Perceptual Impact of Gesture Control of Spatialization

G. MARENTAKIS, IEM, University of Music and Performing Arts, Graz
S. McADAMS, CIRMMT, Schulich School of Music, McGill University

In two experiments, visual cues from gesture control of spatialization were found to affect auditory movement perception depending on the identifiability of auditory motion trajectories, the congruency of audiovisual stimulation, the sensory focus of attention, and the attentional process involved. Visibility of the performer's gestures improved spatial audio trajectory identification, but it shifted listeners' attention to vision, impairing auditory motion encoding in the case of incongruent stimulation. On the other hand, selectively directing attention to audition resulted in interference from the visual cues for acoustically ambiguous trajectories. Auditory motion information was poorly preserved when dividing attention between auditory and visual movement feedback from performance gestures. An auditory focus of attention is a listener strategy that maximizes performance, due to the improvement caused by congruent visual stimulation and its robustness to interference from incongruent stimulation for acoustically unambiguous trajectories. Attentional strategy and auditory motion calibration are two aspects that need to be considered when employing gesture control of spatialization.

Categories and Subject Descriptors: H.5.2 [**User Interfaces**] Evaluation/Methodology; H.5.2 [**User Interfaces**] Theory and Methods; H.5.2 [**User Interfaces**] Auditory (non-speech) feedback; H.5.1 [**Multimedia Information Systems**] Artificial, augmented, and virtual realities; H.5.5 [**Sound and Music Computing**] Sound and Music Computing

General Terms: Human Factors

Additional Key Words and Phrases: Document preparation, publications, typesetting

1. INTRODUCTION

Spatiotemporal organization in music has been used since Giovanni Gabrieli in the 16th century, but is receiving increased attention in contemporary music, with composers such as Edgar Varèse, Karlheinz Stockhausen, Pierre Boulez, Iannis Xenakis, John Chowning, Barry Truax, and Roger Reynolds making heavy use of dynamic spatial manipulations of sounds [Reynolds 2002; Harley 1994]. Such manipulations have been until recently achieved by sound engineers during the performance of a musical piece. Advances in human interface technology, however, have led to another interaction paradigm: the *gesture control of spatialization*. Gesture control of spatialization has been conceived since the time of Pierre Schaeffer's *potentiomètre d'espace* (1951),

Author's addresses: G. Marentakis, Institute for Electronic Music and Acoustics, Kunstuniversität Graz, Infelldgasse 10/3, Graz, A-8010; S. McAdams, CIRMMT, Schulich School of Music, McGill University, 555 Sherbrooke Street West, Montreal, QC, Canada H3A 1E3.

but it is becoming increasingly popular nowadays. A variety of interfaces for gesture control of spatialization have been developed by artists such as Michel Waisvisz (the Hand), Tod Machover (Bug Mudra), and Colby Leider (eLBo) to name a few.
Despite the increasing popularity of this technique, the perceptual impact of gesture control of spatialization on the audience has not been investigated. In this work, we contribute by investigating the impact of the cueing from the performer's gestures in connection with the attentional strategies that the audience might employ when exposed to bimodal feedback from the performer's gestures and the sound in the listening area. In addition, the results of the two experiments we present allow certain important observations to be made concerning the extent to which predictions made about auditory motion perception within auditory virtual environments hold under different listening setups and listener placements.

2. BACKGROUND

Relevant results from the psychoacoustics and 3D audio literature, interaction design, and cross-modal integration are reviewed here in order to provide the necessary background for the design of the experiments that follow.

2.1 Psychoacoustic and 3D Audio Reproduction Background

For both stationary and moving sounds, the precision with which humans perceive their direction and distance and detect their displacement is limited [Saberi et al. 1991; Blauert 1997; Mills 1958]. Similar to sound localization, auditory motion perception is more accurate for horizontal vs. vertical movements, frontal vs. lateral incidence, and broadband vs. narrowband sounds. Short onset times and substantial spectral variability over time improve performance [Chandler and Grantham 1992; Grantham 1986; Grantham and Hornsby 2003; Saberi and Perrott 1990]. Auditory motion perception is somewhat sluggish as dynamic auditory localization cues are integrated within a window of about 300ms to yield perceived motion [Chandler and Grantham 1992]. In accordance with this, auditory rotational motion perception is not robust above a speed of about 2 rot/sec [Féron et al. 2010] and, depending on speaker separation, stimulus onset asynchronies between 150 and 200 ms still yield motion direction discrimination above threshold [Lakatos and Shepard 1997]. The accuracy of tracking moving sounds is also limited and deteriorates in the presence of a distractor sound [Grohn et al. 2002].
To avoid the need for a large number of loudspeakers, virtual instead of real sources are synthesized by means of spatial audio algorithms applied to loudspeaker arrays. The prevailing spatialization techniques are Vector Based Amplitude Panning (VBAP) [Pulkki 2001], Ambisonics ([Zotter et al. 2012; Gerzon 1992; Malham 1999]), and Wave Field Synthesis [Berkhout et al. 1993]. Localization cues are degraded for virtual sound sources [Grohn et al. 2002; Pulkki and Hirvonen 2005; Guastavino et al. 2006]. Although it is possible to provide auditory spatial impressions, the accuracy of the perception of sound location, movement, and displacement is somewhat compromised. In comparative studies, in particular between VBAP and lower-order Ambisonics [Pulkki and Hirvonen 2005; Guastavino et al. 2006], VBAP yields the highest localization accuracy for the commonly used 8-speaker circular array.
Auditory motion is implemented in virtual environments by updating virtual sound location in real time according to the desired movement velocity. In this way, changes in intensity and interaural cues are produced. The Doppler frequency shift is not commonly implemented. This is partly to avoid changes in the frequency content of the sound material, but also because auditory motion can be sufficiently perceived without simulating Doppler cues. According to [Rosenblum 1987], and more recently [Lutfi and Wang 1999], intensity cues generated by an approaching or receding source dominate auditory motion perception, followed by interaural cues, and lastly Doppler-effect cues. [Lutfi and Wang 1999] showed that this relationship holds primarily for moderate sound velocities (up to 10 m/sec), and the Doppler effect becomes more important at velocities on the order of 50 m/sec. These findings explain why simulation of motion in virtual applications as well

as in psychoacoustic experiments is often done without considering the Doppler effect. Nevertheless, ways to implement Doppler shift [Chowning 1977; Ahrens and Spors 2011] and techniques to avoid it have been proposed [Peters and Braasch 2011].
VBAP, Ambisonics, and, to a lesser extent, Wave Field Synthesis pose constraints on listener placement. An ideal placement, in the so-called 'sweet spot', occurs when the listener is placed in the geometric center of the speaker array. Then the superposition of the acoustic waves is optimal, producing localization cues that when perceptually integrated yield the desired auditory spatial perception. On the contrary, when a listener is substantially closer to one speaker relative to the others, sound will be localized at the speaker that emitted sound first due to the precedence effect [Wallach et al. 1949]. Localization performance is further affected by early reflections and late reverberation. Early reflections can affect localization if they arrive within a time window of ≈ 1 μs for noise stimuli, but also at substantially longer delays for slow-onset stimuli [Hartmann 1983]. Increased reverberation time reduces localization ability for both real and virtual sounds [Marentakis et al. 2008; Giguere and Abel 1993; Begault 1992]. In addition, the direct-to-reverberant energy ratio, known to contribute to distance perception in rooms [Bronkorst and Houtgast 1999], cannot be easily controlled when sound spatialization systems are deployed in halls of variable reverberation time.
Consequently, it appears that identification problems may emerge when listeners are asked to identify trajectories from a given set, especially in situations where listeners are distributed in a concert hall, and reproduction is done using virtual auditory environments. Evaluation is therefore necessary in order to understand the design of auditory movement trajectories better.

2.2 Interaction Design Aspects

Control of auditory motion in music has been primarily implemented by sound engineers. In this practice, motion is either pre-programmed or performed according to the score in real-time using a mixing console or a human interface device. Gesture control of spatialization [Marshall et al. 2009] integrates control of spatialization within the performance. Performance gestures are typically tracked by sensors and subsequently used to dynamically control sound spatialization. Ancillary performance gestures can be used implicitly for this purpose or new gestures can be explicitly designed to control spatialization [Marshall et al. 2007]. These can be either direct or indirect manipulation gestures. In the first case, similar to direct manipulation, a 'performer of space' can point to a sound in the audience area, select it, and subsequently move it. In the second, performance movements are used to trigger spatialization as in the case of indirect control of spatialization through dancers' movements [Wijnans 2010]. Schacher [Schacher 2007] distinguishes the latter two approaches as top-down and bottom-up control of spatialization.
Direct manipulation of the location of sound by a 'performer of space' provides congruent visual stimulation to the audience. In a performance context, however, the level of congruency may be manipulated by the composer or become limited, e.g., when bottom-up control is used. Our work examines the implications of this link for the audience's perception and listening strategies in the case of direct manipulation of auditory movement. The design of interfaces for gesture control of spatialization for performers is a valid and complex question which is outside the scope of this paper.

2.3 Cross-Modal Integration

Audiovisual cross-modal interference has been mainly studied within the contexts of ventriloquism and of the interference caused by lip movements in speech perception.
Static and dynamic ventriloquism refer to the phenomenon of shifts in the perceived location and direction of movement of auditory stimuli due to the presence of visual stimuli. The interference becomes smaller with increasing intermodal spatial separation or temporal asynchrony and does not depend on the direction of visual fixation denoting a bottom-up process, at least when cognitive load is low [Thomas 1941; Spence 2007; Bertelson and Aschersleben 1998; Spence 2007; Soto-Faraco et al. 2003; Soto-Faraco et al. 2004; Spence 2007;

Meyer et al. 2005; Vroomen et al. 2001; Oruc et al. 2008]. Interference in the opposite direction with auditory stimuli affecting the location of visual stimuli has also been observed, but is much weaker. Auditory detection of spoken sentences in noise improves when sound is presented together with synchronous lip movements, due to the redundant information from speech-reading (also called lip-reading) and the temporal coherence of the stimulation by the two modalities [Sumby and Pollack 1954; Grant and Seitz 2000; Summerfield 1979; Grant and Seitz 1998]. Interference occurs even for spatially separated stimuli, but it is weaker [Driver and Spence 1994]. Destructive interference has been observed for incongruent audiovisual stimuli in the context of the McGurk effect [McGurk and McDonald 1976].

Cross-modal interference have been explained by the stimulus-driven interactions that arise in multisensory neurons. Another interpretation invokes covert spatial attention where a salient but non-predictive cue in one modality may attract multisensory covert attention to its location. Again, interference is normally stronger for spatially proximate stimulation, however other binding properties also play a significant role [Spence and Driver 2004; Spence et al. 2004; Spence and Driver 1994; Spence and Driver 1997]. The *unity assumption* [Vatakis and Spence 2007; Welch and Warren 1980] predicts that the more properties are shared between different modalities, the more likely the brain will be to treat them as originating from a common object or source. Commonality in time is probably the most important property for integration; although commonality in space, association upon co-occurrence or semantic congruency may also be of importance [Vroomen and Stekelburg 2011].

Gesture control of spatialization results in bimodal stimulation from the performer's gestures and the auditory motion. Consider the case of direct manipulation of sound within the listening area, which is the focus of this article. The performer's gestures are temporarily synchronized and spatially define auditory motion, but do not coincide spatially as the performer is on stage, and the audience sits in the concert hall. Despite the spatial discrepancy, stimuli are synchronized and semantically related; interference therefore cannot be ruled out. Similar predictions can also be made within the framework of covert spatial attention. Here, visual stimulation from gestures could be interpreted as a way to endogenously cue auditory attention, thus directing it spatially. The amount of interference might be limited when attention is directed to different regions for each modality, but as long as these regions do not alternate rapidly, a case for interference can be made [Spence et al. 2000].

Oruc et al. [Oruc et al. 2008] found that varying the attentional process led to substantial response variation when examining audiovisual motion perception. When subjects were instructed to report the direction of auditory motion in the presence of conflicting, temporally and spatially aligned visual cues, there was big interference between visual cues and auditory motion perception, which was larger when attention was divided between modalities and smaller when participants selectively attended to the auditory cues. In our context, the essentially semantic nature of the spatial cueing of auditory motion by visual cues might imply that attentional orienting might play an important role and needs thus to be considered.

2.4 Synthesis of the literature review and presentation of the experiments

According to this review, gesture control of spatialization gives rise to a complex perceptual situation where visual cues of variable spatiotemporal, and possibly semantic, congruency are provided simultaneously with auditory motion. It is consequently difficult to predict the extent to which information is integrated or perceived separately for each modality, whereas attentional orienting, resulting from different listener strategies, could be influencing the aforementioned processes. Given that gesture control is used increasingly often, it is important that the impact of the visual cueing on the listener's perception is understood. In addition, composers use extensively sound trajectories (for example Varèse, Stockhausen, Xenakis, Pierre Henry & Pierre Schaffer, but also more contemporary Roger Reynolds, and many more), but the extent to which they are identified is little studied, as most studies focus on detection of auditory motion, tracking or discrimination of its direction. Given the lower fidelity of localization cues in auditory virtual environments and

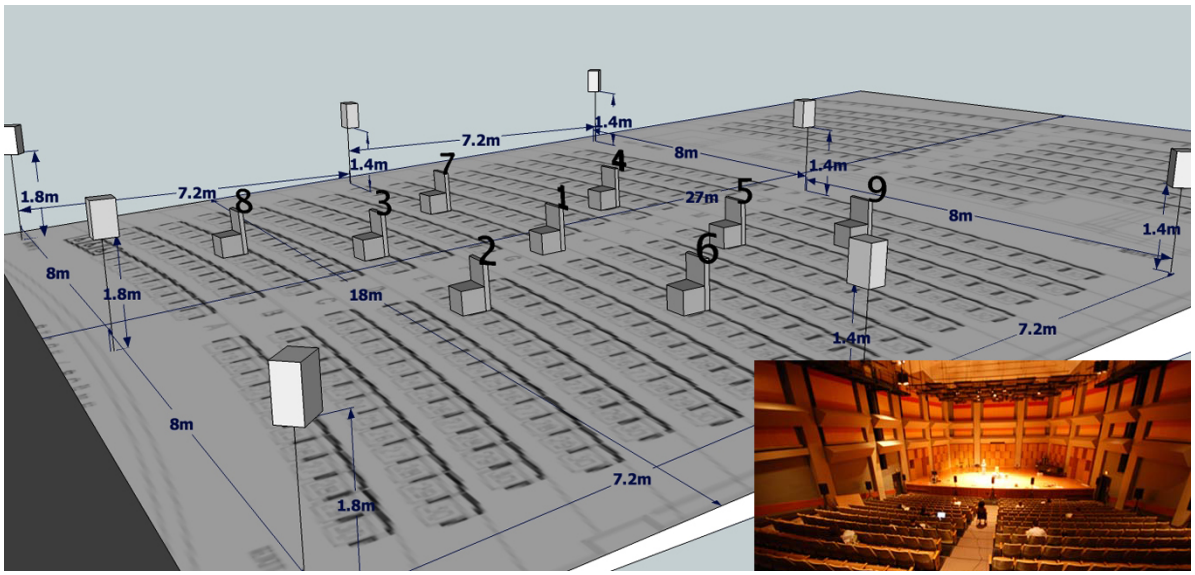

Fig. 1: A 3D view of the hall from the right corner of the stage. Hall dimensions, loudspeaker, and listener placement are illustrated. At the bottom right, a photo of the experimental set-up taken from the back of the hall is shown.

the potential influence of room acoustics and listener placement within the speaker array, it is reasonable to question whether sound trajectories are uniformly identified, and how much their identification is affected by the reproduction space and set-up.

We present two experiments designed to answer the aforementioned research questions. In the first experiment, performed in a concert hall, we evaluate the identification of spatial sound trajectories in the absence and presence of congruent visual cueing from the performer's gestures. This is done for participants seated in different listening locations, thus identification performance is estimated both within and outside of the optimal listening area. The instructional set here is kept open in order to observe the strategy participants adopt. In the second experiment performed in a controlled laboratory space, we examine the impact of listening strategies and cue congruency on auditory motion identification performance. Here, the congruency of audiovisual stimulation, the sensory focus of attention, and the attentional process involved (selective or divided) are manipulated. Contrasting the two experiments, we evaluate the impact of the reproduction setup and reflect on how participants deal with the experimental task. Both experiments employ an identification task: participants are asked to identify spatial trajectory shapes. This is done not only because sound trajectories are commonly used in compositional practice, but also because it allows for a more integrative task that is more relevant to our concern compared to the tasks of motion detection or discrimination.

Table I. : Reverberation times in seconds in the concert hall and the studio

Hz	63	125	250	500	1k	2k	4k
RT_{Hall}	2.3	2.0	1.7	1.8	1.8	1.7	1.4
RT_{Studio}	1.4	0.7	0.34	0.32	0.2	0.18	0.16

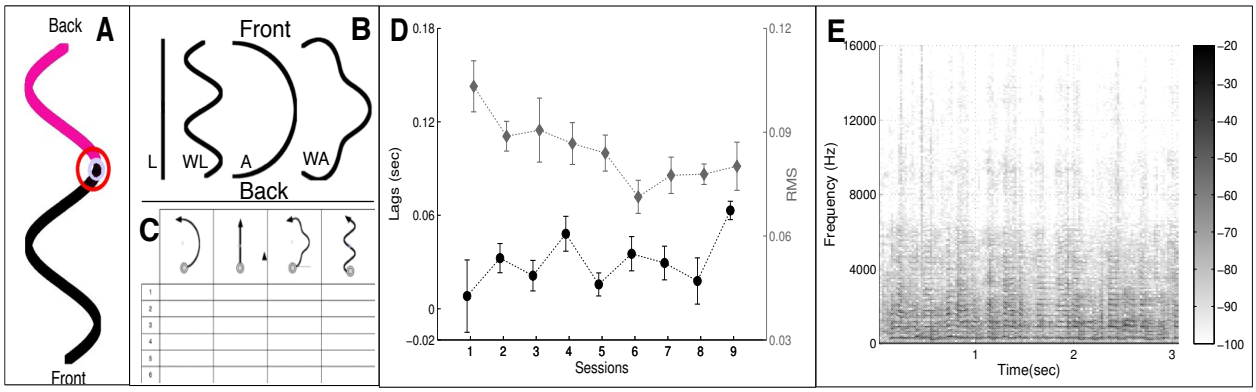

Fig. 2: A: Screenshot of the graphical user interface, B: The four auditory motion trajectories, C: The response sheet used to collect participants' responses. On left hand side is the trial number and on top the possible answers, D: Performer Lags and RMS error over the nine sessions, E: Spectrogram of the target sound stimulus

3. EXPERIMENT 1

3.1 Research Questions & Hypotheses

The experiment addresses the following research questions: 1. Does feedback from the performer's gestures assist the identification of auditory motion trajectories? 2. Does identification performance vary as a function of auditory motion trajectory and are some auditory motion trajectories easier to identify in comparison to others?, 3. Does listener placement in the concert hall affect identification performance?, and 4. Is identification robust to distractor interference?

To investigate our research questions we presented participants seated in nine different listening locations in a concert hall with four auditory motion trajectories shapes of a fixed target sound stimulus, presented with and without distractors, both in the presence and in the absence of congruent visual feedback from the performer's gestures. We measured the identification rate, defined as the percentage of trials in which participants reported each auditory motion trajectory shape in each condition. There were four independent variables: Auditory Motion Trajectory (4 levels), Listening Location (9 levels), Display Modality (2 levels), and Presence of Distractors (2 levels).

We hypothesized that in the absence of congruent visual stimulation the identification rate of the actually reproduced trajectory: H1. would vary as a function of auditory motion trajectory, H2. would deteriorate away from the sweet spot, and H3. would deteriorate in the presence of distractors. In the presence of congruent visual stimulation the aforementioned identification rate: H4. would improve due to visual cueing, H5. would be uniform in the different listening locations of the concert hall, and H6. would not be affected by the presence of distractors.

3.2 Procedure

Display Modality was a between subjects variable; two different groups of nine participants received the first auditory and the second bimodal stimulation in two sessions, each approximately one hour long. Variables Listening Location, Distractors, and Trajectory were within-subjects variables; all levels were presented within each Display Modality session. There were four trial repetitions, yielding 32-trial blocks. In each block, participants were seated as shown in Figure 1, and trials were presented in a random order. Participants moved to the next location after each block until they had performed the experiment in all listening locations, yielding a total of 288 trials per session. A response sheet was used to gather responses [Figure 2C]. The trial number was announced using a synthetic voice before each trial started to help in keeping track of the flow

of the experiment. Participants were instructed to identify the shape of the auditory motion trajectory of the target sound and to tick the appropriate box on the answer sheet. In the bimodal condition, participants were not specifically instructed concerning the modality to attend to, but were to act as if attending a concert. Prior to data collection, we presented the sound stimuli followed by four training trials.

3.3 Participants

Eighteen paid ($10) non-musicians (10 male and 8 female - mean age of 26 years) were randomly assigned in the two experimental sessions (9 in each). All participants reported having normal hearing. A male performer controlled the location of the target sound in the bimodal session.

3.4 Apparatus & Materials

Congruent audiovisual cues were provided in the bimodal condition. To achieve this, the hand movement of a 'performer of space' was tracked and used to control the location of sound in the audience. Hand movements occurred parallel to the floor and in the horizontal plane. For example, a hand movement to the front and away from the body would cause sound to move towards the back of the audience. Similarly, a hand movement from the performer's left to his right would yield a right-to-left auditory movement within the audience. Thus, the visual cues from the performer's hand movement on stage corresponded to the auditory motion, as perceived by the audience. The performer stood on stage and was aided by a graphical user interface [Figure 2A]. The screen was placed below him on the floor, so that hand movements were visible to the audience. In each trial, the desired movement trajectory was drawn on the screen, and a grey circle moved with the desired speed along the trajectory. A second red circle visualized the position of the performer's hand. The performer's task was to continuously capture the grey circle with the red circle, thus moving along the displayed trajectory at the desired speed. At the beginning of each trial, the grey circle was placed at the beginning of the trajectory, and the performer had to capture it for 1 sec before it started moving. To avoid exposing the performer to inconsistent auditory feedback, he wore headphones and listened to an independent binaural rendering of the sound scene that was aligned with his movements. In the unimodal condition, the red circle was simply automatically aligned with the grey one. The performer did practice with the Graphical User Interface, and was experienced in the use of controllers for spatialization and live performance. The accuracy with which the trajectories were performed was monitored throughout the experiment and was validated after the experiment, by estimating the lags that would maximize the cross-correlation between the performed trajectory and the actual trajectory shown on the Graphical User Interface. They had a mean value of 30 ms, standard deviation of 100 ms, and maximum value of 280 msec. The mean RMS error between the performed and shown trajectories after alignment was 0.09 (std=0.03), max = 0.2, within a movement range between -1 and +1 units. Considering the fact that the trajectories lasted about 3 sec, the velocity profile indicated by a mean lag of 30 ms is within an acceptable range, as is the RMS error, which was always less than 10% with a mean value of about 5% (see Figure 2D).
The target sound was an excerpt of a clarinet improvising in a contemporary music fashion and remained the same in all trials [Figure 2E]. It moved along four trajectories: a line straight across the middle of the audience (L), an arc (A) to the right side of the audience, as well as along two modulated variations, called wobbly line (WL) and wobbly arc (WA) [Figure 2B]. When spatialized in the hall, the wobbly line swung from the left to the right, whereas the arc and wobbly arc moved along the right side of the audience. All four trajectories had identical start and end points in the middle of the back and front of the seating area. Two distractors were included in half of the trials: the sounds of percussion and cello improvising, which were stationary and located laterally at the two sides of the hall. The clarinet, percussion, and cello sequences were played at 71, 58, and 58 dBA, respectively, measured at the center of the hall. Distractor levels were adjusted in pilot experiments at about a quarter of the loudness of the target sound, a level that would enable distractor interference with the target sound while maintaining a reasonable level of trajectory identification.

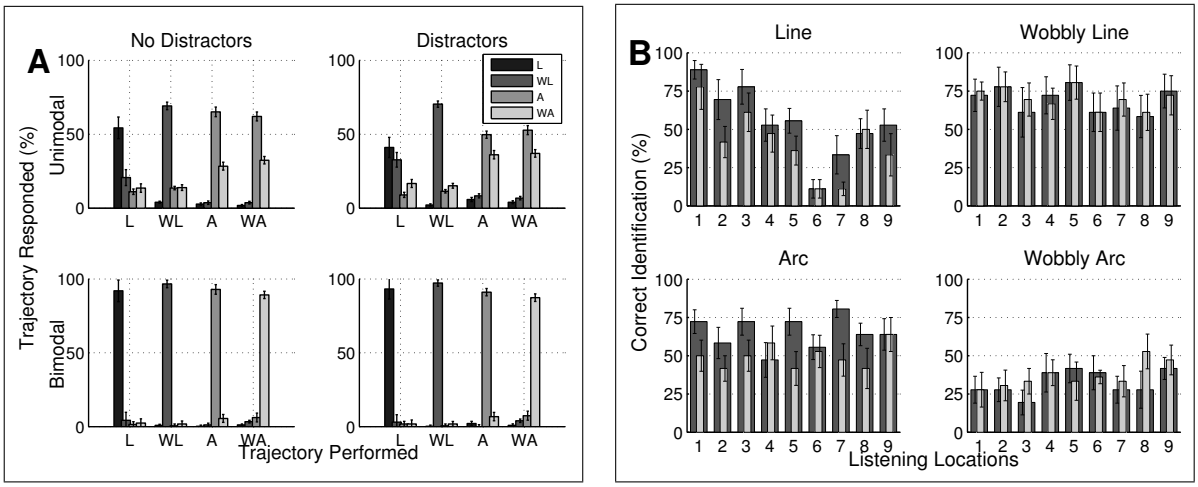

Fig. 3: A: Percent correct identification in the hall averaged across listeners and seats for the bimodal (upper) and unimodal (lower) conditions, without (left) and with (right) distractor sounds. Bars represent standard error. B: Identification performance in the unimodal display as a function of listening locations (seat numbers displayed in 1) and trajectory. Outer bars without and inner bars with distractors.

Auditory motion was simulated by updating the sound position within the speaker array, a procedure that induced changes to the distance-induced intensity cues, the interaural cues within the listening area, and the direct-to-reverberant energy ratio. The SPAT sound spatializer (IRCAM Forum) running in the Max/MSP environment (www.cycling74.com) was used for that purpose in the VBAP mode, as VBAP performs best for the 8-speaker configuration we used [Pulkki and Hirvonen 2005; Guastavino et al. 2006]. Each trajectory took about 3 sec. to complete, yielding a movement speed of 5-7 m/sec depending on the trajectory played. The Doppler effect was not simulated, both because this is a common practice for musical material, and because at these speeds its influence on auditory motion perception is very limited [Lutfi and Wang 1999; Rosenblum 1987].

Eight Meyer UPJ-1P loudspeakers were placed as in Figure 1. Each row in the hall is elevated. The difference in floor height between the last and the first row is 140 cm, seat height is 83 cm, and stage height 1 m. Loudspeaker height was adjusted so that all participants received direct sound, and loudspeaker orientation was aligned to the sweet spot. The three loudspeakers in the front were placed at 1.8m height, and the loudspeakers on the side and behind the audience were placed at a height of 1.40m relative to the floor height of the corresponding row of seats. Four computers and a Polhemus Liberty magnetic position tracker controlled the experiment. One provided the specification of each trial, the second displayed the trajectory tracking GUI, the third rendered the spatialized sound scene according to the data from the trajectory control computer, and the fourth together with the Polhemus Liberty ran the tracking software. The performer held the Polhemus sensor cable between his thumb and index finger so that the sensor extended just slightly out of his hand.

3.5 Results

The Unimodal and Bimodal conditions were analyzed with a repeated-measures Listening Location (9) × Trajectories (4) × Distractors (2) ANOVA with percent correct identification as dependent variable, defined as the identification rate of the actually reproduced trajectory [Table II].

Unimodal Condition: Percent correct identification was highest for the Wobbly line, followed by Arc, then Line, and then Wobbly Arc [Figure 3A] and while the ranking of the four trajectories remained the same in

Table II. : Significant effects and interactions of a repeated-measures Listening Location (9) × Trajectories (4) × Distractors (2) ANOVA performed separately for the Unimodal and Bimodal Conditions.

Condition	Factor	F-value	p-value
Unimodal	Distractors	F(1,8) = 8.038	$p = 0.022$
	Trajectory	F(3,24) = 32.510	$p < 0.001$
	Distractors × Trajectory	F(3,24) = 17.826	$p < 0.001$
	Listening Location × Trajectory	F(24,192) = 2.037	$p = 0.004$
Bimodal	Trajectory	F(3,24) = 5.669	$p = 0.004$

the two distractor conditions, performance dropped in the presence of distractors, but not for all trajectories [Figure 3B]. Significant main effects of Trajectory ($p < 0.001$), Distractors ($p = 0.022$), and significant interactions between Distractors × Trajectory ($p < 0.001$) and Listening Location × Trajectory ($p = 0.004$) were observed. Percent correct identification was significantly different between all trajectory pairs (t-tests, $p < 0.05$). Post-hoc t-tests indicated that each trajectory was influenced in a different way by the distractors, thus explaining the Distractors × Trajectory interaction. Averaged across listening locations, performance deteriorated significantly in the case of the Line (Δ=13%, $p < 0.001$) and Arc (Δ=15%, $p < 0.001$) trajectories, whereas the reduction for the Wobbly Line was negligible. A small (5%), but not significant, improvement was observed for the Wobbly Arc. The Trajectory × Listening Location interaction is explained by the fact that trajectories ranked differently in terms of identification rate at the different Listening Locations. For instance, percent correct identification for the Line trajectory was the highest in Listening Locations 1 and 3, whereas in Listening Locations 6, 7, and 9 it was the lowest. The prevailing of the Wobbly Line trajectory in the global ranking is mainly because it was consistently recognized in all Listening Locations. In detail, Wobbly Line, Arc, and Wobbly Arc were relatively consistently identified across Listening Locations (the range of percent correct identification across Listening Locations was 14%, 20% and 18% respectively), but this was not the case for the Line (range of 72%). There, performance in the sweet spot was significantly better than Listening Locations 5, 6, 7, and 8 (t-tests, $p < 0.01$), and not significantly different from Locations 2, 3, 4 and 9 (marginal effect for Locations 4 and 9, $p = 0.06$). A significant disadvantage occurred for Listening Location 6, where percent correct identification performance was 11%, significantly worse than all Locations save Location 7 ($p < 0.05$) and below chance level (25%). Performance was at chance level for Location 7. No differences in identification rates across Listening Locations were significant when considering the Wobbly Arc and Wobbly Line trajectories, whereas only identification in Location 7 was significantly better than Location 2 for the Arc trajectory.

Improvement Due to Bimodal Feedback: Evidently congruent bimodal stimulation improved performance, reduced confusions, and minimized differences due to Trajectory, Listening Location, or Distractors [Figure 2A]. Identification performance was significantly higher in the Bimodal compared to the Unimodal Display, $M_{Bimodal} = 92.4\%(SE = 0.013)$, $M_{Unimodal} = 52.4\%(SE = 0.13)$, F(1,16) = 487.035, $p < 0.001$. Only the effect of Trajectory was significant here; there was a small but significant identification advantage for Wobbly Line compared to the Arc and the Wobbly Arc ($p < 0.05$), but no other differences were observed.

3.6 Interim Discussion

Hypothesis H1 was verified as auditory motion trajectories were not uniformly identified. Even at the sweet spot position, and in the absence of distracting sounds, significant differences in the correct identification scores were observed: 88% for Line, 72% for Wobbly Line, 72% for Arc, and 27% for Wobbly Arc. Hypothesis H2 was partially verified as identification dropped significantly in Listening Locations away from the sweet spot, but not for all auditory movement trajectory shapes. The azimuth swing of the Wobbly Line, for example, was identified well throughout the listening area, as did, to a great extent, the movement along the circumference of the loudspeaker array of the Arc trajectory. The Line trajectory, whose reproduction

required a balance of inter-aural cues, was not identified well at locations away from the sweet spot. For example, in Listening Location 6 a bias towards reporting Wobbly Line occurred at a rate of 58%, and identification for the Line trajectory dropped to 11%. Such confusions were not uncommon in the case of the Wobbly Line and Arc trajectories, nonetheless the intended trajectory always received the highest identification score in the different Listening Locations. This was not the case for the Wobbly Arc trajectory. It was systematically misidentified as the Arc trajectory at an average rate of 52% and 62% respectively with and without distractors. The increased reverberation time in the hall, likely affected the reproduction of the intensity and direct-to-reverberant ratio differences that cue auditory distance perception [Bronkorst and Houtgast 1999]. H3 was also partially verified as introducing distractors affected trajectories to a different extent. It reduced performance by approximately 15% for the Line and the Arc trajectories, but did not affect the other two. Averaged across Listening Locations, Arc and Line trajectories were about 10% more likely to be identified as Wobbly Arc and Wobbly Line when distractors were present. For some specific combinations of Distractor, Trajectory, and Listening Location, performance on average increased in the presence of distractors e.g. Location 8 for the Line and Wobbly Line, Location 4 for the Arc, and Locations 2, 3, 7, 8, and 9 for the Wobbly Arc. In addition, averaged across Listening Location performance for the Wobbly Line trajectory increased by 5% in the presence of distractors. Because these effects did not prove to be statistically significant, they are not examined further.

H4, H5 and H6 were verified, as percent correct identification was significantly better in the bimodal compared to the unimodal display and no effect of Listening Location or Distractors appeared. As participants were not explicitly instructed concerning to which modality to attend, and given the magnitude of the improvement, participants may have based their responses on the visual stimulus, which would represent a task-relevant strategy given the spatial nature of the task. As mentioned earlier, the implications of such a strategy on the listening experience are not easy to predict. If spectators focus explicitly on vision, auditory motion information may be suppressed. An alternative strategy would be to leave visual attention spatially diffuse and attend to auditory stimulation. In such a case, visual stimulation would have likely acted to cue rather than dominate perception, giving rise to cross-modal interference, the extent of which cannot be predicted due to the relaxed cross-modal binding conditions in our experimental paradigm. The second experiment was designed in an effort to better understand the impact of the different spectator strategies in the aforementioned situations.

4. EXPERIMENT 2

In the second experiment, we investigated our hypothesis concerning the strategy the participants adopted in the first experiment, but also examined in detail the influence of visual stimulation from the performer's gestures on auditory motion perception, and vice versa, under conditions that manipulate the sensory focus of attention and the attentional process involved: selective or divided. To this end, we presented listeners with video recordings of the performer's movement from the first experiment and contrasted these with congruent and incongruent auditory motion. This allowed the estimation of the extent of the interference between modalities, in an ecologically valid manipulation; cues of mixed congruency can appear in a performance, and it is not always possible to know in advance how the performer's gestures relate to auditory motion, thus making the retrieval of relevant modality-specific information from memory necessary. We asked listeners to focus either on the visual or auditory cues and to report what they saw or heard. This procedure was repeated with listeners being informed on which modality they should attend either in advance or after each trial was completed. This enabled the assessment of how well sensory specific information can be retrieved from memory under selective, but also under divided attention conditions in the context of our experimental task.

Video recordings were used in the experiment for two reasons: Practical ones, such as in order to avoid performer confusion, performer fatigue, and inaccuracy due to the large number of trials in the experiment,

as well as limitations in space, and importantly also because this is a standard experimental protocol for examining cross-modal perception. In lip-reading as well as ventriloquism experiments, it is quite common, if not standard, to use video recordings and loudspeaker playback for visual stimulation, while the results are taken to construct hypotheses with respect to perception in real-world situations; see for example [Vatakis and Spence 2007; Vroomen and Stekelburg 2011; Sumby and Pollack 1954; Grant and Seitz 2000; Summerfield 1979; Grant and Seitz 1998; McGurk and McDonald 1976], to name a few. This study was performed in a studio room, and the same spatial audio software as in the concert hall was used. This time only one participant sat in the sweet spot in the middle of the speaker array in each session. There were six independent variables: Modality: Unimodal/Bimodal; Sensory Focus of Attention: Visual/Auditory; Attentional Mode: Pre/Post Cued, Distractors: Present/Absent; Audio Trajectory: Line, Wobbly Line, Arc, and Wobbly Arc; and Video Trajectory: Line, Wobbly Line, Arc, and Wobbly Arc as in the first experiment. We measured the rate (%) with which each trajectory was reported in each condition of the experiment.

4.1 Research Questions and Hypotheses

The second experiment addresses the following research questions: 1. What is the influence of congruent and incongruent audiovisual stimulation in the context of gesture control of spatialization? Does it give rise to crossmodal interference and to what extent for each modality? 2. How does attentional orienting influence the amount of observed interference? 3. How well can subjects recall modality-specific motion information? An additional research question was established with respect to the findings of the previous experiment: 4. Can we verify based on our results that participants based their responses on the visual feedback from the performer in the first experiment?

We formed the following hypotheses: H1: when attentional resources are directed to visual motion before the beginning of each trial, interference of auditory motion may emerge, but will be minimal; H2: Limited, if any, interference from auditory information will be expected in the a posteriori recollection of visual information due to its higher potency within our task context; H3: when directing attention to audition before the beginning of each trial, visual cues will interfere with auditory motion perception to a larger extent compared to H1. Consequently, we expect a mild performance improvement in the case of congruence, but at the same time a small disadvantage in the case of incongruence; H4: there will be limitations in the recollection of auditory motion when visual motion from the performer's gestures is presented simultaneously; although retrieving auditory motion information in the presence of visual motion can be performed when a priori knowledge concerning the relevant modality is available, it might be hard to perform a posteriori as access to the auditory cues might be limited likely due to post-perceptual interference from the higher-potency visual motion cues; H5: According to the rationale presented in the Interim Discussion, we also hypothesized that participants in the first experiment reported based on the visual feedback from the performer.

4.2 Participants

Sixteen paid ($25) non-musicians, who did not participate in the previous experiment, took part in the study (7 males, 9 females; mean age = 25 ± 7.9 years). All had normal hearing as per a standard audiogram procedure.

4.3 Procedure

Trials were blocked and presented in two experimental sessions on different days, lasting 1.25 hours each. Distractor presentation was counterbalanced across the two sessions: participants who were tested with distractors in the first session were tested without distractors in the second and vice versa. Within each session, the following trial blocks were completed: 1. audio-only, 2. video-only, 3. bimodal pre-cued, and 4. bimodal post-cued conditions. Because the video-only condition did not include distractors, it was split between the two sessions so that they would have the same duration. No video appeared in the audio-only

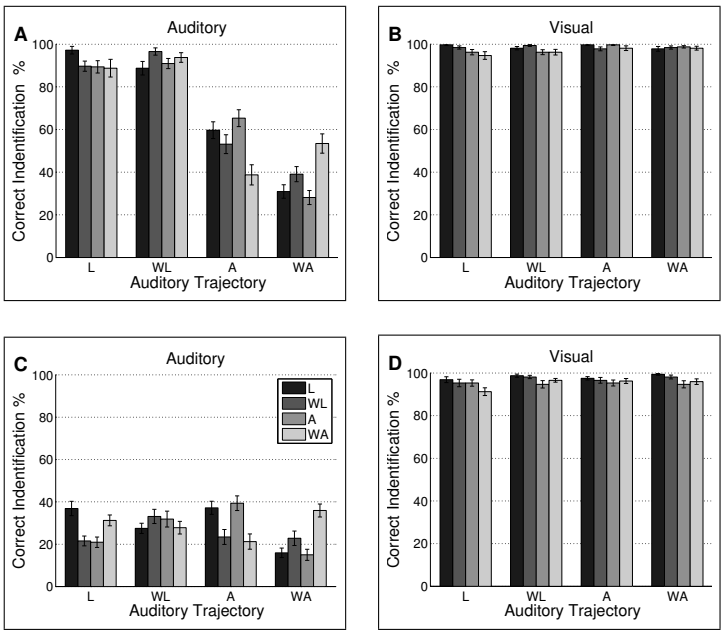

Fig. 4: Percent correct identification and standard error: A. auditory pre-cued, B. visual pre-cued, C. auditory post-cued and D: visual post-cued conditions. On the x-axis is the auditory motion trajectory and the colors in the legend correspond to the visual trajectory provided by the performer's movement. Below: Identification performance for the two modalities, the three attentional manipulations, and the four trajectories used in the experiment for the cases in which audiovisual cues were congruent. Data in both panels are averaged across distractor conditions. (L=Line, WL=Wobbly Line, A=Arc, WA=Wobbly Arc)

Auditory				
	L	WL	A	WA
Unimodal	99.1%	92.8 %	61.5%	38.7%
Pre-Cued	97.1%	96.5 %	65.3%	53.4%
Post-Cued	36.8%	33.1 %	39.4%	35.9%
Visual				
Unimodal	98.1%	95.0%	95.0%	99.3%
Pre-Cued	99.7%	99.3%	99.7%	98.1%
Post-Cued	96.9%	98.1%	95.3%	95.9%

condition. Blocks were counterbalanced across participants. In the bimodal conditions, the sensory focus of attention and the combination of visual and auditory trajectory (out of the 16 possible combinations of the four audio and four visual trajectories) to be displayed were randomly selected at each trial. Participants completed 10 repetitions for each combination of the independent variables within each condition. A break was given between the selective and divided attention tasks to minimize fatigue.

4.4 Apparatus & Materials

A Max/MSP patch on a Mac mini computer controlled the experiment. Participants sat on a chair in the middle of the speaker array, which had a radius of 2m. A monitor was placed on a box in front of them, lower than the speaker array, and participants responded using a mouse placed on a small table. The table and box were covered with a thick tablecloth. A scaled-down version of the audio setup used in the concert hall that would fit in the studio was implemented using the same software and loudspeaker array geometry (speaker height: 1.5 m), but different speakers (Genelec 2022A). The target sound and distractors were played at the same level as measured in the concert hall. Participants responded by clicking on a GUI: the video clip corresponding to each trial appeared at the top, and the icons showing the four trajectories appeared at the bottom. As only one subject participated in each session, we integrated response collection in the software to simplify data collection and provide the experimenter the ability to monitor performance in the course of the experiment. Participants were cued as to which modality to attend by means of an icon, in the form of an ear or an eye, presented on an empty screen for 1 sec before or after the trial was finished depending on the experimental condition. Responses could only be given at the end of each trial. A check mark appeared on the clicked icon to inform the participant that a response was registered. New trials commenced with a delay of 1.5 sec following a response.

Table III. : Results of the statistical analyses discussed in the Congruent Stimulation and Incongruent Stimulation subsections of the results section of Experiment 2

Case	Modal Focus	Factor	Statistic	p-value
I. Congruent Stimulation	Auditory	Condition	F(2,30) = 293.99	$p < 0.001$
		Trajectory	F(3,45) = 46.77	$p < 0.001$
		Condition × Trajectory	F(3,45) = 46.77	$p < 0.001$
	Visual	Condition	$\chi^2(2,30) = 7.48$	$p = 0.02$
II. Incongruent Stimulation	Auditory Pre-Cued	Auditory × Video Trajectory	F(9,135) = 12.72	$p < 0.001$,($\epsilon = 0.4$)
	Auditory Post-Cued		F(9,135) = 9.08	$p < 0.001$
	Visual Pre-Cued		F(9,135) = 2.74	$p = 0.03$, ($\epsilon = 0.4$)
	Visual Post-Cued		F(9,135) = 1.77	$p = 0.15$

4.5 Results

The presentation of the results and the statistical analyses is organized in three subsections. The first examines congruent stimulation, the second incongruent stimulation, the third reflects on the impact of vision on audition, and the fourth relates to the results of the first experiment. Whenever violation of sphericity was observed, a Greenhouse-Geiser correction has been applied and the ϵ value is given.

Congruent Stimulation: The percentage with which the displayed trajectory was correctly identified was analyzed separately for each modality in the three conditions: 1. Unimodal stimulation, 2. Pre-cued: containing trials where participants received congruent audiovisual stimulation and were pre-cued to the response modality, and 3. Post-cued: containing trials where participants received congruent audiovisual stimulation and were post-cued to the response modality [see Figure 4 and the affixed Table]. Data were analyzed by a repeated-measures Conditions (3) × Distractors (2) × Trajectories (4) ANOVA in the auditory modality and by a Friedman test in the visual modality, as identification scores were too close to the ceiling (Table III, Case I).

When the sensory focus of attention was *auditory*, percent correct identification varied in the examined conditions and depended on the auditory motion trajectory [Figure 4A, 4C]. The effects of Trajectory and Condition were significant ($p < 0.001$). Conditions differed significantly from each other in pair-wise comparisons (t-tests, $p < 0.01$). Trajectories Line and Wobbly Line were significantly better identified than Arc and Wobbly Arc (t-tests, $p < 0.001$), and Arc better than Wobbly Arc ($p = 0.01$). Percent correct identification depended on trajectory in the Unimodal and the Pre-cued bimodal condition, but not in the Post-cued condition. The improvement in identification in the bimodal Pre-cued condition relative to the unimodal one was negligible in the cases of Line and Wobbly Line, small in the case of Arc, and substantial in the case of Wobbly Arc. These two findings explain the significant Condition × Trajectory interaction ($p < 0.001$).

When the sensory focus of attention was *visual*, percent correct identification was high [Figure 4B, 4D] and the influence of the attentional process involved and the presence of auditory feedback were weak. Confusions were limited and mainly reflected similarities between the trajectories; averaged across conditions, Line was confused at a rate of 1.2% with the Wobbly Line, Wobbly Line at a rate of 1.8% with the Line and 0.6% with the Arc, Arc at a rate of 3% with the Wobbly Line and 0.6% with the Wobbly Arc, and Wobbly Arc at a rate of 4% with the Arc. The effect of Condition was significant ($p = 0.02$), and there were no effects of Trajectory or Distractors. Performance in the Pre-cued condition was significantly higher than in the Post-cued condition, but no other differences proved significant (Wilcoxon signed rank test).

Incongruent Stimulation: Percent identification identification was higher overall for congruent compared to incongruent cues in both the Auditory Pre- and Post-cued conditions, whereas congruency had little effect in the Visual Pre-cued and none in the Post-Cued condition [Figure 4B, 4D]. Each Attentional Mode and Target Modality was analyzed separately using repeated-measures Distractors (2) × Audio Trajectory (4) × Video Trajectory (4) ANOVAs to examine the impact of cue congruency. There was no significant effect

of Distractors so data are presented averaged over this variable. Of particular relevance is the Audio Trajectory (4) × Video Trajectory (4) interaction which when significant reveals that incongruent cues affected performance [Table III, Case II]. Following a significant interaction, we performed pairwise comparisons to see for how many of the incongruent trajectory combinations, significantly worse performance was observed in the incongruent in comparison to the congruent case.

When the sensory focus of attention was *auditory*, the interaction was significant both in the pre-cued and post-cued conditions. The mean correct identification score was significantly higher in the congruent compared to the incongruent condition ($p < 0.05$) in 10 out of the 12 comparisons in the pre-cued condition, and in 7 of the 12 in the post-cued condition. Evidently, cue congruency had a significant impact on identification performance when the sensory focus of attention was auditory.

When the sensory focus of attention was *visual*, the interaction was not significant in the post-cued case, but significant in the pre-cued case. The mean correct identification score was significantly higher in the congruent compared to the incongruent condition ($p < 0.05$) in 5 out of the 12 (3 out of the 12 when Wilcoxon test was used) comparisons in the pre-cued condition and in 2 of the 12 in the post-cued condition. A small effect of auditory motion trajectory appears in the pre-cued case and almost none in the post-cued case.

Impact of Vision on Audition: In order to understand the impact of visual cues on auditory motion judgments, we categorized the responses in the pre- and post-cued conditions in which the sensory focus of attention was *auditory*, and incongruent visual cues were provided as: vision-induced, audition-induced or other, depending on the modality in which the trajectory that determined each response was delivered. In the pre-cued condition [Figure 5A], the majority of the received responses were auditory-induced, except when the auditory motion was cued through the Wobbly Arc and Arc trajectories, and vision cues corresponded to Arc and Wobbly Arc trajectories, but not otherwise. Audition-induced responses ($M_A = 68.9(std = 5)\%$) were significantly higher than vision-induced ($M_V = 31.9(std = 5)\%$) $t(15) = 15.5$, $p < 0.001$ or other responses ($M_O = 8.4(std = 2)\%$) $t(15) = 42.2$, $p < 0.001$ ($p < 0.001$) in all cases, apart from the aforementioned exceptions in which vision-induced responses were significantly higher than auditory-induced ones ($p < 0.001$).

In the post-cued condition [Figure 5B], vision-induced responses ($M_V = 33.2(std = 5.4)\%$) were on average significantly more likely than audition-induced ($M_A = 27.6(std = 1.8)\%$), $t(15) = 3.71$, $p = 0.002$ or other responses ($M_O = 21.5(std = 1.8)\%$), $t(15) = 6.56$, $p < 0.001$. T-tests showed that in only 2 out of the 12 conditions did audition-induced responses occur more often than vision-induced ones. When broken down into the different conditions, vision-induced responses occurred significantly more often than did auditory responses when the visual motion trajectory was the Arc (Auditory Trajectory: Line, Wobbly Arc) and the Line (Auditory Trajectory = Wobbly Arc). In the other conditions, although the same tendency was observed, the difference was not statistically significant.

Comparison to Experiment 1: The response pattern that was observed in the bimodal condition in the concert hall was very similar to the one obtained in this experiment in the conditions in which participants were cued to attend to vision. In support of this, there was no significant difference between the identification rate in the bimodal condition in the concert hall and the identification rate in the pre- and post- cued conditions in the studio (Mann-Whitney U test, responses averaged over distractors and trajectories).

4.6 Discussion

H1 and H2, postulating weak interference from auditory cues when participants were pre- or post- cued to attend to vision, were supported by the lack of a significant difference between both the pre- and post-cued bimodal conditions and the unimodal visual condition where sensory focus of attention was visual. Interestingly, performance in the pre-cued condition was to a small, albeit significant, extent (2.7%) higher than in the post-cued condition. This indicates that attentional orienting may provide a small advantage for visual motion judgments in multimodal environments. H3, postulating a significant interference from visual

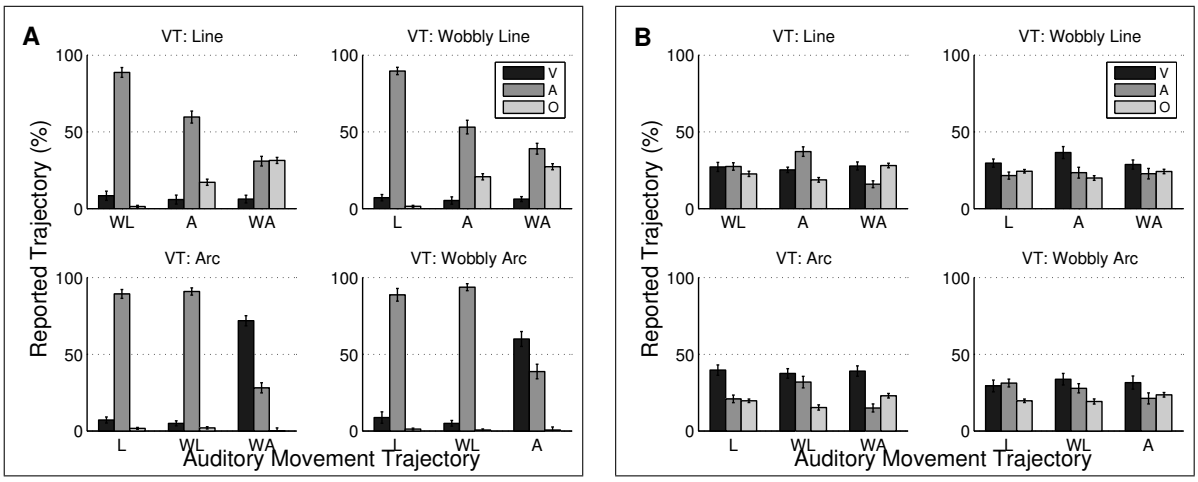

Fig. 5: Reported Trajectory Type [vision-induced (V), audition-induced (A) or other (O)] as a function of the auditory motion trajectory heard and the visual motion trajectory seen when participants were instructed to report auditory motion. In panel A the Pre-cued and in panel B the Post-cued Attentional Mode conditions are plotted.

cues in the pre-cued bimodal conditions when the sensory focus of attention was auditory, was supported since performance improved significantly compared to the unimodal auditory condition with congruent visual stimulation, and deteriorated significantly with incongruent. The improvement was greater ($\approx 20\%$) for the trajectory that had the poorest identification score: the Wobbly Arc. In the case of the Arc, although the trajectory was also not well identified in the unimodal condition, the magnitude of the improvement was smaller and not significant. This was because 2 out of the 16 participants showed the inverse tendency compared to the rest for this particular condition, with their performance worsening for bimodal compared to unimodal stimulation. When these two participants were excluded, the improvement for the Arc trajectory was also significant, $t(13) = 2.27$, $p = 0.04$, an improvement of 10%. Constructive and destructive interference is strongest therefore for auditory motion trajectories that were not well identified in the unimodal condition. This is also evidenced by the fact that the majority of the responses in the auditory pre-cued condition were audition-induced, apart from the case where auditory motion was cued by the poorly identified Arc and Wobbly Arc trajectories and visual motion by the Wobbly Arc and Arc, where vision-induced responses were the most likely. We hypothesize that this was because visual cues could be interpreted as potentially matching the ambiguous auditory percept. H4, postulating difficulties in the post-hoc recollection of auditory motion in a bimodal context, was also confirmed. This is evidenced by the significantly lower identification scores in the bimodal post-cued condition, where the sensory focus of attention was auditory, compared to the unimodal and pre-cued conditions, and the fact that vision-induced responses were significantly more frequent than audition-induced ones in this condition. Compared to [Oruc et al. 2008], where collocated and synchronized audiovisual motion information was provided, the detrimental impact of divided attention on the perception of auditory motion when attention was divided is much higher, even in the case of congruent cues. H5 was also verified, as no difference between the conditions where the sensory focus of attention was visual in this experiment and the responses in the bimodal condition in the sweet spot in the concert hall experiment emerged.

5. GENERAL DISCUSSION

The findings of the two studies can be interpreted to show that the listener's attentional strategy considerably influences the outcome of the exposure to audiovisual stimulation in the context of gesture control of

spatialization. Directing attention to vision yields very limited interference from auditory motion on visual motion judgments. Good identification and recall emerge due to visually induced judgments which are relatively unaffected by competing auditory motion. Directing attention to audition yields interference from the visual stimulation from the performer's gestures on auditory motion judgments. This works synergistically for congruent, and detrimentally for incongruent, stimulation. The amount of interference depends on the identifiability of auditory motion trajectories in the absence of visual cues, and is negligible in the case of unambiguous auditory motion trajectories. When ambiguity with respect to an auditory motion trajectory exists, interference from visual motion arises and is highest for visual stimulation that can be interpreted as compatible with the ambiguous auditory motion percept. Although congruent visual motion improved auditory motion judgments in the auditory pre- and post-cued conditions, the handicap induced by incongruent stimulation was not the same for all incongruent visual motion trajectories. This indicates that certain visual motion trajectories were easier to exclude than others, likely on the grounds of their perceptual distance from the perceived auditory motion trajectory. Compared to other studies in the literature, the amount of interference of visual cues on auditory motion judgments in the pre-cued condition is moderate. In [Oruc et al. 2008], the handicap between congruent and incongruent pre-cued bimodal conditions, where the sensory focus of attention was auditory, was more than 60%, whereas in our case it was never more than 30%. When no attentional orienting occurs, there is a tendency for visual motion to either override or heavily interfere with auditory motion. This is evidenced by the good recollection of visual motion in the post-cued condition, where the sensory focus of attention was visual and where no interference of auditory motion was observed, by the poor recollection of auditory motion information in the post-cued conditions where the sensory focus of attention was auditory, and by the increased tendency to respond based on visual information in the aforementioned condition. The results can be interpreted according to the modality appropriateness hypothesis, as vision has the highest potency in the context of a spatial task.

We would interpret the low recollection of auditory motion information in the post-cued condition, where the sensory focus of attention was auditory, as an indication that auditory motion information was suppressed when participants were not a priori advised to attend to it. As participants were explicitly instructed to report auditory motion in this condition, a hypothesis towards a response bias in the direction of reporting the visual trajectory is not plausible, rather it appears the auditory motion was not sufficiently encoded. This hypothesis is also supported by the fact that the same tendency occurred even for auditory trajectories that obtained high identification rates in the absence of visual cues and by the fact that although vision-induced responses were on average significantly more likely than auditory ones in this condition, they did not occur often enough to support a hypothesis towards response bias. Although congruent visual stimulation increased correct identification, this remained surprisingly low in the post-cued condition of our experiment. This interesting finding is in agreement with the observations of [Oruc et al. 2008]. We observed a difference of 50% between the post- and pre- cued conditions vs. about 10% in [Oruc et al. 2008]. This magnitude difference can be attributed both to the semantic nature of visual cueing in our task and to task specificity, because the trajectory combination and the cue congruency were randomized in each trial. [Spence et al. 2000] showed that when the spatial focus of attention changes between modalities across trials, participants have difficulty following, by comparison with the situation in which it remains fixed in trial blocks. Had a blocked design been used here too, the recollection rate may have improved, but would still remain lower than in the pre-cued condition, as the results of [Oruc et al. 2008] imply.

The results of the experiments show that gesture control of spatialization is promising, as congruent visual stimulation can assist the identification of auditory motion trajectories, irrespective of attentional orienting. In line with the observations that visual cues improve the perception of expressiveness, tension, and phrasing in music [Davidson 1993; Vines et al. 2006], it can be hypothesized that it would give rise to additional qualities that are not present when gesture control of spatialization is not performed. However, it can bias listener attentional strategy. The similar pattern that was observed in the bimodal condition in the sweet

spot in the concert hall and the audiovisual conditions where attention was cued towards vision in the studio, could be interpreted as an inclination to attend to visual cues in the bimodal condition in the concert hall, a tendency that may also emerge in a concert situation. Such a practice can significantly affect auditory motion perception, especially in the case of incongruent stimulation, as focusing on visual cues from the performer's movement will likely override incongruent auditory motion information, as evidenced by the inability of listeners to recall auditory motion trajectories when not cued to attend to them in advance. Our findings indicate that an auditory attentional strategy is optimal in such a context, because a synergistic advantage emerges in which listeners benefit from the congruent visual cues while the influence of the incongruent ones is mitigated, especially when auditory movement is carefully designed. The latter observation is important for artists as well as designers who are willing to employ such practice in their works. Incongruent audiovisual motion stimulation under conditions of no attentional orienting mainly leaves a visual motion memory, while auditory motion is only poorly preserved. The spatial focus of attention for each modality needs to remain consistent and change slowly in order to achieve good recollection in situations in which participants are expected to divide their attention across modalities.

The reproduction setting affected the identifiability of the trajectories used in our experiment. The confound between the difference in reverberation time (Table I), the different distances between the speakers and the listeners, and the different speaker types does not allow us to separate these factors. We can only attribute the lower identification rates observed in the sweet spot in the concert hall compared to the studio to differences in the reproduction setting, $M_{Studio} = 71\%$ to $M_{Hall} = 61.5\%$, t(23)=-3.53, $p = 0.0018$. As the identifiability of an auditory motion trajectory directly affects the interference from visual cues, perceptual calibration needs to be done with the listening setting in mind, which importantly also includes the location of the listener relative to the speaker array. The effect of distractors in the studio was negligible in comparison to the significant effect in the concert hall. There, an effect of distractors was observed even with static distractors at -13dBA relative to the target sound, and although it was detrimental for some trajectories, for others there was no effect, or even a tendency towards a small improvement was observed. It appears that apart from acoustic interference, geometrical relations between trajectories and distractor positions might need to be taken into account when interpreting the impact of Distractors, as especially static distractors can act as spatial references that can potentially aid identification. Unfortunately we cannot explain the effect of Distractors on auditory motion identification based on our limited experimental manipulation.

The spatial audio trajectory set used in this experiment was limited, and the results cannot be easily translated to design directives. Indeed, the trajectory set allowed us to observe that not all auditory movement trajectories share the same identification potential. This proved to be a critical aspect when examining the integration of audiovisual cues in the context of gesture control of spatialization. We argue that our results are generalizable in the sense that some ambiguity is to be expected when considering auditory motion trajectories, and in this sense the experiments provide insight into how such ambiguity gives rise to increased crossmodal interference. Concerning auditory movement trajectory design, within our forced-choice task and the small trajectory vocabulary, trajectories that fundamentally differed in their geometry such as the line, the wobbly line, and the arcs were relatively easy to differentiate and could be potentially used by artists and designers. The confusion between the Arc and Wobbly Arc trajectories indicates a certain difficulty in using distance cues for auditory motion identification, likely due to the fact that the ratio of direct-to-reverberant energy [Blauert 1997; Bronkorst and Houtgast 1999] cannot be easily controlled in enclosed spaces, as the surfaces in a room will inadvertently interact in a complex way with the cues the auditory virtual environment is creating. The success of more complex auditory motion vocabularies, containing for example two wobbly line trajectories that bounce in slightly different points or two lines starting from different offsets, is difficult to predict without further experimentation in relation to specific listening settings. As the identifiability of auditory motion trajectories is a critical aspect for both composers and designers, it is important that auditory motion trajectory design is researched further taking both spatial as well as acoustic parameters

into consideration.

6. CONCLUSION

We presented two studies that investigated the perceptual impact of gesture control of spatialization as a function of the identifiability of auditory motion trajectories, the congruency of audiovisual stimulation, the sensory focus of attention, and the attentional process involved (selective or divided attention). We found that visual cues from the performer's gestures significantly assisted the identification of spatial audio trajectories, so that it did not depend on the shape of the trajectory played, and it was not influenced by listener placement and room acoustics. It resulted though, in a tendency for spectators to visually orient their attention. Such an attentional orienting results in very limited interference from auditory motion and poor retention of auditory motion information. This can be problematic in the case of incongruent audiovisual motion stimulation, as vision-oriented spectators will retain a visual motion memory and auditory motion will be suppressed. On the contrary, selective attention to audition was found to yield good recollection of auditory motion that was little affected by the interference from the visual cues in the performer's gestures in the case of unambiguous auditory motion trajectories. The attentional strategy that maximizes auditory motion identification is therefore this of maintaining an auditory focus of attention. However, because this attentional strategy is subject to increased interference from incongruent visual motion trajectories in the case of ambiguous auditory motion trajectories, it is important that auditory motion trajectories are perceptually calibrated. Trajectories whose geometry differs in a plain and fundamental way are easy to differentiate, however more subtle differences could be difficult to register. Furthermore, the identification of spatial sound trajectory shapes in the absence of visual cueing is affected by the accompanying sound material and variations in the reproduction setting. Consequently, perceptual calibration might be difficult to achieve without perceptual evaluation, that importantly also takes listener placement in the speaker array into account.

We believe that the semantic information communicated by spatial sound trajectories is useful for artists as well as designers. The interference due to visual motion cues can unfortunately jeopardize auditory motion identification, even if these are not collocated with the auditory movement. This can be partially overcome by directing attention to auditory stimulation and leaving visual attention diffuse. In addition, it is important that: 1. auditory movement trajectory identification is maximized by using fundamentally different auditory motion trajectory shapes, 2. the amount of incongruent visual motion is minimized, as this will likely attract visual attention leading to poor encoding of auditory motion information, 3. congruent visual cues are used to improve ambiguous auditory motion trajectory identification, and 4. the listener is located in the sweet spot.

7. ACKNOWLEDGMENTS

This work was supported by Stephen McAdams's Canada Research Chair and a New Media Initiative grant to Stephen McAdams funded jointly by the Natural Sciences and Engineering Research Council of Canada and the Canada Council for the Arts. The experiments were performed while the first author was a postdoctoral fellow in CIRMMT at McGill University. We are grateful to Nils Peters for assisting with the implementation of the first experiment, to Joe Malloch for designing the user interface that was used by the performer in the first experiment, to Cathryn Gryffiths for testing the participants in the second experiment, and to three anonymous reviewers for invaluable comments.

REFERENCES

Ahrens, J. and Spors, S. 2011. Wave field synthesis of moving virtual sound sources with complex radiation properties. *The Journal of the Acoustical Society of America 130,* 5, 2807–2816.

Begault, D. 1992. Perceptual effects of synthetic reverberation on three-dimensional audio systems. *J. Audio Eng. Soc 40,* 11, 895–904.

Berkhout, A., de Vries, D., and Vogel, P. 1993. Acoustic control by wavefield synthesis. *The Journal of the Acoustical Society of America 93,* 5, 2764–2778.

Bertelson, P. and Aschersleben, G. 1998. Automatic visual bias of perceived auditory location. *Psychonomic Bulletin & Review 5*, 482–489. 10.3758/BF03208826.

Blauert, J. 1997. *Spatial Hearing: Psychophysics of Human Sound Localization.* Mit Press.

Bronkorst, A. and Houtgast, T. 1999. Auditory distance perception in rooms. *Letters to Nature 397*, 517–520.

Chandler, D. and Grantham, W. 1992. Minimum audible movement angle in the horizontal plane as a function of stimulus frequency and bandwidth, source azimuth and velocity. *The Journal of the Acoustical Society of America 91,* 3, 1624–1636.

Chowning, J. 1977. The simulation of moving sound sources. *Computer Music Journal* 3, 48–52.

Davidson, J. 1993. Visual perception of performance manner in the movements of solo musicians. *Psychology of Music 21*, 103–113.

Driver, J. and Spence, C. 1994. *Attention and Performance: conscious and non-conscious information processing.* Vol. 15. MIT Press, Chapter Spatial synergies between auditory and visual attention, 311–331.

Féron, F., Frissen, I., Boissinot, J., and Guastavino, C. 2010. Upper limits of auditory rotational motion perception. *The Journal of the Acoustical Society of America 128,* 6, 3703–3714.

Gerzon, M. 1992. Panpot laws for multispeaker stereo. In *92nd Convention of the AES.*

Giguere, C. and Abel, S. M. 1993. Sound localization: Effects of reverberation time, speaker array, stimulus frequency, and stimulus rise/decay. *The Journal of the Acoustical Society of America 94,* 2, 769–776.

Grant, K. W. and Seitz, P. F. 1998. The use of visible speech cues (speechreading) for directing auditory attention: Reducing temporal and spectral uncertainty in auditory detection of spoken sentences. *The Journal of the Acoustical Society of America 103*, 3018(A).

Grant, K. W. and Seitz, P. F. 2000. The use of visible speech cues for improving auditory detection of spoken sentences. *The Journal of the Acoustical Society of America 108,* 3, 1197–1208.

Grantham, D. 1986. Detection and discrimination of simulated motion of auditory targets in the horizontal plane. *The Journal of the Acoustical Society of America 79,* 6, 1624–1636.

Grantham, D. and Hornsby, B. 2003. Auditory spatial resolution in the horizontal, vertical and diagonal planes. *The Journal of the Acoustical Society of America 114,* 2, 1009–1022.

Grohn, M., Lokki, T., and Takala, T. 2002. Static and dynamic sound source localization in a virtual room. In *Proceedings of the 22nd AES International Conference: Virtual, Synthetic and Entertainment Audio.*

Guastavino, C., Larcher, V., Catusseau, G., and Boussard, P. 2006. Spatial audio quality evaluation: Comparing transaural, ambisonics and stereo. In *International Conference on Auditory Display.*

Harley, M. A. 1994. Space and spatialization in contemporary music. Ph.D. thesis, Schulich School of Music, McGill University, Montreal, Canada.

Hartmann, W. 1983. Localization of sound in rooms. *The Journal of the Acoustical Society of America 74,* 5, 1380–1391.

Lakatos, S. and Shepard, R. 1997. Contraints common to apparent motion in visual, tactile and auditory space. *Journal of Experimental Psychology: Human Perception and Performance 23,* 4, 1050–1060.

Lutfi, R. and Wang, W. 1999. Correlational analysis of acoustic cues for the discrimination of auditory motion. *The Journal of the Acoustical Society of America 106,* 2, 919–928.

Malham, D. 1999. Higher order ambisonics systems for the spatialiation of sound. In *International Computer Music Conference.* 484–487.

Marentakis, G., Corteel, E., and McAdams, S. 2008. Wave field synthesis evaluation using the minimum audible angle in a concert hall. In *AES 124th Convention, Amsterdam, NL.*

Marshall, M., Malloch, J., and Wanderley, M. 2007. Non-conscious control of spatialization. In *International Conference on Enactive Interfaces.*

Marshall, M., Malloch, J., and Wanderley, M. 2009. *Gesture-Based Human-Computer Interaction and Simulation.* Vol. 5085. Springer Berlin/Heidelberg, Chapter Gesture Control of Sound Spatialization for Live Musical Performance, 227–238.

McGurk, H. and McDonald, J. 1976. Hearing lips and seeing voices. *Nature 264*, 746–748.

Meyer, G., Wuerger, S., Röhrbein, F., and Zetzsche, S. 2005. Low-level integration of auditory and visual motion singals requires spatial co-localisation. *Experimental Brain Research 166*, 538–547.

Mills, A. W. 1958. On the minimum audible angle. *The Journal of the Acoustical Society of America 30,* 4, 237–246.

Oruc, I., Sinnet, S., Biscof, W. F., Soto-Faraco, S., Lock, K., and Kingstone, A. 2008. The effect of attention on the illusory capture of motion in bimodal stimuli. *Brain Research 1242*, 200–208.

Peters, N. and Braasch, J. 2011. Compensation of undesired doppler artifacts in virtual microphone simulations. In *Tagungsband der Deutsche Jahrestagung für Akustik*. DAGA.

Pulkki, V. 2001. Spatial sound generation and perception using amplitude panning techniques. Ph.D. thesis, Helsinki University of Technology, Espoo, Finland.

Pulkki, V. and Hirvonen, T. 2005. Localization of virtual sources in multichannel audio reproduction. *IEEE Transactions on Speech and Audio Processing 13,* 1, 105–119.

Reynolds, R. 2002. *Form and Method: Composing Music (The Rothschild Essays).* Routledge, New York, NY.

Rosenblum, L. 1987. Relative effectiveness of three stimulus variables for locating a moving sound source. *Perception 16,* 2, 175–186.

Saberi, K., Dostal, L., Sadralodabai, T., and Perrott, D. 1991. Minimum audible angles for horizontal, vertical and oblique orientations: Lateral and dorsal planes. *Acustica 75*, 57–61.

Saberi, K. and Perrott, D. 1990. Minimum audible movement angle as a function of sound source trajectory. *The Journal of the Acoustical Society of America 88,* 6, 2639–2644.

Schacher, J. 2007. Gesture control of sounds in 3d space. In *Proceedings of the Conference on New Interfaces for Musical Expression*.

Soto-Faraco, S., Kingstone, A., and Spence, C. 2003. Multisensory contributions to the percpetion of motion. *Neuropsychologia 41*, 1847–1862.

Soto-Faraco, S., Spence, C., Lloyd, D., and Kingstone, A. 2004. Moving multisensory research along: Motion perception across sensory modalities. *Current Directions in Psychological Science 13,* 1, 29–32.

Spence, C. 2007. Audiovisual multisensory integration. *Acoustical Science and Technology 28,* 2, 61–70.

Spence, C. and Driver, J. 1994. Covert spatial orienting in audition: Exogenous and endogenous mechanisms. *Journal of Experimental Psychology: Human Perception and Performance 20,* 3, 555–574.

Spence, C. and Driver, J. 1997. Audiovisual links in exogenous covert spatial orienting. *Perception and Psychophysics 59,* 1, 1–22.

Spence, C. and Driver, J. 2004. *Crossmodal space and cross modal attention*. Oxford University Press, Chapter Crossmodal Spatial Attention: evidence from human performance, 179–220.

Spence, C., McDonald, J., and Driver, J. 2004. *Crossmodal space and cross modal attention*. Oxford University Press, Chapter Exogenous spatial-cueing studies of human crossmodal attention and multisensory integration, 277–320.

Spence, C., Ranson, J., and Driver, J. 2000. Crossmodal selective attention: On the difficulty of ignoring sounds at the locus of visual attention. *Perception and Psychophysics 62*, 410–424.

Sumby, W. H. and Pollack, I. 1954. Visual contribution to speech intelligibility in noise. *The Journal of the Acoustical Society of America 26,* 2, 212–215.

Summerfield, Q. 1979. Use of visual information for phonetic perception. *Phonetica 36,* 314-331.

Thomas, G. J. 1941. Experimental study of the influence of vision on sound localizations. *Journal of Experimental Psychology 28*, 163–177.

Vatakis, A. and Spence, C. 2007. Crossmodal binding: Evaluating the unity assumption using audiovisual speech stimuli. *Perception and Psychophysics 69*, 744–756.

Vines, B., Krumhansl, C., Wanderely, M., and Levitin, J. 2006. Cross-modal interactions in the perception of musical performance. *Cognition 101*, 80–113.

Vroomen, J., Bertelson, P., and de Gelder, B. 2001. The ventriloquist effect does not depend on the direction of automatic visual attention. *Perception and Psychophysics 63*, 651–659.

Vroomen, J. and Stekelburg, J. 2011. Perception of intersensory synchrony in audiovisual speech: Not that special. *Cognition 118*, 78–86.

Wallach, H., Newman, B., and Rosenzweig, R. 1949. The precedence effect in sound localization. *The Journal of the Acoustical Society of America 21,* 4, 468.

Welch, R. and Warren, D. 1980. Immediate perceptual response to intersensory discrepancy. *Psychological Bulletin 88,* 3, 638–667.

Wijnans, S. 2010. The body as a spatial sound generating instrument, defining the three dimensional data interpreting methodology (3dim). Ph.D. thesis, Bath Spa: Creative Music Technology Department.

Zotter, F., Pomberger, H., and Noisternig, M. 2012. Energy-preserving ambisonic decoding. *Acta Acustica united with Acustica 98,* 1, 37–47.