures showed that instrumental training differentially enhanced ability to distinguish between pairs of five-tone musical phrases that differed either in melody, that is, in pitch sequence, or in rhythm (Overy et al., 2004). Analysis of magnetic resonance images captured at the start and end of the study revealed that training-induced improvement on the melodic/rhythmic discrimination test were correlated with deformation changes in a key auditory area of subjects' right hemispheres, that is, the lateral aspect of Heschl's gyrus.

There are obvious limitations to the criteria we adopted in defining Musician and Non-musician. In fact, the challenge of precisely defining what constitutes a "musician" is well-known (Levitin, 2012). Following the lead of Skoe and Kraus (2012), we defined musical training mainly by the time for which subjects self-reported that they had played an instrument. Needless to say, not every single person who takes many years of music instruction

Please cite this article in press as: Aizenman AM et al. Multisensory integration in short-term memory: Musicians do rock. Neuroscience (2017), http://dx.doi.org/10.1016/j.neuroscience.2017.04.031

or engages in years of continuous practice achieves a level of proficiency that would satisfy commonly held definitions of ‘‘musician.” Conversely, some individuals might possess sufficient talent that he or she could achieve a very high level of proficiency in very short order. Additionally, there are undoubtedly multiple differences between the two groups we tested, such as differences in beat perception (Grahn and Rowe, 2009), auditory imagery (Brown and Palmer, 2013; Keller et al., 2010), perceptual grouping (Kung et al., 2011) or the ability to detect duration or pitch deviants in a sequence (Seppänen et al., 2013), and we did not directly test Musicians and Non-musicians for differences in these skills. By assessing the perception of multiple dimensions of auditory experience, future studies could make a start toward a more complete definition of musicianship.

Of course, playing a musical instrument does not require processing quasi-random luminance sequences like the ones with which we tested subjects. However, music training can entail translating visual information into temporal sequences, as one does, for example, in learning to read music. This form of learning might build on the spontaneous mapping of pitch onto the visual feature of vertical location (Evans and Treisman, 2010). Mindful of the role that visual information could play in music training, Bergstrom et al. (2012) tested the speed and accuracy as subjects made key presses to each of a series of targets presented at different locations on a computer screen. Unbeknown to the subjects, embedded in the sequence was a sub-sequence in which events' locations were governed by the rules of an artificial grammar (e.g., Reber and Millward, 1968). Using much the same definition of ‘‘Musician” as we did, Bergstrom et al. (2012) showed that compared to Non-musicians, Musicians exhibited greater implicit learning of sequential regularities. This suggests that the skills music training is associated with, includes the skill of implicitly learning and remembering quasi-random visual-temporal sequences.

Fig. 4 shows that even though subjects were instructed to focus exclusively on the Visual aspect of any Audiovisual sequence, the presence of a concurrent, congruent Auditory sequence boosted performance considerably over what was seen with either unimodal sequence alone. Although we lack the parametric data that would be needed for formal model selection, it is useful to compare this Audiovisual effect against what would be expected from one simple, widely used benchmark. Imagine that two orthogonal signals were processed by independent mechanisms, *A* and *V*, whose noise was uncorrelated. Under such conditions, with each sensitivity expressed as d', the response to the combination of the two signals would be $\sqrt{d'V^2 + d'A^2}$ (Green and Swets, 1966; Green, 1958; Viemeister and Wakefield, 1991). A *t*-test confirmed that AVcongruent sequences boosted performance well above the predicted value ($t_{26} = -3.46$, $p < 0.001$). For the sake of completeness, we also compared performance with AVincongruent sequences against performance with each type of unimodal sequence. Neither comparison was statistically significant (p = .51 and.33, for *t*-tests against Auditory and Visual sequences, respectively). Returning to the surprisingly powerful advantage seen with AVcongruent sequences, it should be noted that the super-additivity of Auditory and Visual components in such sequences was produced despite the fact that those unimodal aspects were designed to be strongly correlated, that is, distinctly non-orthogonal. As this surprising result may be valuable in informing theories of multisensory integration, the boundary conditions on this result deserve further study. For example, it may be this apparent super-additivity reflects the engagement of mechanisms specialized for multisensory coincidence or congruence (e.g., Bushara et al., 2003; Kayser et al., 2010; Orchard-Mills et al., 2013). Alternatively, it might be that temporal information arising from the succession of auditory items helps subjects parse the accompanying visual items, thereby making it easier to identify whether visual items were replicated or not. Obviously, further experiments would be required to select among these, and possibly other, alternatives.

Earlier, we cited one recent report that music training did not seem to impact visual memory. However, differences between that study and our own are worth considering. For stimuli in their study, Cohen et al. (2011) chose pictures of objects, speech clips and clips of familiar music. Stimuli were presented one at time, each for five seconds. After all the stimuli of one class had been presented, the researcher tested recognition memory by presenting intermixed old (previously presented) and new (novel) stimuli, noting how well subjects correctly categorized these intermixed stimuli as ‘‘old” or ‘‘new”. With either kind of auditory stimuli (speech or music clips), recognition memory was significantly poorer than recognition memory for the visual stimuli (pictures). More importantly for the present discussion, Musicians and Non-musicians did not differ in recognition memory for the visual stimuli. Multiple differences between tasks make it difficult to compare Cohen et al.'s (2011) results to the ones we report. These differences include (i) the types of stimuli used (low-level, elemental features *vs.* higher level stimuli, such as familiar tunes), (ii) the temporal characteristics of stimulus presentation (rapid presentation of item sequences, which worked against online rehearsal, *vs.* five seconds per individual item), and (iii) the task (recognizing within-trial repetitions of items *vs.* longer term recognition of single items). Although any or all of these differences could account for the difference between Cohen et al.'s (2011) result and our own, when researchers try to assess music-training's impact on visual memory, the modality appropriateness hypothesis (Welch and Warren, 1980) should influence their choice of stimuli and test task.

In our study, Audiovisual congruence was defined by a positive monotone function relating an item's luminance and the frequency of an accompanying tone. Of course, Audiovisual congruence could take other forms (see review in Evans and Treisman, 2010). One obvious form would exploit the normal Audiovisual congruence characteristic of speech production. Speech production involves movements of the mouth and face, which produces a reliable correlation between the auditory output of the vocal tract, on one hand, and visual motion

Please cite this article in press as: Aizenman AM et al. Multisensory integration in short-term memory: Musicians do rock. Neuroscience (2017), http://dx.doi.org/10.1016/j.neuroscience.2017.04.031

cues, on the other. It has long been known that speech-related visual cues alter the intelligibility and detectability of heard speech (e.g., Campbell, 2008). In fact, the congruence between a speaker's mouth and lip movements and the accompanying sound is the basis of the well-known McGurk–MacDonald effect (McGurk and MacDonald, 1976) in which altering the normal relationship between a spoken sound and the accompanying movements of the mouth distorts a listener's perception of that sound. Face-to-face speech can be described as a phenomenon that is inherently multisensory (Chandrasekaran et al., 2013; Chandrasekaran et al., 2011).

It is noteworthy that this form of Audiovisual congruence extends to situations seemingly far removed from face-to-face speech. In particular, a related form of Audiovisual congruence has been incorporated into a first-person fisherman computer game (Goldberg et al., 2015; Sun et al., 2017) in which subjects categorized fish on the basis of the rate at which their size modulated. Responses to computer-generated, swimming fish were considerably speeded when the amplitude modulation of a sound emitted by a fish was correlated with the periodic fluctuations in the fish's size. Additionally, in the present study Audiovisual congruence between components in an Audiovisual sequence could be described as all-or-none: while components of an AVcongruent sequence were perfectly correlated, components within an AVincongruent sequence were completely unrelated (on average). This was reminiscent of the arrangement Agus et al. (2010) used to construct their unimodal auditory stimuli. On each trial in their study, the auditory noise samples comprising one half of the stimulus were either replicated identically in the second half, making the within-stimulus correlation 1.0, or were paired with new, randomly generated noise samples, making the within-stimulus correlation 0.0. Of course, one can imagine sequences in which correlation, either within unimodal sequences, and/or between separate multimodal components, is neither 1.0 nor 0.0, but some value between these extremes. Such partially correlated stimuli could be leveraged to identify what strategies subjects use in an Audiovisual paradigm like ours.

REFERENCES

Agus TR, Thorpe SJ, Pressnitzer D (2010) Rapid formation of robust auditory memories: insights from noise. Neuron 66:610–618.

Ahumada Jr AJ, Beard BL (1997) Image discrimination models predict detection in fixed but not random noise. J Opt Soc Am A Opt Image Sci Vis 14:2471–2476.

Ahumada Jr AJ, Marken R, Sandusky A (1975) Time and frequency analyses of auditory signal detection. J Acoust Soc Am 57:385–390.

Barrett KC, Ashley R, Strait DL, Kraus N (2013) Art and science: how musical training shapes the brain. Front Psychol 4:713.

Berger CC, Ehrsson HH (2013) Mental imagery changes multisensory perception. Curr Biol 23:1367–1372.

Bergstrom JCR, Howard JH, Howard DV (2012) Enhanced implicit sequence learning in college-age video game players and musicians. Appl Cogn Psychol 26:91–96.

Brainard DH (1997) The psychophysics toolbox. Spatial Vis 10:433–436.

Brown RM, Palmer C (2013) Auditory and motor imagery modulate learning in music performance. Front Human Neurosci 7:320.

Burgess AE, Wagner RF, Jennings RJ, Barlow HB (1981) Efficiency of human visual signal discrimination. Science 214:93–94.

Bushara KO, Hanakawa T, Immisch I, Toma K, Kansaku K, Hallett M (2003) Neural correlates of cross-modal binding. Nat Neurosci 6:190–195.

Campbell R (2008) The processing of audio-visual speech: empirical and neural bases. Philos Trans Royal Soc Lond B Biol Sci 363:1001–1010.

Chan AS, Ho YC, Cheung MC (1998) Music training improves verbal memory. Nature 396:128.

Chandrasekaran C, Lemus L, Ghazanfar AA (2013) Dynamic faces speed up the onset of auditory cortical spiking responses during vocal detection. Proc Natl Acad Sci U S A 110:E4668–E4677.

Chandrasekaran C, Lemus L, Trubanova A, Gondan M, Ghazanfar AA (2011) Monkeys and humans share a common computation for face/voice integration. PLoS Comput Biol 7:e1002165.

Chen YC, Spence C (2010) When hearing the bark helps to identify the dog: semantically-congruent sounds modulate the identification of masked pictures. Cognition 114:389–404.

Cohen MA, Evans KK, Horowitz TS, Wolfe JM (2011) Auditory and visual memory in musicians and nonmusicians. Psychon Bull Rev 18:586–591.

Cousineau M, Demany L, Pressnitzer D (2009) What makes a melody: the perceptual singularity of pitch sequences. J Acoust Soc Am 126:3179–3187.

Deliége I (1996) Cue abstraction as a component of categorisation processes in music listening. Psychol Music 24:131–156.

Deliége I, Mélen M, Stammers D, Cross I (1996) Musical schemata in real time listening to a piece of music. Music Percept Interdiscip J 14:117–160.

Efron B, Tibshirani R (1991) Statistical data analysis in the computer age. Science:253.

Evans KK, Treisman A (2010) Natural cross-modal mappings between visual and auditory features. J Vis 10:11507–11510.

Francois C, Schön D (2011) Musical expertise boosts implicit learning of both musical and linguistic structures. Cereb Cortex 21:2357–2365.

Gold JM, Aizenman A, Bond SM, Sekuler R (2014) Memory and incidental learning for visual frozen noise sequences. Vis Res 99:19–36.

Goldberg H, Sun Y, Hickey TJ, Shinn-Cunningham B, Sekuler R (2015) Policing fish at boston's museum of science: studying audiovisual interaction in the wild. iPerception 6. 2041669515599332.

Grahn JA, Rowe JB (2009) Feeling the beat: premotor and striatal interactions in musicians and nonmusicians during beat perception. J Neurosci 29:7540–7548.

Green DM (1958) Detection of multiple component signals in noise. J Acoust Soc Am 50:904–911.

Green DM, Swets JA (1966) Signal detection theory and psychophysics. New York: Wiley.

Guttman S, Gilroy LA, Blake R (2005) Hearing what the eyes see: auditory encoding of visual temporal sequences. Psychol Sci 16:228–235.

Hyde KL, Lerch J, Norton A, Forgeard M, Winner E, Evans AC, Schlaug G (2009) Musical training shapes structural brain development. J Neurosci 29:3019–3025.

Kahana MJ, Sekuler R (2002) Recognizing spatial patterns: a noisy exemplar approach. Vis Res 42:2177–2192.

Kayser C, Logothetis NK, Panzeri S (2010) Visual enhancement of the information representation in auditory cortex. Curr Biol 20:19–24.

Keller PE, Dalla Bella S, Koch I (2010) Auditory imagery shapes movement timing and kinematics: evidence from a musical task. J Exp Psychol Human Percept Performance 36:508–513.

Please cite this article in press as: Aizenman AM et al. Multisensory integration in short-term memory: Musicians do rock. Neuroscience (2017), http://dx.doi.org/10.1016/j.neuroscience.2017.04.031

Kraus N, Chandrasekaran B (2010) Music training for the development of auditory skills. Nat Rev Neurosci 11:599–605.
Kung SJ, Tzeng OJL, Hung DL, Wu DH (2011) Dynamic allocation of attention to metrical and grouping accents in rhythmic sequences. Exp Brain Res 210:269–282.
Levitin DJ (2012) What does it mean to be musical? Neuron 73:633–637.
Magnussen S (2000) Low-level memory processes in vision. Trends Neurosci 23:247–251.
Marks LE (1974) On associations of light and sound: the mediation of brightness, pitch and loudness. Am J Psychol 87:173–188.
McGurk H, MacDonald J (1976) Hearing lips and seeing voices. Nature 265:746–748.
Michalka SW, Kong L, Rosen ML, Shinn-Cunningham BG, Somers DC (2015) Short-term memory for space and time flexibly recruit complementary sensory-biased frontal lobe attention networks. Neuron 87:882–892.
Miller MB, Gazzaniga MS (1998) Creating false memories for visual scenes. Neuropsychologia 36:513–520.
Mueller G, Hall JW (1998). Audiologist's desk reference: audiolologic management, rehabilitation and terminology, vol. II. Singular Publishing Group Inc..
Murray RF, Bennett PJ, Sekuler AB (2002) Optimal methods for calculating classification images: weighted sums. J Vis 2:79–104.
Nisbett R, Wilson TD (1977) Telling more than we can know: verbal reports on mental processes. Psychol Rev 84:231–259.
Orchard-Mills E, Leung J, Burr D, Morrone MC, Wufong E, Carlile S, Alais D (2013) A mechanism for detecting coincidence of auditory and visual spatial signals. Multisens Res 26:333–345.
Overy K, Norton AC, Cronin KT, Gaab N, Alsop DC, Winner E, Schlaug G (2004) Imaging melody and rhythm processing in young children. NeuroReports 15:1723–1726.
Oxenham AJ, Fligor BJ, Mason CR, Kidd Jr G (2003) Informational masking and musical training. J Acoust Soc Am 114:1543–1549.
Pasternak T, Greenlee MW (2005) Working memory in primate sensory systems. Nat Rev Neurosci 6:97–107.
Pelli DG (1985) Uncertainty explains many aspects of visual contrast detection and discrimination. J Opt Soc Am A 2:1508–1532.
Rammsayer T, Altenmüller E (2006) Temporal information processing in musicians and nonmusicians. Music Percept Interdiscip J 24:37–47.
Reber AS, Millward RB (1968) Event observation in probability learning. J Exp Psychol 77:317–327.
Sekuler R, Sekuler AB, Lau R (1997) Sound alters visual motion perception. Nature 385:308.
Seppänen M, Häläinen J, Pesonen AK, Tervaniemi M (2013) Passive sound exposure induces rapid perceptual learning in musicians: event-related potential evidence. Biol Psychol 94:341–353.
Skoe E, Kraus N (2012) A little goes a long way: How the adult brain is shaped by musical training in childhood. J Neurosci 34:11507–11510.
Strait DL, Parbery-Clark A, Hittner E, Kraus N (2012) Musical training during early childhood enhances the neural encoding of speech in noise. Brain Language 123:191–201.
Sun Y, Shinn-Cunningham B, Hickey TJ, Sekuler R (2017) Catching audiovisual interactions with a first-person fisherman video game. Perception. http://dx.doi.org/10.1177/0301006616682755.
Thomas G (1941) Experimental study of the influence of vision on sound localization. J Exp Psychol 28:167–177.
Viemeister NF, Wakefield GH (1991) Temporal integration and multiple looks. J Acoust Soc Am 90:858–865.
Welch RB (1999) Meaning, attention and the unity assumption in the intersensory bias of spatial and temporal perceptions. In: Aschersleben G, Bachmann T, Müsseler J, editors. Cognitive contributions to the perception of spatial and temporal events. Amsterdam: Elsevier. p. 371–387.
Welch RB, Warren DH (1980) Immediate perceptual response to intersensory discrepancy. Psychological Bulletin 88:638–667.

(Received 1 April 2017, Accepted 20 April 2017)

Please cite this article in press as: Aizenman AM et al. Multisensory integration in short-term memory: Musicians do rock. Neuroscience (2017), http://dx.doi.org/10.1016/j.neuroscience.2017.04.031