Atten Percept Psychophys (2016) 78:1512–1528
DOI 10.3758/s13414-016-1088-5

The unity assumption facilitates cross-modal binding of musical, non-speech stimuli: The role of spectral and amplitude envelope cues

Lorraine Chuen[1] · **Michael Schutz**[1,2]

Published online: 15 April 2016

Abstract An observer's inference that multimodal signals originate from a common underlying source facilitates cross-modal binding. This 'unity assumption' causes asynchronous auditory and visual speech streams to seem simultaneous (Vatakis & Spence, *Perception & Psychophysics, 69*(5), 744–756, 2007). Subsequent tests of non-speech stimuli such as musical and impact events found no evidence for the unity assumption, suggesting the effect is speech-specific (Vatakis & Spence, *Acta Psychologica, 127*(1), 12–23, 2008). However, the role of amplitude envelope (the changes in energy of a sound over time) was not previously appreciated within this paradigm. Here, we explore whether previous findings suggesting speech-specificity of the unity assumption were confounded by similarities in the amplitude envelopes of the contrasted auditory stimuli. Experiment 1 used natural events with clearly differentiated envelopes: single notes played on either a cello (bowing motion) or marimba (striking motion). Participants performed an un-speeded temporal order judgments task; viewing audio-visually matched (e.g., marimba auditory with marimba video) and mismatched (e.g., cello auditory with marimba video) versions of stimuli at various stimulus onset asynchronies, and were required to indicate which modality was presented first. As predicted, participants were less sensitive to temporal order in matched conditions, demonstrating that the unity assumption can facilitate the perception of synchrony outside of speech stimuli. Results from Experiments 2 and 3 revealed that when spectral information was removed from the original auditory stimuli, amplitude envelope alone could not facilitate the influence of audiovisual unity. We propose that both amplitude envelope and spectral acoustic cues affect the percept of audiovisual unity, working in concert to help an observer determine when to integrate across modalities.

Keywords Multisensory processing · Music cognition · Audition · Audiovisual integration · Amplitude envelope

Introduction

Our perceptual experience of environmental events is inherently multisensory. To reduce variance in perceptual estimates, observers combine redundant information from different modalities into a single unified percept (Ernst & Banks, 2002). Low-level stimulus features, such as the degree to which amodal properties are shared between modalities, influence multisensory binding (Welch, 1999). For instance, observers are more likely to integrate temporally coincident (Bertelson & Aschersleben, 1998; Radeu & Bertelson, 1987) and/or spatially proximate (Gepshtein, Burge, Ernst, & Banks, 2005; Körding et al., 2007; Slutsky & Recanzone, 2001) cross-modal signals. Top-down processes also play an important role in facilitating integration, where semantic congruency regarding the signals' underlying event (Vatakis & Spence, 2007), synesthetic correspondences (Parise & Spence, 2009), and even learned statistical associations between otherwise arbitrary signals (Ernst, 2007), facilitate integration. These factors contribute to the perceptual system's decision as to whether multimodal signals originate from common underlying causes or events—a process known as the *unity assumption* (Welch, 1999).

✉ Michael Schutz
schutz@mcmaster.ca

[1] Department of Psychology, Neuroscience, and Behaviour, McMaster University, Hamilton, Ontario, Canada

[2] School of the Arts, Department of Psychology, Neuroscience, and Behaviour, McMaster University, 424 Togo Salmon Hall, Hamilton, Ontario, Canada

Although much audiovisual integration research focuses on integration in the spatial domain, the unity assumption also facilitates audiovisual integration in the temporal domain. For instance, *temporal ventriloquism* is a process by which a cross-modal temporal conflict (asynchrony) is resolved through cross-modal attraction in time (Bertelson & Aschersleben, 2003; Aschersleben & Bertelson, 2003; Morein-Zamfir, Soto-Faraco, & Kingston, 2003). The temporal window in which we are tolerant to cross-modal asynchrony is malleable—it can be widened by bottom up factors such as repeated exposure to asynchronous stimuli (Navarra et al., 2005), and shortened through experience such as long-term musical training (Lee & Noppeney, 2011) or short-term perceptual learning (Powers, Hillock & Wallace, 2009). Here we explore how this temporal window can also be influenced by the *unity assumption*.

An observer's sensitivity to cross-modal asynchrony can be measured through a temporal order judgment (TOJ) task, where participants determine the order of asynchronous signals from different modalities. One compelling example of the unity assumption's influence on temporal ventriloquism is the finding that observers are less sensitive in judging the temporal order of asynchronous audiovisual speech stimuli when the face and voice are gender-matched, compared to when they are gender-mismatched (Vatakis & Spence, 2007). An observer's assumption that audiovisual signals belong to the same underlying event causes stronger cross-modal coupling, reducing sensitivity to audiovisual lags in congruent conditions. Comparable effects of audiovisual congruency (i.e., unity) can be found for simple stimuli with synesthetic correspondences (Parise & Spence, 2009), for semantically matched speech/gesture pairs (Margiotoudi, Kelly & Vatakis, 2014), and for bouncing human-like point light displays paired with rhythmic up- and down-beat tones (Su, 2014).

However, findings conducted with simple tone-light pairings do not necessarily generalize to more complex stimuli. Natural stimuli such as the speech used by Vatakis & Spence (2007) differ from such pairings in many ways: they are more realistic, more dynamic, longer in duration, and richer in information. Given these differences, the study of complex stimuli is important as it allows for a greater understanding of multisensory processing in naturalistic environments (De Gelder & Bertelson, 2003).

Within the range of complex stimuli, the influence of the unity assumption on temporal ventriloquism has only been documented for speech (Vatakis & Spence, 2007), and for gestures paired with speech (Margiotoudi et al., 2014). Furthermore, explorations using non-speech stimuli such as musical and object/action events suggest the unity assumption may be confined to speech (Vatakis & Spence, 2008). For instance, temporal order sensitivity was no different when viewing "congruent" stimuli with a guitar pluck motion paired with a guitar pluck sound compared to viewing "incongruent" stimuli with the same motion paired with a note played on the piano. The influence of audiovisual unity on temporal binding was also not replicated when stimuli consisted of monkey calls or humans imitating monkey calls (Vatakis, Ghazanfar, & Spence, 2008).

From these null results, Vatakis & Spence (2008) proposed that the influence of the unity assumption on temporal audiovisual integration might be unique to speech. The authors suggest that this specificity might reflect the special nature of speech perception, perhaps due to our heightened familiarity and exposure to this stimulus class or the greater temporal correlation between auditory and visual signals in dynamic speech (Vatakis & Spence, 2008). This 'speech is special' hypothesis is consistent with a body of research suggesting that perception of speech signals is specialized (Baart, Stekelenburg, & Vroomen, 2014; Jones & Jarick, 2006; Tuomainen, Andersen, Tiippana, & Sams, 2005), and presents a plausible explanation for the null results found in experiments conducted with both with complex music and object-action stimuli.

However, there is reason to believe that the influence of the unity assumption on cross modal binding is not specific for speech. As previously mentioned, research indicates the influence of audiovisual unity for non-speech, synesthetically congruent simple stimuli (Parise & Spence, 2009) as well as the influence of unity and semantic relatedness on the audiovisual integration of speech/gesture pairs Margiotoudi, et al. (2014). A somewhat different argument against the 'speech is special' hypothesis by Vroomen & Stekelenberg (2010) suggests that differences between speech and non-speech in temporal ventriloquism as a function of audiovisual congruency have less to do with higher-order percepts of 'unity', and more to do with low-level differences between these stimulus classes. For instance, when presented with audiovisual sine-wave speech stimuli, audiovisual temporal sensitivity is no different for participants regardless of whether they perceive the stimuli as speech, or sine-waves (Vroomen & Stekelenberg, 2010).

In order to clarify the role of the unity assumption in non-speech stimuli, we consider the possibility that previous null results could instead be explained by the limitations imposed by the stimuli used in Vatakis and Spence's (2008) studies. Given the high degree of perceptual similarity between audiovisual conditions (Vatakis & Papadelis, 2014), it is possible that a sense of unity existed even in what was labeled the "mismatched" condition. The guitar auditory stimulus used by Vatakis and Spence (2008) might have been perceived as consistent with the piano visual stimulus, given that they share an important similarity—both guitar and piano sounds are produced by impacting a string. This similarity in event origin leads to an acoustic similarity with respect to *amplitude envelope*: the changes in energy of a sound over time. Singular piano and guitar

notes both exhibit a rapid attack, followed by an exponential decay that is characteristic of most impact, percussive sounds (Fig. 1).

We propose that amplitude envelope plays an important role in facilitating integration of auditory and visual signals, because it provides useful event-related information that allows an observer to evaluate whether a sound should be integrated with an accompanying visual signal. Changes in amplitude over time provide rich information about events in our environment, and are used as source identification cues, allowing us to differentiate breaking from bouncing (Warren & Verbrugge, 1984) as well as shaping how we perceive material properties of a sounding object (Klatzky, Pai & Krotkov, 2000).

The event-related information provided by amplitude envelope may explain why this parameter also influences how (and whether) we integrate audiovisual stimuli. For example, in the *marimba illusion*, the length of a musician's visual gesture influences the perceived auditory duration of the accompanying sound: a listener who watches a longer sweep of the arm will hear a longer note played than when watching a short arm movement (Schutz & Lipscomb, 2007). This illusion requires congruency between sight and sound. The percussive sound produced by the marimba integrates with the gestures only when they mimic impact motions, and fail to integrate with motions of similar speed that do not imply impacts (Armontrout, Schutz & Kubovy, 2009). Similarly, the illusion occurs with sounds synthesized using decaying amplitude envelopes (resembling impact sounds), but not spectrally matched sounds using flat amplitude envelopes (Schutz, 2009).

The novel pattern of integration found in the marimba illusion derives from importance of the unity assumption, rather ambiguity in the decaying auditory information. Although the moment of acoustic completion is clearer for flat vs. decaying sounds, duration assessments of decaying tones such as those produced by the marimba and piano are no more ambiguous than those from more abruptly ending sounds (Schutz & Kubovy, 2009). Similarly, pure tones synthesized with gradually decaying amplitude envelopes are no more ambiguous than those with abruptly ending envelopes (Schutz, 2009). These counter-intuitive findings can be explained by the use of different duration assessment strategies for decaying tones affording the ability to predict the moment of completion, and abruptly ending flat tones requiring detection of an explicit moment of offset (Vallet, Shore, & Schutz, 2014).

Amplitude envelope's role in audio-visual integration can also be seen in a variation of The *Metzger Motion Display*, a bi-stable display in which two disks streaming past one another can either be perceived as streaming or bouncing (Metzger, 1934). Bounce percepts increase significantly more when a decaying, impact-like sound occurs concurrent with the dots' overlap than when this sound increases in amplitude by a similar amount (Grassi & Casco, 2009). This suggests that amplitude envelope, in acting as a cue for event identification, helps facilitate integration of auditory and visual signals.

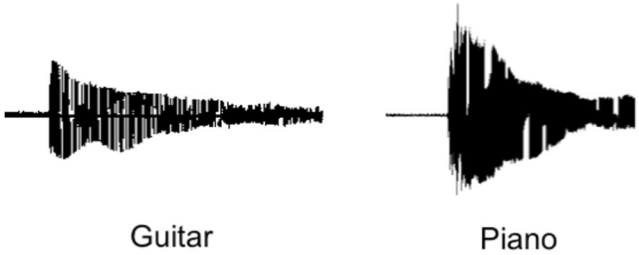

Fig. 1 Amplitude envelope of guitar and piano notes. Generated from stimuli used by Vatakis & Spence (2008), with permission from A. Vatakis

The present study explores amplitude envelope's role in triggering the perception of audiovisual unity, linking the sound produced by a particular object to an action consistent with the sound. Building on the TOJ paradigm used by Vatakis and Spence (2007, 2008), here we investigate the role of the unity assumption on temporal cross-modal binding for realistic audiovisual stimuli consisting of musical instruments. We chose to use music-related stimuli rather than generalized object-action stimuli, as sounds produced by musical instruments are wider in range, allowing us to more conveniently manipulate amplitude envelope as a parameter of interest. To assess amplitude envelope's crucial role in cross modal binding, we contrast the striking of a marimba to the bowing of a cello. Whereas the long bowing gesture of a cello produces a relatively flat amplitude envelope with a period of sustain, the impact of a mallet on a marimba produces a decaying amplitude envelope characteristic of impact percussive sounds (Figure 1).

As in Vatakis and Spence's (2007, 2008) experiments, participants in this study were required to indicate which modality was presented first for matched and mismatched asynchronous audiovisual stimuli. We predicted that, unlike previous research with musical stimuli, there would be decreased sensitivity (greater Just Noticeable Differences, JNDs) for temporal order in matched conditions vs. mismatched conditions, as a result of a greater percept of audiovisual unity.

We chose musical stimuli in part to allow for investigating whether long-term experience affects the unity assumption. Accuracy is superior for familiar stimuli (Vatakis & Spence, 2006a), and musical training increases sensitivity to audiovisual asynchrony (Lee & Noppeney 2011). Past findings also suggest that musicians, through their long-term training, acquire internal forward action models that allow them to better predict the auditory outcomes of visual actions (Petrini, Russell & Pollick, 2009). Our primary research question concerning musical expertise, however, is motivated by Vatakis & Spence's (2008) explanations for the unique effect

of the unity assumption on speech stimuli, specifically regarding the heightened familiarity of this stimulus class.

If familiarity plays a role in our ability to judge unity (Vatakis & Spence, 2008), increased exposure to musical stimuli may lead to a greater influence of unity with musical stimuli. Previous research investigating the link between musical expertise, audiovisual congruency, and synchrony perception in point-light drumming displays has shown an effect of expertise. Whether a moving dot's velocity is correlated with the temporal properties of a drumming sound affects the audiovisual integration window only for novices and not expert drummers (Petrini et al., 2009). Given the challenge of understanding the role of long-term experience in influencing audiovisual unity perception, we designed our task to afford comparison of musicians and non-musicians to contribute to ongoing exploration of this complex topic.

Experiment 1

In Experiment 1, we tested whether the unity assumption influenced temporal cross-modal binding when contrasting musical events with dissimilar amplitude envelopes. We compared cello and marimba audiovisual stimuli in an un-speeded, TOJ task. Participants viewed audio-visually matched and mismatched versions of these stimuli at various stimulus onset asynchronies, and indicated which modality was presented first. We predicted that sensitivity to audiovisual asynchrony would be lower in matched conditions due to cross-modal binding facilitated by unity between auditory and visual signals.

Methods

Participants

Fifteen musicians (M = 18.9 years, SD = 1.2, six males) and 15 non-musicians (M = 18.5 years, SD = 1.1 years, five males) were included in the final analysis. Nine additional non-musicians and one additional musician participated in the experiment but were excluded, as they could not properly carry out the task. Trained musicians were defined as those having 6 years or more of formal music training, and non-musicians were defined as those having less than 2 years. Musicians listed piano (N = 11), violin (N = 3), and guitar (N = 1) as their primary instrument.

All participants included in the analysis passed a standard audiometric screening test. All participants were students enrolled in undergraduate psychology courses at McMaster University, and received course credit for participation. The McMaster Research Ethics Board approved the study, and all participants gave informed consent to participate.

Apparatus

The experiment was conducted in a sound attenuating testing booth. The experiment was run in Psychopy (Pierce, 2007): a Python-based stimulus presentation program, which ran on a Macintosh computer (OS X 10.9.3). The visual stimuli were displayed on a Dell M993 monitor, and auditory stimuli were presented over Sennheiser headphones at a system level of 3.

Materials

The audiovisual stimuli consisted of color video clips presented on a grey background, each with a male performer playing either a cello or a marimba, in front of a red curtain (Fig. 2) within a room designed for stimulus creation within the MAPLE Lab. The stimuli were recorded on a Samsung HMX-F80 HD camcorder. A trained cellist and trained marimbist were recruited and instructed to naturally play a single note (Middle C, 261.63 Hz) in front of a red curtain. Stimulus auditory and visual components were separated and then edited using Final Cut Pro (waveforms of stimuli appear in Fig. 3). Each clip began and ended with a still frame from the start/end of the video clip. During the still frame, each clip also began and ended with background acoustic noise (rather than silence), so that participants could not infer the stimulus onset asynchrony (SOA) from the duration of silence accompanying the still frame. The still frame at the beginning and end of each stimulus was always 300 ms long, regardless of audiovisual condition and audiovisual SOA. To manipulate SOA, the auditory track was shifted forward or backward according to the SOA of interest, and the duration of acoustic noise at the beginning and end of each stimulus was adjusted to accommodate the SOA, while keeping overall video clip length constant across SOA conditions. In this way, regardless of SOA, total marimba video stimulus duration (including the still frame) was always 4.2 s in duration, and all cello video stimuli were 4 s in duration. The marimba tone (independent of acoustic noise) was 2.7 s in duration and 60 dB(A) in intensity, and the cello 1.3 s in duration and 65 dB(A) in intensity. The durations of each tone were approximately equivalent in perceived duration. Previous work has shown that sounds with sustained (e.g., cello) envelopes are perceived as longer than those with decaying (e.g., marimba) envelopes (Grassi & Darwin, 2006; Grassi, 2010; Grassi & Pavan, 2012). Auditory components of the stimuli had 16 bits per sample, and a sample rate of 48 kHz.

Design

There were nine different SOAs between the auditory and visual stimuli: ±300 ms, ±200 ms, ±130 ms, ±70 ms, 0 ms, where positive SOAs indicate that the visual stimulus was presented first and negative SOAs indicate that the auditory

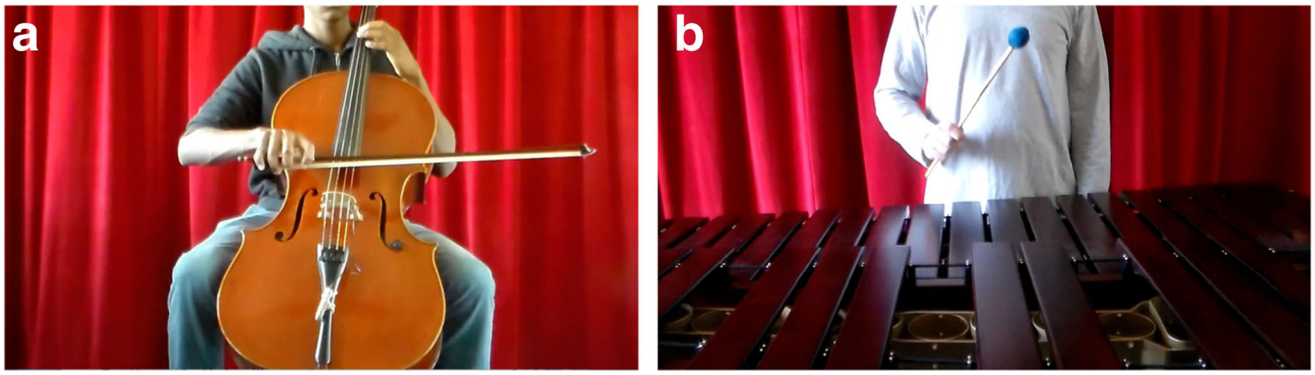

Fig. 2 **a**,**b** Stills taken from the video stimuli used in all experiments, **a** cello, **b** marimba

stimulus was presented first. There were ten repetitions of each SOA for each audiovisual combination condition (cello matched, cello mismatched, marimba matched, marimba mismatched), leading to a total of 350 ($9 \times 4 \times 10$) trials. Trial order was presented randomly.

Procedure

Participants were told that they would be viewing videos of a man playing either a cello or marimba, and that sometimes the auditory content would not match the video content. They were told that either the auditory or visual information would lead on a given trial, and that their task was to report whether the auditory or visual event occurred first by pressing "A" on the keyboard if they thought the auditory event was presented first, and "V" if visual event was first. The responses were un-speeded. To ensure they understood the instructions, before the main experiment, participants first completed eight practice trials, which consisted of the maximal SOAs (300 ms) for auditory leads and visual leads for each of the audiovisual combinations.

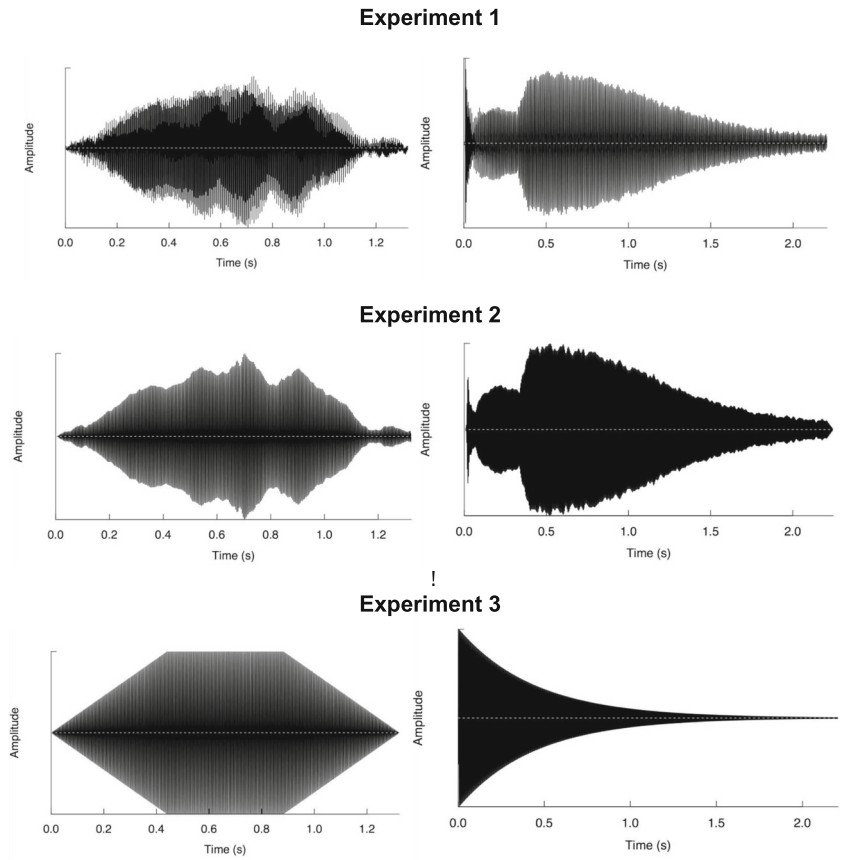

Fig. 3 Amplitude envelopes used in Experiments 1, 2 and 3 (from *top* to *bottom*) for cello (*left*) and marimba (*right*) auditory conditions

Analysis

Assuming a cumulative normal distribution, the proportion of vision first responses was fit to a psychometric curve (Fig. 4) using probit analysis (Finney, 1947); responses were converted to z-scores and the line of best fit was calculated per participant, for each of the four audiovisual combinations. The slope and intercept of each line was then used to calculate the JND (0.675/slope, given that ±0.675 represents the 75% and 25% points on the cumulative normal distribution) and point of subjective simultaneity (PSS; –intercept/slope). A separate ANOVA was conducted for both JND and PSS, with match condition and visual condition as within subjects variables, and musicianship as a between subjects variable. We reported *F*-values and *P*-values, and, for effect size, we reported generalized eta squared, notated η^2_G, –as this allows for comparison of effect sizes across both between-subjects and within-subjects designs (Olejnik & Algina, 2003; Bakeman, 2005).

Coefficients of determination (r^2) were also calculated per participant per condition, which indicated the goodness-of-fit of the data. In conjunction with visually inspecting psychometric curves per participant, r^2 values were used to index participants who were unable to adequately carry out the task; participants with average r^2 values (across conditions) less than 0.4 were excluded from the final analysis (nine non-musicians, one musician) as their poor fits rendered their data non-interpretable for our final analysis.

Just noticeable differences

There was a significant main effect of match condition (Fig. 5), $F(1,28) = 10.27$, $P = .0034$, $\eta^2_G = .062$, where overall, JNDs in matched conditions ($M = 213$ ms) were greater than JNDs in mismatched conditions ($M = 160$ ms). There was no effect of musicianship or visual condition on JND.

Point of subjective simultaneity

There was a significant main effect of match condition on PSS, $F(1,28) = 14.66$, $P = .0006$, $\eta^2_G = .10$, where matched

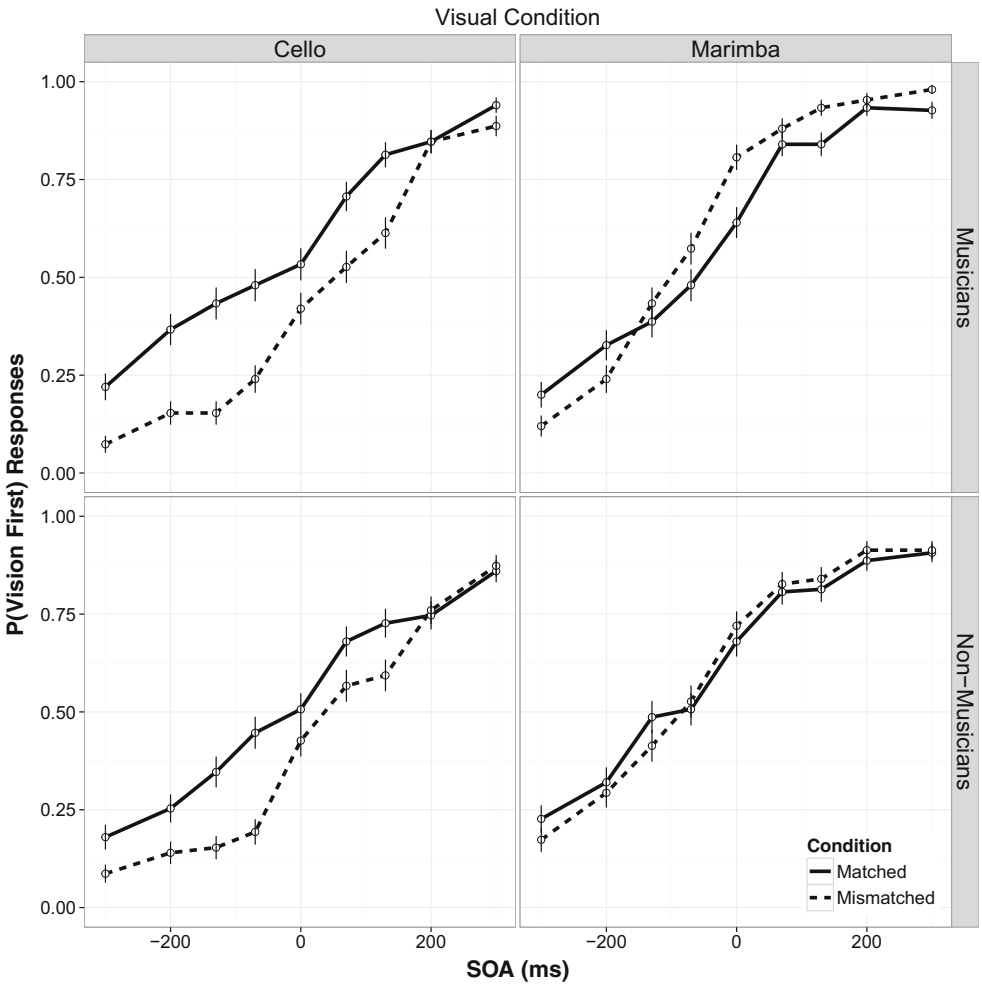

Fig. 4 Mean proportion of 'vision first' responses as a function of stimulus onset asynchrony (SOA) between visual and auditory stimuli, for musicians (*top panel*) and non-musicians (*bottom panel*) in Experiment 1. *Error bars* indicate within-subjects ± 1 SE about the mean

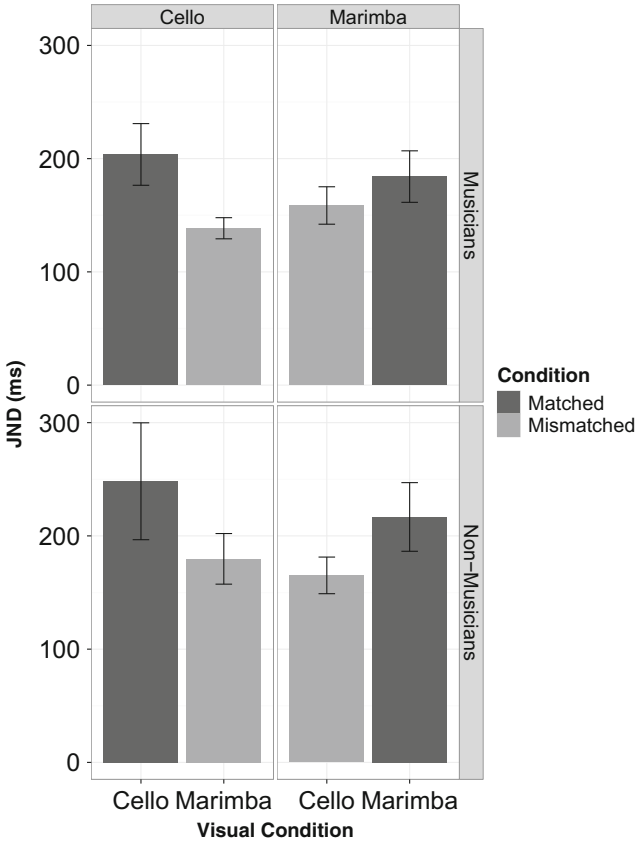

Fig. 5 Average just noticeable differences (JNDs) for audiovisual matched and mismatched musical stimuli for musicians (*top panel*) and non-musicians (*bottom panel*) in Experiment 1. *Error bars* indicate within-subjects ± 1 SE about the mean

conditions required an earlier auditory lead (M = –100 ms) than mismatched conditions (M = –25 ms) for simultaneity perception (Fig. 6). There was a significant effect of visual condition, $F(1,28) = 24.2$, $P < .0001$, $\eta_G^2 = .2$, where marimba visual conditions required a greater auditory lead for audiovisual simultaneity to be perceived (M = –120 ms), than the cello

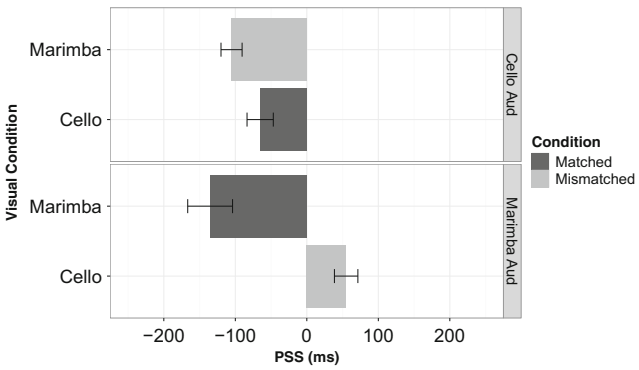

Fig. 6 Mean point of subjective simultaneity (PSS) values for matched and mismatched conditions, organized by visual and auditory conditions for Experiment 1. *Error bars* indicate within-subjects ± 1 SE about the mean

visual conditions, which on average, had PSSs that were closer to veridical (M = –5 ms).

Finally, there was a significant interaction between visual condition and match condition, $F(1,28) = 11.3$, $P = .0022$, $\eta_G^2 = .037$. In the cello visual conditions, there was a significant ($P < .0001$) difference between matched and mismatched PSS after correcting for multiple comparisons using the Bonferroni-Holm method, where the mismatched cello visual/marimba auditory condition requires a visual lead (M = 55 ms), and the matched, cello visual/cello auditory condition requires an auditory lead (M = –65 ms). There was no significant difference between match conditions in marimba visual conditions, which both required auditory leads, regardless of the auditory stimulus pairing.

Discussion

The results from Experiment 1 provide important evidence that the influence of unity on audiovisual binding of realistic stimuli is not specific to speech: participants were more sensitive to audiovisual temporal order when the audiovisual information was mismatched, and less sensitive to temporal order when matched. The decrease in sensitivity in matched conditions was presumably due to the stronger cross-modal integration resulting from an observer's assumption that the auditory and visual signals came from the same underlying musical event.

Although we had no predictions concerning the PSS, we observed a main effect of match condition, where matched conditions required a greater auditory lead than mismatched conditions. We also found that conditions in which the video consisted of the marimba strike required a much greater auditory lead than conditions in which the video consisted of the cello bowing, in which the PSS was closer to veridical. The PSS appeared to be affected by stimulus-driven features, varying with each particular audiovisual combination, as indicated by the interaction between visual and match condition.

In the current experiment, many non-musicians (nine of ten excluded participants) were unable to complete the task. This may be explained by musicians' more narrow temporal windows for audiovisual integration in general (Petrini et al., 2009), and for musical stimuli, more specifically (Lee & Noppeney, 2011), or from differences in motivation between musicians and non-musicians (McAuley, Henry, & Tuft, 2011). Importantly, the effect of match condition did not differ as a result of musical training. Contrary to our predictions, this suggests that familiarity with a particular stimulus class does not affect the strength of perceived unity. Alternatively, the effects of stimulus familiarity might operate only when the differences between conditions are more nuanced. In the present experiment, these differences may have been sufficiently salient for even a non-expert to distinguish match from mismatch conditions.

Finally, we believe the difference between our clear effect of audiovisual congruency on JND and the null effects found in previous work (Vatakis & Spence, 2008) reflects the importance of amplitude envelope—which differs markedly between cello and marimba notes but much less so between the guitar and piano notes used previously. This suggests that amplitude envelope may play an important role in influencing whether an observer perceives an acoustic signal as belonging to the same event as an accompanying visual event. It is also possible, however, that the percept of 'unity' was modulated by overall differences in timbre—the unique quality characterizing an instrument's sound.

Although the duration of the attack, a parameter related to amplitude envelope, is one factor that shapes perception of timbre, so are spectral properties such as spectral centroid and spectral flux (McAdams, Winsberg, Donnadieu, Krumhansl & De Soete, 1995). When amplitude envelope is kept constant, differences in spectral cues can determine whether an observer integrates auditory and visual information (Grassi & Casco, 2010). If spectral cues are important for integration, amplitude envelope alone may not provide enough event information to facilitate cross-modal integration for an observer. This possibility that spectral cues contribute to the percept of audiovisual unity was tested in Experiment 2.

Experiment 2

The results of Experiment 1 suggest that amplitude envelope may act as an important cue for 'unity' between auditory and visual signals. However, it is also possible that spectral cues may contribute to an observer's percept of audiovisual unity. In Experiment 2, we test the hypothesis that amplitude envelope can serve as a primary cue for facilitating the percept of unity for realistic object-action events, independent of spectral differences between the cello and marimba. We contrast pure tone versions of marimba and cello amplitude envelopes in a TOJ task, keeping spectral information constant across all audiovisual conditions.

Methods

The apparatus, design, and procedure were identical to Experiment 1, with the following differences: a new group of participants were tested, and spectral differences were removed from the original auditory signals.

Participants

Fourteen musicians (M = 18.6 years, SD = 0.65 years, five males) and 14 non-musicians (M = 18.9 years, SD = 1.5 years, four males) from McMaster University participated in the experiment. Musicians listed piano (N = 10), clarinet (N = 2), violin (N = 1) and flute (N = 1) as their primary instrument. Six additional non-musicians and one musician participated but were not included the final analysis as they were unable to properly complete the task. Participants were naïve to the task and did not participate in Experiment 1.

Stimuli

Pure tone versions with marimba and cello amplitude envelopes were synthesized in R using the *seewave* package (Sueur, Aubin, & Simonis, 2008). For each sound, the envelope was extracted as a vector, and then smoothed with a low pass (50 Hz) filter. The envelope was then filled with a 250 Hz pure tone carrier with a sample rate of 48 kHz. This is comparable to Grassi and Casco's (2010) method for removing spectral information from sound signals, and is presumed to neutralize a sound's meaning. For each SOA condition, the pure tone stimulus was aligned at the onset of the original complex auditory stimulus. Although each video started and stopped with a still frame (as in Experiment 1), here there was no acoustic noise flanking each sequence. Adding noise at the beginning and end of the auditory sequence was unnecessary since there was no background noise in the artificially generated auditory stimulus to cue the participant of the SOA. The marimba-like tone was 70 dB(A) in intensity, and the cello-like tone was 65 dB(A) in intensity. The durations of each tone were the same as in Experiment 1.

Apparatus

To roughly approximate loudness levels with Experiment 1, auditory stimuli were presented over at a system volume level of 5.

Results

The analysis was identical to that conducted in Experiment 1. Here, six non-musicians and one musician did not meet the inclusion criterion due to poor goodness of fit of their data (average r^2 values below 0.4). The psychometric curve for these data appear in Fig. 7.

Just noticeable differences

There was a significant main effect of visual condition on asynchrony sensitivity (Fig. 8), $F(1,26) = 14.38$, $P = .0008$, $\eta^2_G = .059$, where participants had greater JNDs for cello visual conditions (M = 201 ms) than marimba visual conditions (M = 156 ms). Moreover, there was a significant interaction between visual condition and musical training, $F(1, 26) = 4.61$, $P = .041$, $\eta^2_G = .02$: for non-musicians, JNDs

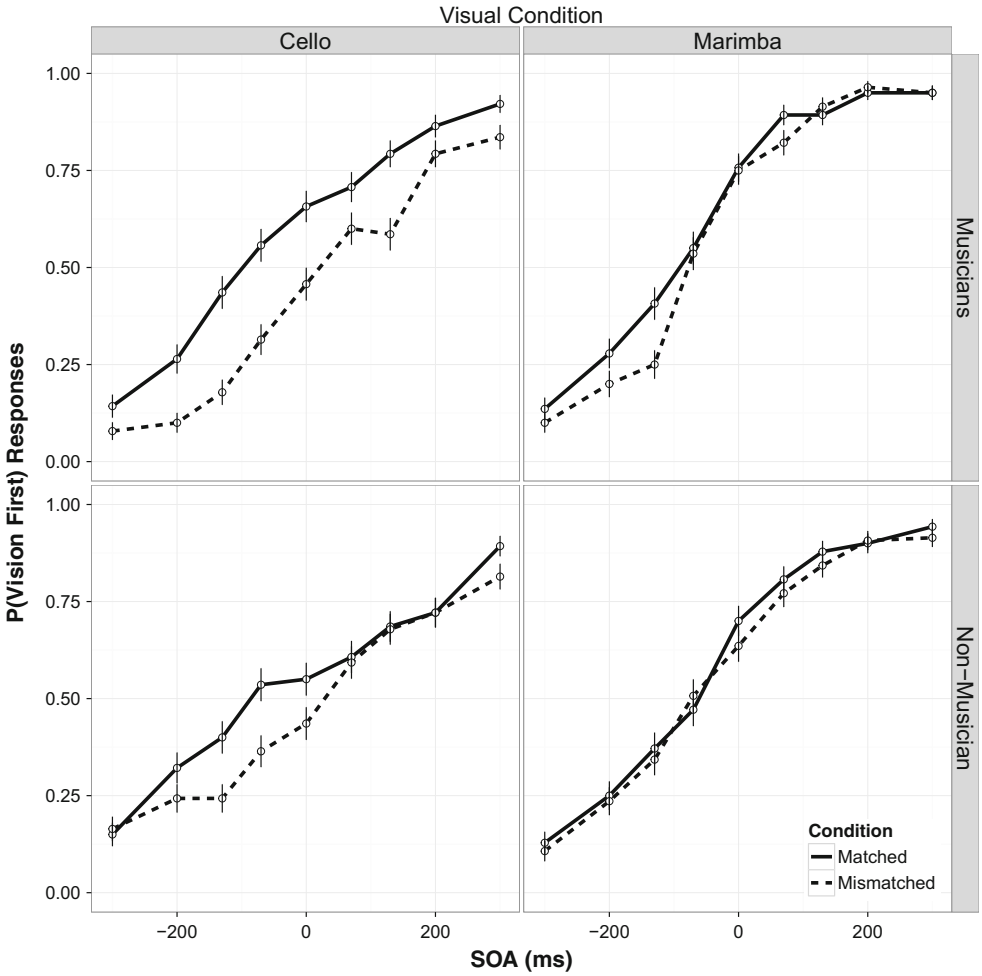

Fig. 7 Mean proportion of 'vision first' responses as a function of SOA for musicians (*top panel*) and non-musicians (*bottom panel*) in Experiment 2. *Error bars* indicate within-subjects ± 1 SE about the mean

were significantly ($P = .018$) greater in the cello visual condition ($M = 229$ ms) than the marimba visual condition ($M = 158$ ms), after correcting for multiple comparisons using the Bonferroni-Holm method. There was no difference between visual conditions for musicians.

Point of subjective simultaneity

There was a main effect of visual condition on the PSS, $F(1,26) = 8.55$, $P = .0071$, $\eta_G^2 = .10$, where the marimba visual conditions require a greater auditory lead (M = –91 ms) than the cello visual conditions (M = –7 ms), which have PSSs that are, on average, closer to veridical (Fig. 9). There was also a main effect of match condition on the PSS, $F(1,26) = 10.17$, $P = .0037$, $\eta_G^2 = .05$, where matched conditions, on average, required a greater auditory lead ($M = –78$ ms) than mismatched conditions ($M = –20$ ms) for audiovisual simultaneity perception.

Discussion

Without spectral cues, there was no influence of match condition on asynchrony sensitivity. This suggests that within the context of the current study, amplitude envelope alone could not facilitate a percept of audiovisual unity. Rather than a main effect of match condition, we observed a main effect of visual condition on sensitivity: participants had lower JNDs when the visual stimulus consisted of a marimba compared to a cello gesture. Moreover, this effect of visual condition was only significant for non-musicians, who exhibited significantly lower temporal sensitivity for cello visual trials than marimba visual trials. For musicians, there are no significant differences in JND between visual conditions, suggesting similar degrees of cross-modal binding across audiovisual combinations.

Given that the visual stimuli used in this experiment were identical to those in Experiment 1 (there was no effect of visual condition in Experiment 1), the effect of visual condition cannot be explained by visual stimulus-driven factors, such as the ambiguity of a cellist's gesture. It is more likely that

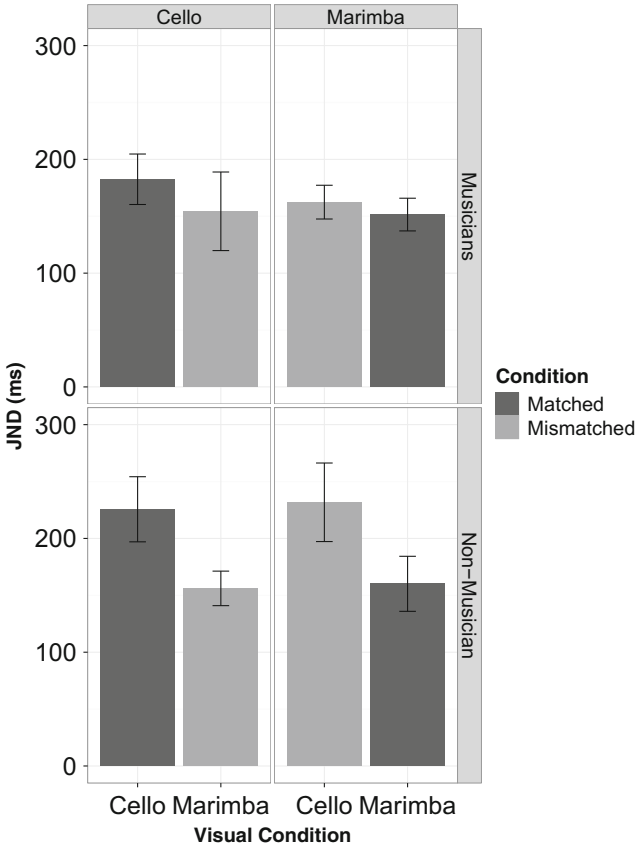

Fig. 8 Average JNDs for audiovisual matched and mismatched musical stimuli for cello visual (left) and marimba visual (right) conditions in Experiment 2. *Error bars* indicate within-subjects ± 1 SE about the mean

without spectral information, the marimba amplitude envelope may have integrated with the cello visual gesture in the mismatched condition. Impact sounds begin with a click consisting of a complex frequency spectrum (van den Doel & Pai, 1998), and removing such spectral cues from the marimba-envelope may have rendered the 'impact' quality of the sound less salient. The marimba-envelope pure tone may have also been given a cello-like, sustained quality when the period of reverberation was filled with a pure tone.

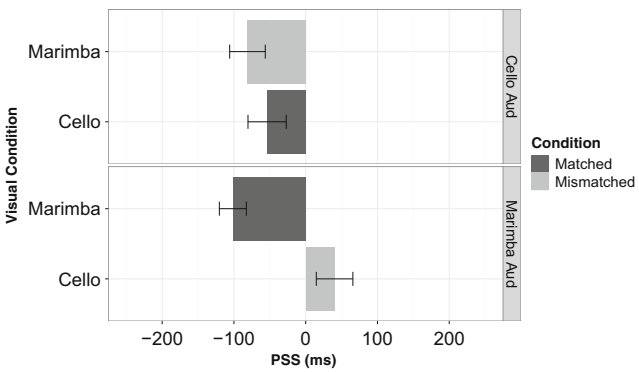

Fig. 9 Mean PSS values for matched and mismatched conditions, organized by visual and auditory conditions for Experiment 2. *Error bars* indicate within-subjects ± 1 SE about the mean

Research on amplitude envelope and audiovisual integration has successfully used auditory stimuli that simulate impact events using pure tones without a complex frequency spectrum at the onset. The marimba illusion occurs with impact-like, computer-generated exponentially decaying pure tones, but not with flat tones with a period of amplitude sustain (Schutz, 2009). Similarly, the audiovisual bounce inducing effect is increased when streaming disks are paired with a pure tone with an impact-like, computer-generated decaying envelope, compared to one with a gradual amplitude attack (Grassi & Casco, 2009). Using synthesized envelopes resembling categorical event types can be a useful avenue for elucidating audiovisual integration processes. In Experiment 3, we also manipulated amplitude envelope using 'percussive' and 'sustained' pure tones that resemble marimba and cello envelopes, respectively.

Finally, PSS values were comparable to those of Experiment 1, where the PSS was affected by both audiovisual congruency and stimulus-driven (visual) information. For simultaneity to be perceived, matched conditions required greater auditory leads than mismatched conditions, and marimba visual conditions required greater auditory leads than cello visual conditions.

Experiment 3

Here, we again contrast marimba and cello stimuli in an unspeeded TOJ task similar to Experiments 1 and 2, while keeping spectral information constant. Our goal was to assess whether amplitude envelope alone as a cue can facilitate the percept of unity, while accounting for the possibility that the marimba-envelope auditory stimulus in Experiment 2 did not resemble an impact acoustic event due to the reverberation in the signal. As such, we manipulated the parameter of amplitude envelope using computer-generated, categorically 'percussive' (marimba-like) and 'sustained' (cello-like) pure tones.

Methods

The apparatus, design, and procedure were identical to Experiments 1 and 2. Only the following differences existed in Experiment 3.

Participants

Fourteen musicians (M = 19.1 years, SD = 1.4 years, six males) and 14 non-musicians (19.6 years, SD = 2.2 years, two males) from McMaster University participated in the experiment. Musicians listed piano (N = 12), guitar (N = 1) and drums (N = 1) as their primary instrument. Three additional non-musicians participated but were not included in the final

analysis as they were unable to complete the task. Participants were naive to the task and did not participate in Experiment 1 or Experiment 2.

Stimuli

Pure tone auditory stimuli with a "marimba-like" (decaying) and "cello-like" (sustained) amplitude envelope were synthesized in R using the *seewave* package (Sueur et al., 2008). The marimba stimulus was 2 s in duration, and its envelope was defined with a 4 ms risetime, followed by an exponential decay to 0 in which amplitude (a) as a function of time (t) is given by the formula: $a(t) = e^{(-5t/T)}$, where T refers to the total duration of the decay (Schlauch, Ries, & DiGiovanni, 2001). The cello-like envelope, which was 1.3 s in duration, was defined with a gradual rise time of 440 ms, a sustain period of 440 ms, and a gradual fall-time of 440 ms. This artificial envelope approximated the amplitude changes of the original cello stimulus. The maximum amplitude of the cello-like envelope was half of the maximum of the marimba-like envelope to control for perceived overall loudness differences between the two envelope types. Each envelope was then filled with a 250 Hz pure tone carrier with a sample rate of 48 kHz. Both tones were 80 dB(A) in intensity.

As in Experiment 2, for each SOA condition, the onset of the pure tone stimuli was aligned at the onset of original (Experiment 1) complex auditory stimulus. As in Experiment 2, each video started and stopped with a still frame, but there was no acoustic noise flanking each sequence.

Results

The analysis was identical to that conducted in the previous experiments (the psychometric curve for these data appear in Fig. 10). Three participants, all non-musicians, did not meet the inclusion criterion (had average r^2 values below 0.4), and were thus excluded from the JND and PSS analyses.

Just noticeable differences

There was a significant effect main effect of match condition on JND (Fig. 11), $F(1,26) = 5.2$, $P = .031$, $\eta^2_G = 0.028$, where participants had greater JNDs in matched conditions (M = 185 ms) than in mismatched conditions ($M = 160$ ms). There was a main effect of visual condition on JND, $F(1, 26) = 8.3$, $P = 0.0078$, $\eta^2_G = 0.064$, where participants had greater JNDs for cello visual conditions ($M = 193$ ms) relative to marimba visual conditions ($M = 150$ ms). These main effects were qualified by a significant interaction between visual condition and match condition, $F(1,26) = 5.61$, $P = .025$, $\eta^2_G = .045$. After correcting for multiple comparisons using the Bonferroni-Holm method, there was a marginally significant ($P = .06$) difference between the matched and mismatched condition only for cello visual trials. For marimba visual trials, there was no difference between match conditions, suggesting that the main effect of match condition was simply driven by differences in the cello visual conditions. There was no effect of musical training on JND.

Point of subjective simultaneity

There was a significant main effect of match condition, $F(1, 26) = 6.45$, $P = .0174$, $\eta^2_G = .029$, on the PSS (Fig. 12): matched conditions required a greater auditory lead ($M = –56$ ms) than mismatched conditions, where the PSS was closer to veridical ($M = –15$ ms). There was also a main effect of visual condition, $F(1,26) = 28.6$, $P < .0001$, $\eta^2_G = 0.22$, where on average, cello visual conditions required a visual lead ($M = 28$ ms), marimba visual conditions required an auditory lead ($M = –99$ ms). Finally, there was a significant interaction between visual condition and match condition, $F(1,26) = 18.38$, $P = .0002$, $\eta^2_G = .12$. The mismatched cello visual/marimba auditory condition requires a significantly ($P = .0026$) greater visual lead ($M = 93$ ms) than the matched cello condition, which requires a slight auditory lead ($M = –37$ ms). However, in marimba visual conditions, there is no difference between match conditions; both require moderate auditory leads. Multiple comparisons were corrected using the Bonferroni-Holm method.

Discussion

Using pure tone stimuli with synthesized amplitude envelopes *representative* of marimba and cello musical events, we observed a significant effect of audiovisual unity on asynchrony sensitivity (as in Experiment 1). As in Experiment 2, we also observed an effect of visual condition on JND: cello visual conditions had greater JNDs compared to marimba visual conditions. Importantly, however, both the effect of audiovisual unity and the effect of visual condition were driven primarily by the heightened JNDs in the cello matched condition, relative to all other audiovisual combinations (Fig. 11). This was reflected in the significant interaction between match condition and visual condition; there is only a significant effect of match condition for visual cello conditions. As there was no significant increase in JND for the marimba-matched condition, we cannot conclude that amplitude envelope alone facilitates the unity assumption in cross-modal temporal binding.

Contrary to the results of Experiment 2, however, here there were no apparent differences in JND patterns between musicians and non-musicians. This is likely attributed to the lowered JND in the cello visual/marimba audio mismatched condition for non-musicians. This suggests decreased audiovisual binding of marimba-like auditory stimulus with the cello visual stimulus when the sound has a computer-generated, exponentially decaying envelope (Experiment 3), rather

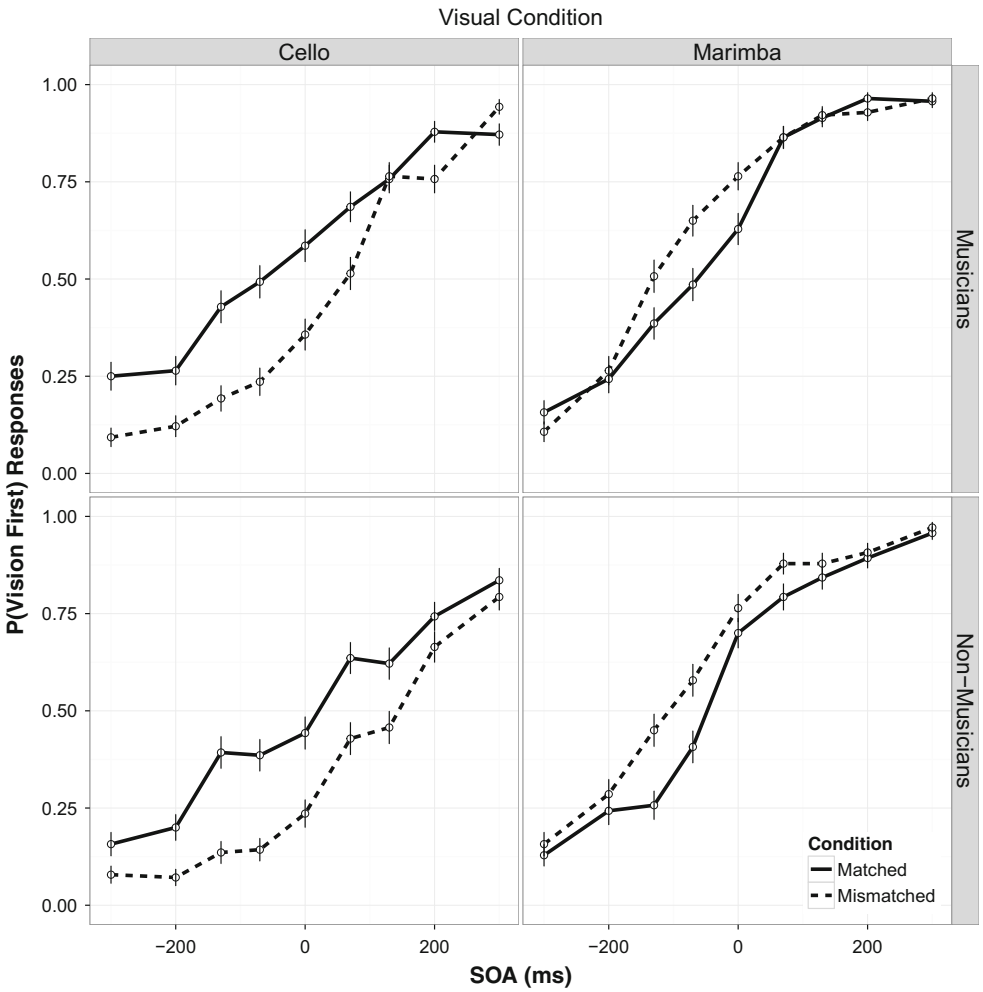

Fig. 10 Mean proportion of "vision first" responses as a function of SOA for musicians (*top panel*) and non-musicians (*bottom panel*) in Experiment 3. *Error bars* indicate within-subjects ± 1 SE about the mean

than a filtered version of the original marimba amplitude envelope (Experiment 2). These results thus support our proposal that in Experiment 2, some physical attributes of the pure-tone marimba envelope likely rendered it less 'marimba-like' and promoted binding with the cello visual stimulus for non-musicians.

Finally, despite spectral changes, PSS values across audiovisual conditions were comparable to those in Experiments 1 and 2. As in the previous experiments, the PSS was affected by both audiovisual congruency and stimulus-driven information: matched conditions required greater auditory leads than mismatched conditions, and marimba visual conditions required greater auditory leads than cello visual conditions for simultaneity perception.

General discussion

Our results indicate an effect of audiovisual congruency on cross modal temporal binding of cello and marimba musical events. As predicted, congruency (i.e., originating from the same instrument) decreased sensitivity to temporal order, consistent with the idea that congruency across modalities increases multi-modal integration. These results indicate that the unity assumption does in fact play an important role in audio-visual integration of non-speech sounds, complementing previous demonstrations of its importance in speech perception (Vatakis & Spence, 2007).

Acoustic factors influencing cross-modal binding

We posited that, by choosing to contrast musical instruments with highly similar amplitude envelopes (i.e., guitar and piano), cross-modal binding may have occurred in both audiovisually matched and mismatched conditions in earlier work (Vatakis & Spence, 2008). Here, we show that that using a similar match/mismatch paradigm that contrasted marimba and cello events, which are highly dissimilar in amplitude cues, reveals an effect of congruency on audiovisual temporal binding of musical stimuli.

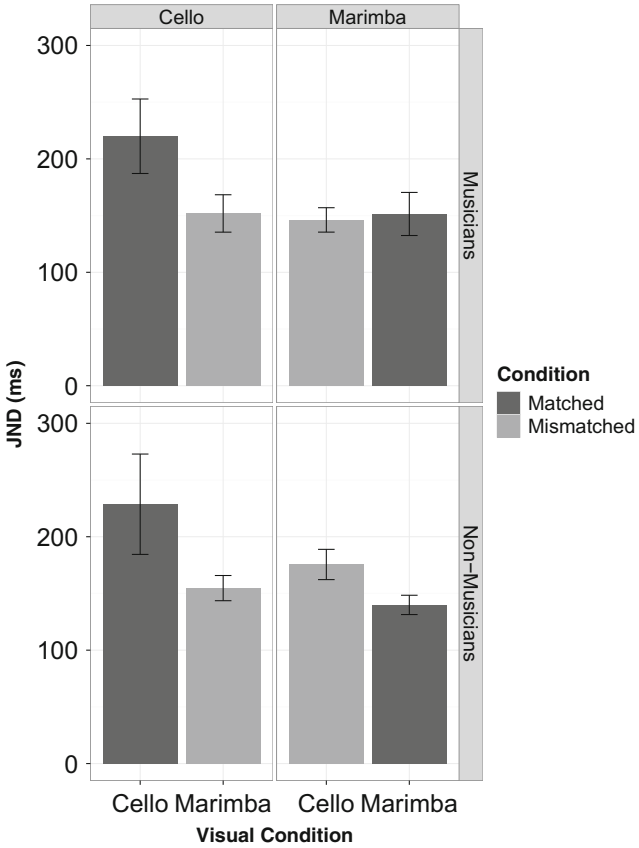

Fig. 11 Average JNDs for audiovisual matched and mismatched musical stimuli for cello visual (*left*) and marimba visual (*right*) conditions and musicians (*top*) and non-musicians (*bottom*) in Experiment 3. *Error bars* indicate within-subjects ± 1 SE about the mean

Experiments 2 and 3 expanded our inquiry by clarifying the role of specific attributes of *timbre*: the unique quality of a sound that distinguishes it from other types of sounds (despite being equivalent in pitch, loudness or duration). Although changes in amplitude affect a sound's timbre, spectral parameters such as spectral flux and spectral centroid are also influential (McAdams, Winsberg, Donnadieu, Krumhansl & De Soete, 1995). To test whether the influence of unity could be solely attributed to differences in amplitude envelope, these experiments isolated amplitude envelope as a factor by removing spectral differences. Without spectral cues, there was no effect of audiovisual unity on cross modal binding, suggesting that amplitude and spectral cues work in concert to help an observer determine the causal source of an auditory event. This is consistent with Grassi and Casco's (2010) work showing spectral information plays an important role in determining whether an observer integrates auditory and visual information. Therefore we conclude that, within the context of this paradigm, distinctions in amplitude envelope are *necessary*, *but not sufficient*, for detection of multi-modal incongruity. Cross-modal binding of incongruous audiovisual events may occur when the events have similar amplitude envelopes (Vatakis & Spence, 2008); however, without spectral cues, amplitude envelope alone is not always sufficient to discriminate congruent vs. incongruent events (as shown by Experiments 2 and 3).

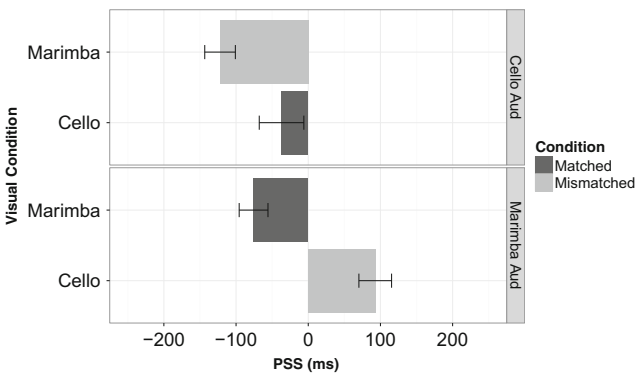

Fig. 12 Mean PSS values for matched and mismatched conditions, organized by visual and auditory conditions for Experiment 3. *Error bars* indicate within-subjects ± 1 SE about the mean

Audiovisual integration typically fails when there is a sufficiently large degree of conflict (spatial, or temporal) between audiovisual stimuli (Slutsky & Recanzone, 2001; Munhall, Gribble, Sacco, & Ward, 1996). Humans are able not only to optimally combine cues from different modalities, but also to use these same cues to deduce whether multimodal signals are linked to a common cause (Körding et al., 2007). Although Körding et al.'s causal inference model only considers spatial cues, our research suggests that multisensory causal inference extends beyond spatiotemporal domains. Although understanding factors shaping cross-modal binding outside these cues requires further research, it is likely that humans form causal links between modalities through more complex, modality-specific information, such as a sound's timbre. Future work could continue to explore what parameters help an observer determine whether two stimuli from different modalities belong to the same event.

Beyond musical stimuli: the role of event categories on audiovisual binding

Although we have focused primarily on experiments involving musical sounds in Vatakis and Spence's (2008) work on non-speech stimuli, their study also included non-musical object/action events. Those stimuli depicted a hammer smashing a block of ice, and a ball being dropped on the ground. Just as the guitar and piano both produce sound through impacts on strings, the ice smash and ball drop are both impact events involving solid objects. An observer, therefore, may have integrated the sound of the ice being smashed with the video of the ball. One area for future investigation would be, then, to re-evaluate whether the influence of audiovisual congruency holds when contrasting object-action events that are highly different.

Our perception of auditory events can be grouped according to the source's interacting materials into three categories: impact, aerodynamic, and liquid (Gaver, 1993). These

categories are perceptually distinct; sounds between groups are rarely confused (Gaver, 1993) and differentially influence audiovisual integration processes (Grassi & Casco, 2010). Contrasting these event types may be a more reasonable audiovisual comparison for a TOJ paradigm that investigates object-action stimuli, and could elucidate the role of higher-level cues in helping an observer determine whether information from two modalities originates from the same event.

Finally, while acoustic factors may be able to account for the null results with respect to object/action stimuli, they do not explain the lack of influence of audiovisual unity on perception of animal calls (Vatakis, Ghazanfar & Spence, 2008). One proposed explanation for these null results is that observers cannot rely on time-varying correlation between auditory and visual streams for animal calls, which consist of only transient onsets with minimal anticipatory information in the visual stimulus (Vroomen & Stekelenberg, 2010). Given that there was time-varying correlation between the musical actions and acoustic stimuli in the current experiment (e.g., the downward motion of the arm preceding the marimba tone), predictive, and time-varying information may also be an important parameter in influencing higher order interpretation of 'unity' between audiovisual streams, and presents a possible area for future investigation.

Stimulus factors influencing PSS shifts

Although our primary interest concerned JND, audiovisual manipulations also had an effect on the modality lead/lags required for audiovisual simultaneity to be perceived (PSS). In all experiments, marimba visual conditions required a greater auditory lead than cello visual conditions, which on average, had PSSs that were closer to veridical. There was also a main effect of match condition in all experiments, where matched conditions required a greater auditory lead than mismatched conditions. Given that the PSS pattern between audiovisual combinations remained relatively constant across experiments (and thus, resistant to spectral changes), our findings suggest that while JND is affected by spectral cues, PSS is not. Rather, it may be low-level time varying features in both the visual domain (e.g., predictive gesture information) and auditory domain (e.g., amplitude envelope) that affect the PSS, given that these were the features that varied between conditions, but were held constant across experiments.

Our conclusion that both spectral and time-varying amplitude cues influence sensitivity given that both types of acoustic cues play an important role in determining timbre is consistent with a wide body of previous research. Assuming that auditory event identification helps determine audiovisual unity, we should expect that both types of cues play an important role in temporal sensitivity. However, the inferred causal source of an auditory event should not be important in determining the PSS, and likely depends, rather, on a number of low level stimulus factors. Indeed, the wide variability in PSS values across experiments in the literature suggests that a number of low-level factors might affect this measure, with considerable disagreement on both the size and direction of the temporal asymmetry window (Vatakis, Maragos, Rodomagoulakis, & Spence, 2012). Some studies with audiovisual speech stimuli have found that simultaneity is perceived when the visual stream leads (Dixon & Spitz, 1980; Grant, Van Wassenhove, & Poeppel, 2004), while other research has found that the modality lead/lag differs depending on whether a phoneme or syllable is presented (Vatakis & Spence, 2006a). For musical stimuli, the PSS differs across instruments, with the piano stimulus requiring an auditory lead, and guitar stimulus requiring a visual lead (Vatakis & Spence, 2006b).

It is possible that visual anticipatory information can partially account for the PSS pattern observed in the current experiments. Previous research has found that the visibility and predictability of a visual speech signal facilitates the speed at which the corresponding auditory speech signal is processed (van Wassenhove, Grant, & Poeppel, 2005), and, similarly, visual anticipatory information has been shown to affect neural audiovisual interactions for non-speech stimuli (Vroomen & Stekelenburg, 2010). In our work, the anticipatory information in the marimba visual stimulus (the arm movement precedes impact) may have facilitated auditory processing, triggering greater auditory leads compared to cello visual conditions, which lacks predictive gesture information. Future work could investigate the role of visual predictive information in object/action events to elucidate the stimulus factors that influence the PSS.

Future directions: multisensory action perception

The current findings are part of a growing body of literature extending the study of audiovisual integration to more ecologically valid, complex stimuli. This complements previous multisensory research using simple and/or arbitrarily paired stimuli (e.g., light flashes and pure tones or clicks), which do not capture the meaningful cross-modal combinations that occur in the natural world (De Gelder & Bertelson, 2003). This focus on artificial sounds is not confined to audiovisual integration research, but appears to be a common trend across a wide range of auditory perception experiments (Schutz & Vaisberg, 2014; Gillard & Schutz, 2013). Although there are a number of more ecologically valid studies on audiovisual speech, the less common study of dynamic audiovisual stimuli (e.g. musical stimuli in the current study) allows researchers to probe additional theoretical questions regarding causality, predictive information, variations in training/ability, and action/gesture perception in multisensory integration. These questions are often not as readily accessible through speech, due to what some researchers have suggested is its 'special' nature (Jones & Jarick, 2006; Tuomainen, Andersen, Tiippana, & Sams, 2005). For instance, there are

few predictive cues in speech stimuli, which does not allow for exploration of dynamic anticipatory cues; a question more easily explored with dynamic action or gesture stimuli.

A parallel can be drawn between our work and research indicating that cross-modal temporal binding occurs between voluntary actions and their effects on the sensory environment. For instance, observers perceive their own intentional actions (a manual button press) and sensory outcomes of these actions (a beep) as occurring closer in time than they actually do (Haggard, Clark, & Kalogeras, 2002). For these self-generated actions, there appears to be a predictive element to action-outcome temporal binding: as long as the probability of a sensory outcome is sufficiently high, the timing of a voluntary action will be pulled toward that expected outcome's position in time, even if the outcome does not occur (Moore & Haggard, 2008). In addition to influencing the perceived timing of actions and their outcomes, research has indicated that congruency between auditory and visual information can also influence the actual speed of motor actions (Castiello, Giordano, Begliomini, Ansuini, & Grassi, 2010).

This temporal binding between actions and their effects is comparable to our findings; the primary difference being that we found temporal binding in externally generated action-effect pairs performed by a third party, rather than self generated actions. A fruitful area for future investigation might consider whether a comparable predictive/inferential mechanism, operating from a forward action model, underlies both phenomena. This would be consistent with the direct-matching hypothesis, which posits that action understanding is due, in part, to internal motor representation of observed actions (Rizzolatti, Fogassi, & Gallese, 2001).

Conclusions and closing thoughts

Our results provide compelling evidence for the influence of the unity assumption on temporal cross-modal binding for non-speech, musical stimuli. As with Vroomen and Stekelenberg (2010), our findings indicate that speech is not "special" when it comes to audiovisual temporal sensitivity. However, contrary to their findings, which argue that higher order interpretations of unity of do not facilitate temporal ventriloquism, our work extends evidence of the unity assumption's influence in audiovisual integration by demonstrating that complex modality-specific information (e.g., timbre) influences multisensory causal inference.

Finally, our findings rest on the assumption that the degree of temporal ventriloquism is an appropriate index for the degree of audiovisual integration. We do feel compelled to note recent research has suggested that audiovisual synchrony perception and integration may in fact be separate mechanisms: when participants are presented with audiovisual speech, auditory *leads* are required for maximized McGurk illusions, but auditory *lags* are required for audiovisual synchrony perception (Freeman et al., 2013). However, the influence of higher order interpretations of 'unity' on synchrony perception, such as in our present work, suggest that these processes are related to some degree. As such, future research is needed to elucidate the relationship between audiovisual integration and audiovisual synchrony perception, in addition to the avenues for future research prompted by our findings. These avenues include multisensory causal inference, the visual and acoustic factors that influence integration, and the potential link between observed and executed sensorimotor relations. We see this research as elucidating the processes underlying multimodal integration by illustrating the role of acoustic cues in contributing to the percept of multi-modal unity.

Acknowledgments The authors would like to thank Emily Gula, Devon Crawford, and Kimberly Germann for helping collect data for this work. This work was supported by grants from the Natural Sciences and Engineering Research Council of Canada (NSERC RGPIN/386603-2010), Ontario Early Researcher Award (ER10-07-195), and Canadian Foundation for Innovation (CFI-LOF 30101) to Dr. Michael Schutz, PI. Parts of this work were presented at the 2015 meeting of the Society for Music Perception and Cognition, Nashville and the 2015 Rhythm and Timing Symposium, London.

References

Armontrout, J. A., Schutz, M., & Kubovy, M. (2009). Visual determinants of a cross-modal illusion. *Attention, Perception, & Psychophysics, 71,* 1618–1627.

Aschersleben, G., & Bertelson, P. (2003). Temporal ventriloquism: crossmodal interaction on the time dimension. 2. Evidence from sensorimotor synchronization. *International Journal of Psychophysiology, 50,* 157–163. doi:10.1016/s0167-8760(03)00131-4

Baart, M., Stekelenburg, J. J., & Vroomen, J. (2014). Electrophysiological evidence for speech-specific audiovisual integration. *Neuropsychologia, 53,* 115–121. doi:10.1016/j.neuropsychologia.2013.11.011

Bakeman, R. (2005). Recommended effect size statistics for repeated measures designs. *Behavior Research Methods, 37*(3), 379–384. doi:10.3758/bf03192707

Bertelson, P., & Aschersleben, G. (1998). Automatic visual bias of perceived auditory location. *Psychonomic Bulletin & Review, 5*(3), 482–489. doi:10.3758/bf03208826

Bertelson, P., & Aschersleben, G. (2003). Temporal ventriloquism: crossmodal interaction on the time dimension. 1. Evidence from auditory-visual temporal order judgment. *International Journal of Psychophysiology, 50,* 147–155. doi:10.1016/s0167-8760(03)00130-2

Castiello, U., Giordano, B. L., Begliomini, C., Ansuini, C., & Grassi, M. (2010). When ears drive hands: the influence of contact sound on reaching to grasp. *PLoS ONE, 5*(8), e12240. doi:10.1371/journal.pone.0012240

De Gelder, B., & Bertelson, P. (2003). Multisensory integration, perception and ecological validity. *Trends in Cognitive Sciences, 7*(10), 460–467. doi:10.1016/j.tics.2003.08.014

Dixon, N. F., & Spitz, L. (1980). The detection of auditory visual desynchrony. *Perception, 9*(6), 719–721. doi:10.1068/p090719

Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature, 415*(6870), 429–433. doi:10.1038/415429a

Ernst, M. O. (2007). Learning to integrate arbitrary signals from vision and touch. *Journal of Vision, 7*(5), 7–7. doi:10.1167/7.5.7

Finney, D. K. (1947). *Probit analysis: a statistical treatment of the sigmoid response curve*. Cambridge: Cambridge University Press.

Freeman, E. D., Ipser, A., Palmbaha, A., Paunoiu, D., Brown, P., Lambert, C., et al. (2013). Sight and sound out of synch: fragmentation and renormalisation of audiovisual integration and subjective timing. *Cortex, 49*(10), 2875–2887. doi:10.1016/j.cortex.2013.03.006

Gaver, W. W. (1993). What in the world do we hear?: An ecological approach to auditory event perception. *Ecological Psychology, 5*(1), 1–29. doi:10.1207/s15326969eco0501_1

Gepshtein, S., Burge, J., Ernst, M. O., & Banks, M. S. (2005). The combination of vision and touch depends on spatial proximity. *Journal of Vision, 5*(11), 7–7. doi:10.1167/5.11.7

Gillard, J., & Schutz, M. (2013). The importance of amplitude envelope: surveying the temporal structure of sounds in perceptual research. In *Proceedings of the Sound and Music Computing Conference* (pp. 62–68). Stockholm, Sweden.

Grant, K. W., van Wassenhove, V., & Poeppel, D. (2004). Detection of auditory (cross-spectral) and auditory–visual (cross-modal) synchrony. *Speech Communication, 44*(1–4), 43–53. doi:10.1016/j.specom.2004.06.004

Grassi, M. (2010). Sex difference in subjective duration of looming and receding sounds. *Perception, 39*(10), 1424–1426. doi:10.1068/p6810

Grassi, M., & Casco, C. (2009). Audiovisual bounce-inducing effect: Attention alone does not explain why the discs are bouncing. *Journal of Experimental Psychology: Human Perception and Performance, 35*(1), 235–243. doi:10.1037/a0013031

Grassi, M., & Casco, C. (2010). Audiovisual bounce-inducing effect: When sound congruence affects grouping in vision. *Attention, Perception, & Psychophysics, 72*(2), 378–386. doi:10.3758/app.72.2.378

Grassi, M., & Darwin, C. J. (2006). The subjective duration of ramped and damped sounds. *Perception & Psychophysics, 68*(8), 1382–1392. doi:10.3758/bf03193737

Grassi, M., & Pavan, A. (2012). The subjective duration of audiovisual looming and receding stimuli. *Attention, Perception, & Psychophysics, 74*(6), 1321–1333. doi:10.3758/s13414-012-0324-x

Haggard, P., Clark, S., & Kalogeras, J. (2002). Voluntary action and conscious awareness. *Nature Neuroscience, 5*(4), 382–385. doi:10.1038/nn827

Jones, J. A., & Jarick, M. (2006). Multisensory integration of speech signals: the relationship between space and time. *Experimental Brain Research, 174*(3), 588–594. doi:10.1007/s00221-006-0634-0

Klatzky, R. L., Pai, D. K., & Krotkov, E. P. (2000). Perception of material from contact sounds. *Presence: Teleoperators and Virtual Environments, 9*(4), 399–410. doi:10.1162/105474600566907

Körding, K. P., Beierholm, U., Ma, W. J., Quartz, S., Tenenbaum, J. B., & Shams, L. (2007). Causal inference in multisensory perception. *PLoS ONE, 2*(9), e943. doi:10.1371/journal.pone.0000943

Lee, H., & Noppeney, U. (2011). Long-term music training tunes how the brain temporally binds signals from multiple senses. *Proceedings of the National Academy of Sciences, 108*(51), E1441–E1450. doi:10.1073/pnas.1115267108

Margiotoudi, K., Kelly, S., & Vatakis, A. (2014). Audiovisual temporal integration of speech and gesture. *Procedia - Social and Behavioral Sciences, 126,* 154–155. doi:10.1016/j.sbspro.2014.02.351

McAdams, S., Winsberg, S., Donnadieu, S., De Soete, G., & Krimphoff, J. (1995). Perceptual scaling of synthesized musical timbres: common dimensions, specificities, and latent subject classes. *Psychological Research, 58*(3), 177–192. doi:10.1007/bf00419633

McAuley, J. D., Henry, M. J., & Tuft, S. (2011). Musician advantages in music perception: an issue of motivation, not just ability. *Music Perception, 28*(5), 505–518. doi:10.1525/mp.2011.28.5.505

Metzger, W. (1934). Beobachtungen über phänomenale Identität. *Psychologische Forschung, 19*(1), 1–60. doi:10.1007/bf02409733

Moore, J., & Haggard, P. (2008). Awareness of action: inference and prediction. *Consciousness and Cognition, 17*(1), 136–144. doi:10.1016/j.concog.2006.12.004

Morein-Zamir, S., Soto-Faraco, S., & Kingstone, A. (2003). Auditory capture of vision: examining temporal ventriloquism. *Cognitive Brain Research, 17*(1), 154–163. doi:10.1016/s0926-6410(03)00089-2

Munhall, K. G., Gribble, P., Sacco, L., & Ward, M. (1996). Temporal constraints on the McGurk effect. *Perception & Psychophysics, 58*(3), 351–362. doi:10.3758/bf03206811

Navarra, J., Vatakis, A., Zampini, M., Soto-Faraco, S., Humphreys, W., & Spence, C. (2005). Exposure to asynchronous audiovisual speech extends the temporal window for audiovisual integration. *Cognitive Brain Research, 25*(2), 499–507. doi:10.1016/j.cogbrainres.2005.07.009

Olejnik, S., & Algina, J. (2003). Generalized eta and omega squared statistics: measures of effect size for some common research designs. *Psychological Methods, 8*(4), 434–447. doi:10.1037/1082-989x.8.4.434

Parise, C. V., & Spence, C. (2009). "When birds of a feather flock together": synesthetic correspondences modulate audiovisual integration in non-synesthetes. *PLoS ONE, 4*(5), e5664. doi:10.1371/journal.pone.0005664

Peirce, J. W. (2007). PsychoPy—psychophysics software in Python. *Journal of Neuroscience Methods, 162*(1-2), 8–13. doi:10.1016/j.jneumeth.2006.11.017

Petrini, K., Dahl, S., Rocchesso, D., Waadeland, C. H., Avanzini, F., Puce, A., & Pollick, F. E. (2009). Multisensory integration of drumming actions: musical expertise affects perceived audiovisual asynchrony. *Experimental Brain Research, 198*(2-3), 339–352. doi:10.1007/s00221-009-1817-2

Petrini, K., Russell, M., & Pollick, F. (2009). When knowing can replace seeing in audiovisual integration of actions. *Cognition, 110*(3), 432–439. doi:10.1016/j.cognition.2008.11.015

Powers, A. R., Hillock, A. R., & Wallace, M. T. (2009). Perceptual training narrows the temporal window of multisensory binding. *Journal of Neuroscience, 29*(39), 12265–12274. doi:10.1523/jneurosci.3501-09.2009

Radeau, M., & Bertelson, P. (1987). Auditory-visual interaction and the timing of inputs. *Psychological Research, 49*(1), 17–22. doi:10.1007/bf00309198

Rizzolatti, G., Fogassi, L., & Gallese, V. (2001). Neurophysiological mechanisms underlying the understanding and imitation of action. *Nature Reviews Neuroscience, 2*(9), 661–670.

Schutz, M., & Lipscomb, S. (2007). Hearing gestures, seeing music: vision influences perceived tone duration. *Perception, 36*(6), 888–897. doi:10.1068/p5635

Schutz, M. (2009). *Crossmodal integration: The search for unity (doctoral thesis)*. Charlottesville, VA: University of Virginia.

Schutz, M., & Kubovy, M. (2009). Causality and cross-modal integration. *Journal of Experimental Psychology: Human Perception and Performance, 35,* 1791–1810.

Schutz, M., & Vaisberg, J. M. (2014). Surveying the temporal structure of sounds used in Music Perception. *Music Perception: An Interdisciplinary Journal, 31,* 288–296.

Schlauch, R. S., Ries, D. T., & DiGiovanni, J. J. (2001). Duration discrimination and subjective duration for ramped and damped sounds. *The Journal of the Acoustical Society of America, 109*(6), 2880. doi:10.1121/1.1372913

Slutsky, D. A., & Recanzone, G. H. (2001). Temporal and spatial dependency of the ventriloquism effect. *NeuroReport, 12*(1), 7–10. doi:10.1097/00001756-200101220-00009

Su, Y.H. (2014). Content congruency and its interplay with temporal synchrony modulate integration between rhythmic audiovisual streams. *Frontiers in Integrative Neuroscience*, 8. doi:10.3389/fnint.2014.00092

Sueur, J., Aubin, T., & Simonis, C. (2008). Seewave: a free modular tool for sound analysis and synthesis. *Bioacoustics, 18,* 213–226.

Tuomainen, J., Andersen, T. S., Tiippana, K., & Sams, M. (2005). Audio–visual speech perception is special. *Cognition, 96*(1), B13–B22. doi:10.1016/j.cognition.2004.10.004

van den Doel, K., & Pai, D. K. (1998). The sounds of physical shapes. *Presence: Teleoperators and Virtual Environments, 7*(4), 382–395. doi:10.1162/105474698565794

Vallet, G., Shore, D. I., & Schutz, M. (2014). Exploring the role of amplitude envelope in duration estimation. *Perception, 43,* 616–630.

Van Wassenhove, V., Grant, K. W., & Poeppel, D. (2005). Visual speech speeds up the neural processing of auditory speech. *Proceedings of the National Academy of Sciences of the United States of America, 102*(4), 1181–1186. doi:10.1073/pnas.0408949102

Vatakis, A., & Spence, C. (2006a). Audiovisual synchrony perception for music, speech, and object actions. *Brain Research, 1111*(1), 134–142. doi:10.1016/j.brainres.2006.05.078

Vatakis, A., & Spence, C. (2006b). Audiovisual synchrony perception for speech and music assessed using a temporal order judgment task. *Neuroscience Letters, 393*(1), 40–44. doi:10.1016/j.neulet.2005.09.032

Vatakis, A., & Spence, C. (2007). Crossmodal binding: evaluating the "unity assumption" using audiovisual speech stimuli. *Perception & Psychophysics, 69*(5), 744–756. doi:10.3758/bf03193776

Vatakis, A., & Spence, C. (2008). Evaluating the influence of the "unity assumption" on the temporal perception of realistic audiovisual stimuli. *Acta Psychologica, 127*(1), 12–23. doi:10.1016/j.actpsy.2006.12.002

Vatakis, A., Ghazanfar, A. A., & Spence, C. (2008). Facilitation of multisensory integration by the "unity effect" reveals that speech is special. *Journal of Vision, 8*(9), 14–14. doi:10.1167/8.9.14

Vatakis, A., & Papadelis, G. (2014). The research on audiovisual perception of temporal order and the processing of musical temporal patterns: associations, pitfalls, and future directions. In D. Lloyd & V. Arstila (Eds.), *Subjective Time*. MIT Press.

Vatakis, A., Maragos, P., Rodomagoulakis, I., & Spence, C. (2012). Assessing the effect of physical differences in the articulation of consonants and vowels on audiovisual temporal perception. *Frontiers of Integrative Neuroscience, 6*(71), 1–18.

Vroomen, J., & Stekelenburg, J. J. (2010). Visual anticipatory information modulates multisensory interactions of artificial audiovisual stimuli. *Journal of Cognitive Neuroscience, 22*(7), 1583–1596. doi:10.1162/jocn.2009.21308

Warren, W. H., & Verbrugge, R. R. (1984). Auditory perception of breaking and bouncing events: a case study in ecological acoustics. *Journal of Experimental Psychology: Human Perception and Performance, 10*(5), 704–712. doi:10.1037/0096-1523.10.5.704

Welch, R. (1999). Meaning, attention, and the "unity assumption" in the intersensory bias of spatial and temporal perceptions. In G. Aschersleben, T. Bachmann, & J. Müsseler (Eds.), *Cognitive contributions to the perception of spatial and temporal events*. Amsterdam: Elsevier Science.