Journal of New Music Research

ISSN: 0929-8215 (Print) 1744-5027 (Online) Journal homepage: www.tandfonline.com/journals/nnmr20

The impact of 3-D sound spatialisation on listeners' understanding of human agency in acousmatic music

Natasha Barrett & Marta Crispino

To cite this article: Natasha Barrett & Marta Crispino (2018) The impact of 3-D sound spatialisation on listeners' understanding of human agency in acousmatic music, Journal of New Music Research, 47:5, 399-415, DOI: 10.1080/09298215.2018.1437187

To link to this article: https://doi.org/10.1080/09298215.2018.1437187

JOURNAL OF NEW MUSIC RESEARCH
2018, VOL. 47, NO. 5, 399–415
https://doi.org/10.1080/09298215.2018.1437187

Routledge
Taylor & Francis Group

The impact of 3-D sound spatialisation on listeners' understanding of human agency in acousmatic music

Natasha Barrett[a] and Marta Crispino[b]

[a]Norwegian State Academy for Music, Norway; [b]Bocconi University, Italy

ABSTRACT

Commonly we hear sound without experiencing the associated visual activity to tell us how the sounds were made or from where they originate. In the larger context of electroacoustic music, and more specifically in acousmatic music where causal visual information is removed, it is interesting to investigate how a sense of human agency may be evoked by sound alone. An intuitive starting point would be to assume that the listener identifies source and cause through a sound's close proximity to a known archetype, yet composers often work with materials offering less obvious source clues, while expanding the spatial image over many loudspeakers. New technologies allow us to accurately control the movement of sound in space, and because we understand our own bodies in spatial terms, it is natural to ask how a sound's spatial behaviour influences our listening understanding. This paper investigates two research questions: can listeners identify human agency in 3-D sound, and if so, what are the most salient features involved in this process? Besides being of interest to electroacoustic composers, the topic is also relevant to the audio industry utilising high density loudspeaker arrays in cinema and other spaces where the projection of high resolution spatial imagery is possible.

ARTICLE HISTORY
Received 27 May 2017
Accepted 19 January 2018

KEYWORDS
Sound spatialisation; human agency; acousmatic music

1. Introduction

Music and motion are undoubtedly connected. Not only does music make us move, but musical parameters have been shown to stimulate listeners' imagined images of motion. Investigations into how listeners associate changes in sound with physical space and their bodies suggest a variety of connections. Simple connections include how temporal features (e.g. tempo or attack rate) are associated with speed or velocity, and how changes in pitch are associated with spatial ascent and descent. Studies also reveal more complex connections, where changes in one domain may stimulate changes in one or more different domains, for example that a crescendo, rather than a pitch rise, may stimulate upwards gestures (Eitan & Granot, 2006). In other words, spatiality and spatial change are implied without any Euclidian variation in the sound itself, where mental images of sound in motion are formed by changes in temporal-spectral information.

Such studies imply that embodied cognition is a central consideration. Embodied cognition is the theory of mentally re-coding sound into multi-modal gestural images involving a re-enactment of whatever we perceive (Godøy, 2006). Developing the discussion, Leman (2012), suggests that the body is a mediator between our environment and our personal experience, through which we accumulate a repertoire of gestures and gesture/action consequences. Directly relevant to our acousmatic musical discourse, Godøy (2010b) clarifies embodied cognition, proposing that it involves 'our capacity for having internal images of the world, as somehow originating in, but not necessarily truthfully reflecting external experience, because bits and pieces from lived experience may be recombined in novel and/or fictional ways'. This process described by Godøy is aligned with the techniques that some composers apply when creating sounds and musical structures.

Connecting embodied and real-world experiences, we can consider how real spatial movement may stimulate internal images of the world. Spatial movement may be experiential in terms of the body itself moving and interacting with its environment, or involve spatially defined elements of the environment either in a dynamic interaction with, or in static relation to, the perceiver. While these real spatial movements may not themselves create spatial sound, spatial sound can allude to these spatial movements. It is with this background that we can discuss the impact of 3-D sound spatialisation on listeners' understanding of human agency, and further, how a sounding spatial representation of the non-sounding physical movement that lies behind the causation of the

CONTACT Natasha Barrett, Postboks 5190 Majorstuen, 0302 Oslo, Norway. E-mail: nlb@natashabarrett.org
Research in this project was also carried out during a research position at the Department of Musicology, University of Oslo.

sound, can carry this information. Although composers and musicologists have at length discussed spatial information in a musical context, as far as we are aware, our study is the first to test for the influence of sound spatialisation on how listeners may hear human agency. Our work is relevant to the larger genre of electroacoustic music, which is an umbrella term for all kinds of contemporary music using computer generated or computer transformed sound (whether live or fixed media). It is of special interest to acousmatic music, which is one art form under the electroacoustic umbrella, whereby visual information is removed so as to free the listening imagination and prevent the eye dictating how we will hear.

2. Background studies

The ways in which human movement may relate to the organisation of sound can be informed by 'sound tracing' experiments (Godøy, Haga, & Jensenius, 2006) where subjects use their bodies to spatially describe what they hear. When considering the whole body, a study by Pedersen and Alsop (2012) showed that many sound stimuli were associated with an expected bodily response and a consistent interpretation of agency in the sound. For example, 'floating' sounds would stimulate 'floating' bodily movements. However, for some stimuli, the interpretation of agency was reversed. For example a 'punching' sound, rather than stimulating a punch action, instead stimulated the response of being punched.

In a study by Marentakis and McAdams (Marentakis & McAdams, 2013) the authors conducted a number of tests to ascertain whether a performer's own gestures assisted in the identification of spatial sound trajectories, and whether some trajectories are easier to identify than others. They also investigated the effect of congruent and incongruent audio-visual information. Although a detailed and relevant study in relation to our own work, their choice of spatialisation method and density of loudspeaker array casts questions over the results. Practical methods for spatialisation over multiple loudspeakers include Vector Base Amplitude Panning (VBAP), Distance-Based Amplitude Panning (DBAP) and ambisonics. A comparison of these methods will be presented later in Section 3.3.1. The authors chose to use VBAP (Pulkki, 1997) to spatialise a variety of trajectories over only eight loudspeakers. VBAP is a panning method, which is suitable for simple trajectories, but without the addition of secondary processing to emulate proximity cues, it is not possible to render perceptually clear trajectories other than panning at the perimeter of the loudspeaker array. Further, use of just eight loudspeakers offers a low angular resolution regardless of the chosen spatialisation method, far lower than that of our auditory perception[1] and insufficient to capture the spatial variations that the test addressed. Although not possible to draw conclusions from their study concerning listening alone, in their experiments involving bimodal feedback (an interaction between sound and sight), congruent information was seen to improve performance. If we assume that the listeners were not able to clearly hear the differences between auditory spatial features, the added visual stimuli appears to have convinced them to hear information that was absent.

In another study concerning how auditory-visual cues affect spatial sound streaming (or the segregation of sound in space), Shestopalova et al. (2014) showed that movement-congruent visual cues did not necessarily strengthen the effects of spatial separation. They also conclude that the congruency between auditory and visual stimuli may use mental resources that could have been utilised for more accurate auditory processing, supporting models of modality-specific competition for perceptual awareness.

More generally, we find studies that assess the effect of bimodality on the subjective experience of the music. Although many studies demonstrate that we primarily use visual information when making judgements about music performance (e.g. Tsay, 2013), cross-modal interactions may involve a more complex network of connections. Vines, Krumhansl, Wanderley, and Levitin (2006) demonstrated three contrasting scenarios: an independence of information transmitted through the visual and auditory domains; that an experience of tension (emotion) and phrasing (structure) could be enhanced by bimodal cues; and that the addition of visual information can dampen the intensity of emotional response.

In our own work, we can we can summarise that, (a) visual information should be removed if listeners are to successfully engage in the challenging task of spatial listening, (b) that an accurate and robust spatialisation method is required when creating the test stimuli and (c) we can anticipate a relationship between agency in the sound and listeners' understanding, but that the mapping between the two may not appear immediately straight forward.

3. Method

3.1. Background

Our aim is to investigate the role of sound spatialisation in listeners' mental representation of human agency when they hear sound without seeing any visual causal

[1] In the horizontal plane, our spatial discrimination has been tested to occupy a range of between 0.75-10 degrees depending on the source angle in relation to the listening direction (Blauert, 1997).

information. Listeners' verbal or written descriptions of what they hear are notoriously inconsistent, even when they are in agreement. Our tests are therefore designed to avoid the need for a descriptive language. For simpler experiments, such as listeners describing qualities of room reverberation or instrumental timbre, it is possible to apply verbal labels, for example in the work of Berg and Rumsey, on the evaluation of perceived spatial quality (Berg & Rumsey, 2003). Yet, even in the field of acoustics, we find that apparently simple concepts such as 'envelopment' are ambiguously defined (Berg, 2009).

Our experiment involves a more complicated situation where listeners are not describing their direct acoustic perception but rather their embodied experience, which in turn adds a layer of abstraction and difficulty. In this situation, it is unrealistic to ascertain the number of test subjects required to obtain meaningful results when using verbal descriptors. We also assessed the effort and shape terminology from the well-known Laban Movement Analysis (Davies, 2007), but the terminology was too vague for our purpose and inappropriate for what we were testing for.

We considered allowing listeners to allocate each sound a score, indicating how strongly each evoked human agency in relation to the other sounds. However, as we already knew that listeners' spatial audio skills would span a large range, that the tests would be challenging, that it would be ideal to ascertain the degree of certainty in the results and to detect any self-contradictions, we designed the following pair-wise evaluation of sound stimuli (an expanded discussion is provided in Section 3.2):

- A number of test stimuli were paired into all possible combinations.
- Each listener was presented with a selection of pairs of sounds. They were asked to select which, from each pair, most evoked a feeling of human causation or human physicality (see Section 3.4 for more detail as to the choice of words).
- The pairs were chosen randomly and independently for each user.
- The order in which the sounds were played was randomised.

The results were analysed by a generalised version of the Bayesian Mallows model (Vitelli, Sørensen, Crispino, Frigessi, & Arjas, in press), able to take into account the non-transitive patterns of the data (Crispino, Vitelli, Barrett, Arjas, & Frigessi, 2017). In particular, from the pair-wise results the method is capable of detecting the uncertainty of the order for all stimuli, as well as indicating the degree of non-transitivity or misreporting within a single subject's test results.

3.2. Statistical method

In the context of ranked data, given a set of n items, a ranking of these items is simply an ordering of the items by preference. For example, imagine that a customer needs to rank, from the most preferred to the least one, the soft drinks in the set $D = \{$Coke, Fanta, Sprite, Tonic Water$\}$. A possible outcome of her choices can be the following ordering $\boldsymbol{o} =$ (Fanta, Sprite, Coke, Tonic water). If each item is represented by an integer indicating its original position in the set D, the ranking would be $\boldsymbol{r} = (3, 1, 2, 4)$. To clarify, $r(i)$ is the rank given by the customer to the item that occupies the $i-$th position in D. For example $r(2) = 1$, means that Fanta (that occupies position 2 in D) is ranked first by the customer.

Sometimes it may be possible for a customer (assessor in general) to compare many items at the same time, to assign ranks to all of them, and thus to produce a unique ranking, like in the previous example. However, when differences between items are small, or the number of items under comparison is too large, it can be difficult for a customer to provide a full ranking. In such cases, a possible experiment is to let the customer repeatedly compare the items in pairs, that is to rely on the method of paired comparisons (David, 1963). It has been shown that the method of paired comparisons is the best option in the case of perceptually misleading problems. As David mentions, '(...) This method is used primarily in cases when the objects to be compared can be judged only subjectively; that is to say, when it is impossible or impracticable to make relevant measurements in order to decide which of two objects is preferable'.

In addition, it is well established (see e.g. Huber, Payne, & Puto, 1982; Simonson & Tversky, 1992) that in decision-making the number of choices available affects the outcome of a decision (e.g. which sound is preferred), not necessarily because of a true preference, but often because people get confused and are unable to make a rational decision.

For these reasons, we decided to rely on a pair comparison experiment in order to collect pair-wise preferences data, which are then used to infer a common ranking of the items.

However, this method makes it possible for the assessors to contradict themselves: for example, one may prefer A over B, B over C, and yet C over A. We refer to these pathological situations as non-transitive data patterns. Clearly, if such a pattern occurs in an assessor's set of pair-wise preferences, a unique ranking cannot be readily identified. The method we exploit to analyse the data, explained in detail in the next section, is able to take into account this problem.

3.2.1. The Bayesian Mallows model for non-transitive data

The Mallows model (Mallows, 1957) is one of the most used models for rankings. It can be seen as the equivalent to the Gaussian model but for rankings, in that the likelihood of a given ranking, $\boldsymbol{r}$, decreases geometrically as the distance between $\boldsymbol{r}$ and a mean (or consensus) ranking, $\boldsymbol{r}^*$, increases. The main parameter of the model is $\boldsymbol{r}^*$, which represents the consensus ranking of a pool of assessors, that is, the ranking with highest probability.

The Bayesian Mallows model, recently developed by Vitelli et al. (in press), is simply the Bayesian[2] version of the original model. For the scope of this paper it is important to keep in mind that the Bayesian paradigm quantifies probabilistically the uncertainty relative to the parameters of interest. This feature is of fundamental importance, especially when the experiment involves human perception, which is often noisy and ambiguous.

The Bayesian Mallows model can be adapted to pair-wise comparisons data. In this case it estimates not only the consensus ranking, $\boldsymbol{r}^*$, but also the individual latent rankings, here denoted $\boldsymbol{r}_1, \ldots, \boldsymbol{r}_N$, of the N assessors involved in the experiment, with uncertainty. For example, suppose that an assessor, when comparing items in the set {A,B,C,D,E}, gives the following preferences: {(A < B), (B < C), (C < D)}.[3] These three preferences are not enough to obtain the assessor's ordering of all the items under consideration. In fact, she didn't express any preference about item E. As a consequence many orderings of the items are possible, for instance the ordering (E,D,C,B,A) is as plausible as (D,C,B,A,E). In such cases, the Bayesian Mallows model exploits information given by other assessors to estimate the individual and latent orderings of all the assessors, through a procedure commonly known in the Bayesian field as hierarchical modelling. The estimated individual rankings can be of enormous interest, for example to study how individual preferences change with user related covariates.

Importantly, the Bayesian Mallows model is capable to handle heterogeneity in the population, and in particular to cluster the assessors into more homogeneous subsets, each sharing a consensus ranking of the items.

A recently developed generalisation of the Bayesian Mallows model (Crispino et al., 2017) has also the ability to handle non-transitive pair-wise patterns in the data, and to estimate the individual latent rankings of the assessors, with uncertainty , even if they contradict themselves. The main assumption of this model is that non-transitive patterns occur because assessors make mistakes and switch the order between two compared items.

The choice to rely on these newly developed models is crucial. First, to our knowledge, there are no other methods that have the ability to, simultaneously, estimate the consensus ranking, the individual rankings, and the clustering patterns of data in the form of pair-wise comparisons. The consensus rankings represent the main interest in this work, but the possibility to inspect the individual preferences, when correlated with covariates, helps in the interpretation of the results. Second, the ability to detect and correct the non-transitive patterns in the data, and to quantify uncertainty, is of enormous importance in our experiment, since the collected data are very noisy and ambiguous. Not considering this aspect, would have led to biased results.

3.3. Creating the test stimuli

When making the test stimuli we considered spatialisation method, sound source, and motion source for the spatial trajectory. Each of these factors are connected in some way. For example, sine-tones are spatially vague regardless of precision in spatialisation method, while a clearly recognisable source, such as that of running footsteps on gravel, will carry extra information biasing listeners' spatial interpretation. The following sections explain these three sides of the test stimuli.

3.3.1. Spatialisation method

Three spatialisation methods can be considered for our experiment: VBAP, DBAP (Lossius, Baltazar, & de la Hogue, 2009) and higher-order ambisonics (HOA) (Daniel & Moreau, 2004). Sound diffusion of stereo or multichannel composed fixed media, also known as performed sound spatialisation, is common practice in the acousmatic music tradition as a way of creating spatial sound imagery over a number of loudspeakers set up in an acousmonium configuration (Harrison, 1999). The method is useful in artistic work, but the spatial projection is too inaccurate for our purpose, and moreover we must avoid performed spatial interpretation. Sound diffusion performance was therefore not considered as one of our possible methods. VBAP is a panning method for an equally spaced loudspeaker array that produces virtual sources using one, two, or three loudspeakers at any one time. If we were interested in a listener hearing a point sound at a specific angle on the perimeter of a loudspeaker array, then VBAP would be a good choice of spatialisation method. Circular motions of a mono sound can be easily created. An illusion of the sound appearing at a varying proximity can be somewhat achieved by applying post-processing that emulates real-world cues. This process-

[2] For a simple review of Bayesian statistics, we refer to the excellent book by Hoff (2009).
[3] The symbol '<' is here used as a preference relation: A<B must be read 'B is preferred to A'.

ing includes the addition of artificial reverberation, gain attenuation and low-pass filtering. However, changes in proximity also involve the perception of changes in image size (where in the real-world, we hear sounds that are in close proximity as as larger), and although we can modify the VBAP method to spread the image, achieving a convincing change in proximity is far from easy. Another problem is that fast moving sources may create audible discontinuities due to the precedence effect, although this is more problematic when using fewer loudspeakers and non-symmetrical geometries. Also, if the spatialisation experiment is to be repeated in different spaces (moving spaces inevitably involves changes in loudspeaker array) then all spatialisation and post-processing needs to be either re-rendered, or implemented in real-time using suitable computer resources.

DBAP is a method that extends the principle of equal intensity panning from a pair of speakers to a loudspeaker array of any size and geometry. Speakers can be placed in any location in the space (angle and proximity to a defined location). DBAP can be particularly useful for custom loudspeaker arrays especially when the speakers themselves help define the trajectory, such as in a sound-art installation. However, non-restrained spatial projections require an equally spaced loudspeaker array, where DBAP will suffer from similar problems as VBAP.

HOA is a way of synthesising 3-D sound over a loudspeaker array. HOA involves a two-step process of spatially encoding the information in a spherical harmonic representation irrespective of the loudspeaker array, and then applying the appropriate decoder for the array to be used. The success of HOA over VBAP will partly depend on the chosen order of ambisonics (the higher the order, the greater the spatial resolution), the desired listening area, and in choosing the correct decoding strategy (which requires basic knowledge of ambisonics). A number of different decoding strategies are available, and the most appropriate is chosen for the given loudspeaker layout, room acoustic and type of music (e.g. Marentakis, Zotter, & Frank, 2014; Frank, 2014). At a sufficient order of ambisonics, over sufficient loudspeakers, HOA creates an accurate projection of spatial information, especially for a centrally located listener. Although HOA suffers from similar distance projection problems as VBAP, the same post-processing can be applied. Additionally, the apparent size of the image can be adjusted by scaling the relative weights of the spherical harmonics for efficient phantom source widening correlated with change in proximity. Although fast moving sources may create audible discontinuities similar to VBAP, ambisonics addresses all loudspeakers even when the sound is localised at one point in space, and the effect of these discontinuities are decreased. Further, the independence of encoding and decoding stages allows spatial encodings to be decoded over different loudspeaker arrays without relying on real-time spatialisation.

Having evaluated the options we therefore applied a 6th order 3-D spatial encoding, decoded over the 47-loudspeaker 3-D array at the motion capture lab at the Department for Musicology, University of Oslo (Figure 1(a)–(c)). The decoder used was the Max-rE dual-band energy preserving decoder with a cross over frequency set to 400 Hz (Zotter, Pomberger, & Noisternig, 2012). This method had been tested in a previous project and was found to be the best option for the available loudspeaker system (Barrett, 2016). Although nearfield coded HOA (Favrot & Buchholz, 2012) can in theory recreate the sensation of changes in proximity resulting from variations in the curvature of the approaching wave-front, based on the author's previous experience this method is unsatisfactory in a practical application, requiring the use of regularisation functions found to have an impact on the reproduced sound field, explained briefly in Carpentier, Barrett, Gottfried, and Noisternig (2017). Instead, variations in proximity were projected via three sets of perceptual cues: (a) by changing the relative weight of the spherical harmonic components: when close-up, the sound is heard as larger and more enveloping, when further away it is heard more as a point source, (b) by changing the gain of the source in relation to distance and (c) by changing the high frequency content of the source in relation to distance. All three modulations correlate with our perceptual understanding of proximity variations in the real-world. The Doppler effect (which is a pitch shift resulting from a sound moving at speed in relation to the listening position) was not used. In our everyday experience, motorised vehicles are the most common source of Doppler shifts, which would add an inappropriate layer of connotation to the test sounds. Reverberation was also avoided so as to not risk the addition of inappropriate source implications.

3.3.2. Motion sources

If our study shows that a human imprint can be reflected in spatialised sound, we are then interested to know what features may be carrying this information. We begin with a hypothesis that human motion archetypes may encapsulate what we are interested in finding.

Human motion archetypes are those with which we are most familiar in terms of our own bodies, and serve as a natural starting point. These archetypes are actions produced by, to name few examples, a swing of the arm, swaying the head and torso, turning, throwing, catching, hitting, the drumming of fingers, punching, stroking, jumping or running. Connected to these archetypes are further actions that are the causation of sound produc-

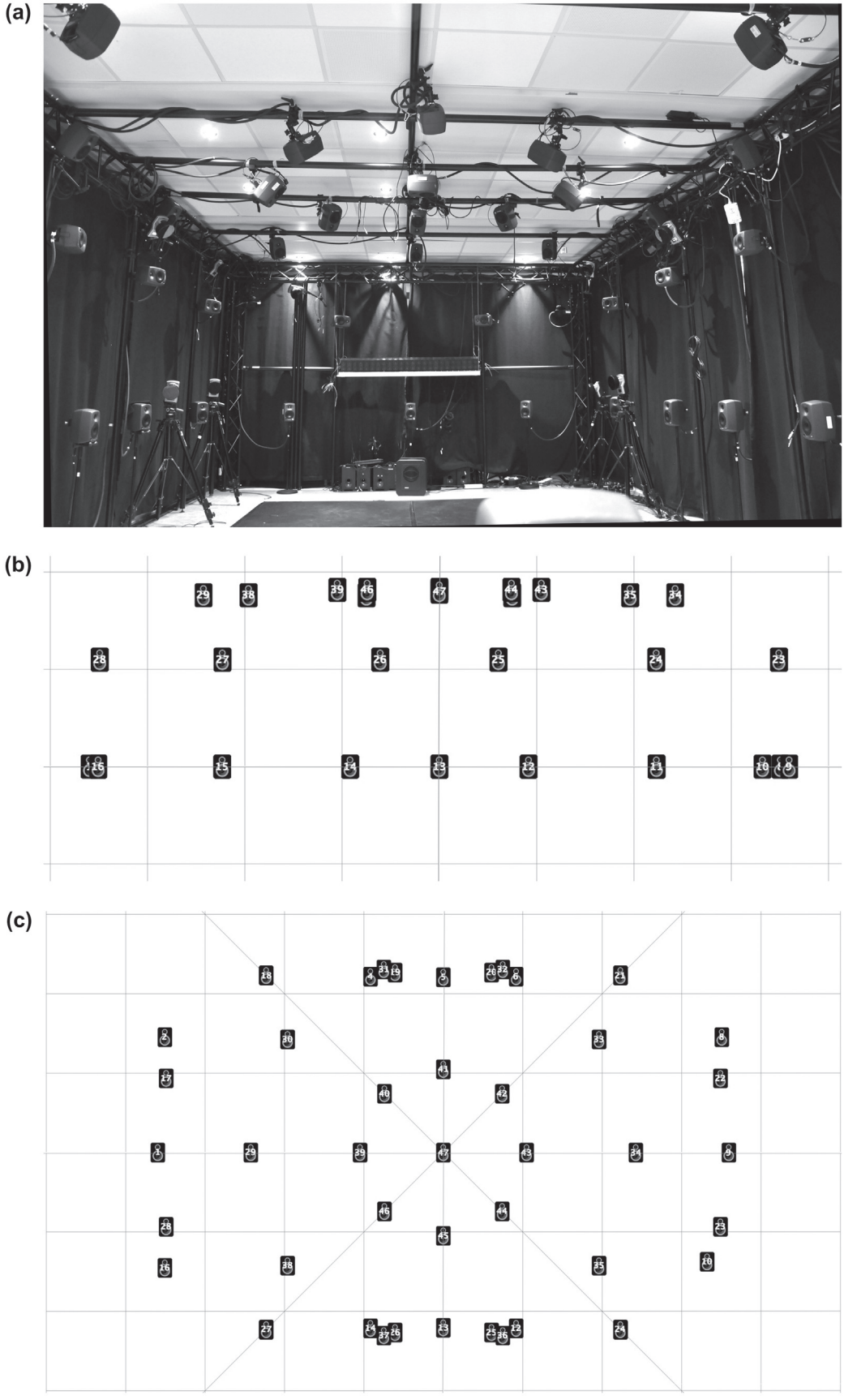

Figure 1. (a) 47-loudspeaker 3-D array at the motion capture lab at the Department for Musicology, University of Oslo. (b) 47-loudspeaker 3-D array modelled from the side, rotated 45 degrees, visualised in IRCAM's spat.viewer (part of the Spat package, IRCAM 2017). (c) Speaker layout from above.

tion: hit, scrape, push, pull, blow and a variety of smaller and larger motoric variations on these themes. In all cases, movement is constrained within a peripersonal space even though the whole body itself may be moving at speed in relation to a larger space, as in the case of running. For most archetypes, macro- and meso-movements are felt and seen, whereas micro-movements are more likely to be felt in space and time and are features less characteristic of mechanical or non-human action.

After considering these possibilities we chose the articulation of a music instrument as the source motion. This causation archetype was useful as it would be naturally controlled within a defined spatial zone, would give the human body an object with which to interact (or in other words afford the motion a 'purpose'), and being performed by a musician, would inevitably contain micro-variations correlated with the larger action of instrumental sound production.

To 'hear' human agency, we need to somehow 'hear' human movement. Normally we hear the result of human movement, and not the movement itself. Therefore, to create the test sounds, it is necessary to capture movement, or describe it, as spatial data, and then sonify the data to make it audible.

3.3.3. Motion capture

In a recording session, different performers executed various motion archetypes on their instruments: a cellist played a variety of bowed articulations; a percussionist articulated a cymbal with his hands, and for a non-musical contrast, one person threw a tennis ball. 3-D motion data from these performances was captured using the Qualisys optical motion-capture system and eight Oqus 300 cameras. Passive markers were placed at critical locations over the performers bodies, and motion data was recorded at a rate of 250 Hz with a spatial resolution of <1 mm. The camera system tracks the location of each marker, recording a dataset of 3-D coordinates at intervals of 4 ms. The temporal and spatial resolution was important so as to capture micro activity. Although we may not see micro movements when watching another person, these movements are understood in our own bodies, and can further be made audible when sonifying data that has captured these movements. After assessing all recordings, a cellist bowing a single down-bow action over a 'double-stop' was chosen as the motion archetype, using only the one marker located on the lower side of the right hand. Listeners would not be expected to identify the origins of this action in terms of a cellist or a cello. Rather, the action embodied 'push' (with some friction resistance), changes of direction, changes of speed, a 'throw' as the bow leaves the strings, as well as creating a human motion archetype embodying small micro-movements.

3.3.4. Sonification: Listening to the data

The data is sonified so that the spatial trajectory can be heard. Sonification is a process where data are mapped to sound, and for our data, parameter mapping sonification is the appropriate method, described in Chapter 15 of The Sonification Handbook (Hermann, Hunt, & Neuhoff, 2011), using the first author's previously designed software called Cheddar (Barrett, 2016).

It is necessary to make appropriate decisions as to how the parameters of the data are mapped to sound parameters, as well as exploring scaling ranges that most clearly reveal the qualities in the data. Cheddar draws on the perceptual aspects of our spatial hearing, as well as allowing interesting sounding results.

To isolate the effect of spatial movement it is necessary not only to remove visual information, but to also avoid any metaphors, or connotations of real-world events that may be implied by the sound prior to spatialisation. The sonification is therefore designed to create sounds that are unrecognisable as to their acoustic source, so as to avoid listeners attaching human causation by way of identifying non-spatial information. For example, when hearing a recognisable musical instrument, most listeners intuitively connect this with a human performance. These measures are also valid for non-instrumental acoustic sources or for sounds created by synthesis methods. For example, if we believe we hear the sound of a door being closed, we may project a spatial action onto this event even if the sound is spatially static. This causation-action connection is regularly exploited by composers of acousmatic music. More recently, Caramiaux, Bevilacqua, Bianco, Schnell, Houix, and Susini (2014) indeed showed that for sounds where a causal action could be clearly identified, listeners mimicked the action with their bodies even though the sounds were only heard in stereo. Caramiaux's sound sources and sound causations were clearly recognisable, such as champagne glasses clinks and the closing of a cupboard door. Our sonified sounds should be designed such that any connection to an absolute source (e.g. glasses or cupboards) is vague, while maintaining the sense of human causation that we are interested in capturing.

The motion source is contained within a relatively small spatial volume. There are a number of reasons why it is both advantageous and necessary to scale up this spatial volume in the sonification. Johnson (2008, Chapter 7) discusses that it is the shape, relations and patterns that are important. Elaborating further, in listeners embodied experiences of sound and music there is rarely a direct one-to-one mapping of mode or scale. When mentally re-coding sound into multi-modal gestural images, the mental re-enactment is on a scale of personal awareness, or the near-body peripersonal space. This space may

appear larger to the actor than the real-scale seen by a visual observer. Repeating Godøy's (2010b) clarification of embodied cognition, he proposes that it involves 'our capacity for having internal images of the world, as somehow originating in, but not necessarily truthfully reflecting external experience'. A number of studies reinforce this view: for example a multi-modal transfer is evident in the sound-tracing studies, mentioned in Section 2, where listeners are asked to draw or make spontaneous free movements to a set of musical excerpts, where the results showed a reasonable degree of consensus in the overall character of the gestures despite both modal and scaling variations (Godøy, 2010a). Kirsh (2013) also discusses the theory of embodied cognition and how peripersonal space can be extended to larger scales. Rather than attempting to evoke in the listener the experience of watching an action from afar, we are interested in their experiential, embodied understanding which may evoke first hand physical experiences. These experiences involve movement sensations that may be clearly evident to the actor, but invisible to the viewer. Scaling up our spatial projection and placing the listener central to this motion is therefore necessary so we may explore a parallel to this phenomenon.

In our work, the change in scale is therefore a logical conclusion: close-up physical experiences will occupy an alternative scale to that of the original source seen or heard from afar, and what is most important is to ensure that the spatial information is clearly experienced by the listeners' auditory perception. Scaling will mean that smaller spatial motions in the source are more likely to be audible. Finally, we are interested in acousmatic music, where in a concert, spatialisation occurs throughout the audience space. Scaling-up the spatial motion is in keeping with the way composers perform spatial ideas in their music, and maintains an alignment between our experimental and general artistic practice.

A number of studies (e.g. Bigand & Parncutt, 1999; Krumhansl, 1996) show that many structural features of music contribute to the experience of tension, such as loudness dynamics, note density and harmonic relations. The mapping of motion velocity to grain volume and grain duration results in changes of timbre and loudness, leading to changes in perceived tension which may enhance associations with human agency. Changes in timbre, loudness and centroid (or pitch focus) are also allied with our perception of space beyond that of spatial movement per se. Although sometimes regarded as a cliché, pitch changes are easily interpreted as spatial height changes. Further, vertical movement has been shown to be related to sound frequency, both in terms of spectral centroid and pitch (Nymoen, Torresen, Godøy, & Jensenius, 2012). Changes in frequency correlated with changes in spatial height are also relevant to the way in which the shape of our ears and head filters sound from different directions.

With these considerations, the following mapping of data to sound was used:

- Each data point triggered a sound grain. The sound grains were initially identical, and chosen from a spectrally rich source.
- 3-D spatial-data points determined 3-D grain location in a 6th order 3-D ambisonics synthesis. The dimensions of the motion source, which traversed a volume of 0.5 m × 0.4 m × 0.3 m, were scaled up to occupy the width of the listening space, resulting in a sonification occupying a volume of 5 × 4 × 3 metres projected over the loudspeaker array of size 8 × 5 × 3 metres (see Figure 1(b) and (c)).
- Motion velocity was also mapped to grain duration of a range between 15–80 ms, where higher velocity data values resulted in longer grains. As the data rate was constant, higher velocities would therefore result in a denser grain overlap and subsequent timbral changes, as well as a volume increase. The grains are themselves fixed points in space, so we may hear a slight increase in trajectory width for fast or high variation trajectories which may be heard as a change in image size. Our short grain duration minimises this side effect. Moreover, in the dataset, accelerating sounds are approaching the listener, whereby any change in image size resulting from the grain-size and velocity correlation will serve to enhance any dynamic changes in proximity. (If the accelerating sounds were departing from the listener, then we may experience a conflict in information).
- Velocity of motion was mapped to the volume of each grain using the decibel scale, so that a doubling in data values resulted in a doubling of perceptual volume. This added an extra volume variation to the result of increased grain overlap, and will hereby be termed volume-2.
- Vertical movement was mapped to pitch.
- A fixed attack and decay envelop of 5 ms was added to each grain. Maintaining this short attack and decay envelop regardless of grain size ensured a textural quality conducive to spatial localisation.
- The timeline of the data was mapped to the timeline of the sonification.

To achieve useful statistical results, it was necessary to find a balance between the number of test stimuli, the number of pairs of stimuli that each listener needed to evaluate, and the number of participants that could be obtained locally. The test stimuli also needed to reflect

combinations of movement feature suppression reflected in space, timing, amplitude, pitch and timbre. We chose to make 12 test stimuli, paired into a possible 66 combinations.

In the first sonification, the mapping ranges were scaled to most clearly enhance features in the data. In a pilot study where listeners were provided with information about the motion source, larger scaling ranges were in general experienced as clarifying micro variations. Eleven more sonifications were made from modifications of this dataset, where each version either suppressed features of the original movement captured in the data, or reduced the scaling ranges of the sonification or both. Although the 11 stimuli have a label (from 2 to 12), they are not arranged in any specific order, meaning that stimuli 12 could be rated higher than stimuli 3.

In ambisonics, it is necessary to specify the view point from which the spatial synthesis is calculated. We can think of this as the location of a 'virtual' listener inside the data, which will then also be the perspective of a real listener. For all but two test stimuli, the sonifications were made for the real listener located in the center of the motion mean, aligning with our discussion of embodiment and spatial scales discussed above. The spatial source followed a path that was laterally biased to the human body that created it – following a skewed diagonal including height variation. It is logical to maintain the same orientation in the sonification, not only to be true to the original human action, but also because our spatial-auditory perception is more sensitive to left–right than front-back movements. It was also important that although listeners were orientated in this way, they were free to move their heads and torso, and gain more information as to the relationship between the moving sound and their bodies.

The remaining two test stimuli served as controls: one was reduced to mono so as to test for the strength of spatially congruent timbral and dynamic processing alone, while the other was spatialised with the virtual listener on the edge of the spatial domain, as if the motion occurred in front. In all but two cases, time was treated in original tempo (one data point triggering a sound grain every 4 ms), where the duration of each sonification was 5 s.

Summary of the test sounds:

- S1 Original data sonified to optimise features of the spatial projection. Pitch, volume-2, grain duration and spatial variations are set to their most dynamic ranges (before sounding comical).
- S2 The same as S1, but where spatial motion is placed in front of the listening location.
- S3 Pitch, volume-2 and grain duration are set to their most dynamic ranges, but all spatial motion is removed and the sound rendered as mono. This test sound is played over one loudspeaker located directly in front of the listener.
- S4 3-D spatial variation in the original data is partially reduced, leaving global direction changes. The sonification mapping is the same as for S1, but as there is less dynamic variation in the source data, the resulting sonification is also reduced in pitch, volume-2 and grain duration variations.
- S5 3-D spatial variation in the original data is flattened further, consisting of just three changes in direction between two points in space. Although the sonification mapping is the same as for S1, the lack of spatial dynamic variation results in very little variation in pitch, volume-2 and grain duration.
- S6 The same as S1, but with volume-2 variation in the sonification removed.
- S7 The same as S1, but with pitch variation in the sonification removed.
- S8 The same as S1, but with pitch and volume-2 variation in the sonification removed.
- S9 The same as S4, but with pitch and volume-2 variation in the sonification removed.
- S10 The same as S5, but with pitch and volume-2 variation in the sonification removed.
- S11 The same as S1 played 30% slower in tempo.
- S12 The same as S1 played 50% slower in tempo (half speed).

Ideally, the list of test stimuli would have consisted of the three spatial variations, each sonified with all permutations of pitch and volume-2 variation. This would however have resulted in four more stimuli (for a total of 16), with a knock-on effect of increasing the number of pairs to 120. It is unrealistic for any one listener to evaluate 120 test pairs, and here we also experience a trade-off: the less pairs each listener evaluates, the more listeners required to achieve meaningful statistical results, where the increase is significant and non-linear. Moreover, in such a case the difficulty of the test would have increased, since the differences between the stimuli would have been smaller and more difficult to hear.

With the hypothesis that larger sonification ranges enhance spatial information in the data, S1 should be ranked top, while S10 should be ranked at the bottom. We can also speculate that as the added pitch variation may enhance vertical motion, yet is not a true part of the 3-D sonification, S7 may be evaluated similarly to S1.

3.4. Test procedure

A pilot listening session was carried out on two listeners who were not participating in the final experiment: one experienced in electroacoustic music and one inexperienced. Both listeners were aware of the aims of the project, were asked to assess whether the sonifications made audible the intended information, whether the proposed questions were clear, and whether the total duration of the test and number of test stimuli was realistic. Based on listener feedback a number of changes were made:

- Evaluating all possible 66 pairs of sounds exceeded the listeners' attention span. The total duration of the test was therefore reduced to 30 pairs of sounds (which is 45% of the total number of possible pairs out of 12 stimuli).
- Sonification mapping ranges were adjusted for pitch and volume-2 scaling, so as to be sure that the qualities of test stimuli S1 were clearly audible.
- The original text asked the listener to identify 'human agency'. This term was discussed as too specialised, and instead replaced with the phrase 'human physical action'.

Forty-six listeners spanning a broad range of ages (21–65 years) and musical abilities, from non-musicians to professional performers, completed the test. One participant was turned down after reporting known hearing loss. The tests were carried out in a darkened black box room. Each listener was located at the centre of the space, which is the most accurate 3-D spatial listening point.

Listeners were presented with the following statement: 'This test investigates the role of sound spatialisation in how we may associate sound that we hear, with human physical action.' They were then told that they would be presented with 30 pairs of short sounds, and for each pair, they should choose the one that, 'most evokes a feeling of human causation or human physical origins'. They were asked to judge the sound as they experienced it in relation to their own body, this being to avoid the listener trying to ascertain a possible real source, which in the pilot was shown to be a rationalisation that halted the intuitive process. The instructions were read in English, where a few listeners asked for Norwegian translations. Listeners were also informed that the sounds were made by sonification (with a brief explanation), that this process results in the sounds appearing abstract and could be experienced as somewhat strange. The test began with a training session, during which the listeners were asked to familiarise themselves with these strange qualities. See Appendix 1 for a copy of the original and revised spoken scripts.

Table 1. The estimated consensus ordering of the three clusters.

Cluster 1	Cluster 2	Cluster 3
$\alpha_1 = 2.65$ (1.15,4.91)	$\alpha_2 = 5.14$ (3.16,9.18)	$\alpha_3 = 5.29$ (3.61,7.56)
$\eta_1 = 0.31$ (0.21,0.41)	$\eta_2 = 0.33$ (0.22,0.43)	$\eta_3 = 0.37$ (0.27,0.48)
S8	S5	S1
S10	S4	S7
S5	S12	S11
S9	S2	S2
S6	S11	S4
S4	S3	S12
S7	S6	S6
S11	S1	S3
S12	S7	S5
S2	S9	S9
S3	S8	S8
S1	S10	S10

Listeners noted their answers on a chart, selecting the first or the second from each pair of unlabelled sounds, and were requested to always make a choice even if they found it difficult to decide. They were also allowed to repeat a test pair, but only in sequence and not at the end of the experiment. At the end, they were asked to complete two questionnaires that probed their background musical and spatial-audio experience. One questionnaire resulted in a musical sophistication index score (MSI) and the other rated spatial-audio awareness (SAA). The MSI used was the Ollen Musical Sophistication Index, which is an online survey that tests the validity of 29 indicators of musical sophistication used in published music research literature (Ollen, 2006). The SAA, or spatial audio awareness index consisted of five questions as indicators of how aware listeners were of spatial audio regardless of musical background. Such a test did not already exist in the literature and was custom designed for the experiment.

The spoken introduction and training session lasted for four minutes, the test lasted 16 minutes, and the questionnaires were completed in an average of five minutes.

4. Results

4.1. Results from the listening test

We analysed the data using the Bayesian Mallows model for non-transitive pair-wise comparisons data (Vitelli et al., in press; Crispino et al., 2017). The algorithm detected three clusters, corresponding to three categories of listeners, each sharing a similar perception of the sounds. In Table 1 are listed the estimated consensus orderings of the three clusters.

At the top of the table are also reported the maximum a posteriori estimates for the proportion of listeners in each cluster, η, and the dispersion parameter of each cluster, α, together with their 95% high posterior density intervals,

that is the Bayesian version of confidence intervals. The parameter α measures the agreement of the listeners in each cluster; the higher α is, the more the listeners agree with the consensus ordering of the cluster.

In Figure 2 we represent the consensus orderings of Table 1 through heat-plots. The ordering of the stimuli is shown on the x-axis and their rank on the y-axis. Each cell represents, through greyscale, the probability that the corresponding sound (on the x-axis) has the rank reported on the y-axis. A dark cell, represents high probability, a light one low probability, and the different shades of grey all intermediate probabilities (as in the legend on the side of each plot).

The explanation of the clusters assignments is as follows:

Cluster 1 (C1)

Top three stimuli: S8, S10 and S5.

Bottom three stimuli: S2, S3 and S1.

The results in C1 are more uncertain than in the other two clusters. This is evinced by both the small value of the parameter α (that indicates a low agreement of the listeners), and the small values (never larger than 0.35) of the probabilities for each cell of Figure 2 (left). Despite this, a trend can be evinced. Rankings 1 and 12 (corresponding to S8 and S1), that have the highest probabilities (as indicated by the darker colour of the corresponding cells), are the most interesting to evaluate. Although the probability of S8 and S1 being placed exactly in those positions is quite low, the uncertainty of the values of ranks very close to them renders S8 with the probability 0.32 of being ranked 1st, 0.21 of being ranked 2nd or 3rd, and 0.19 of being ranked 4th. The probabilities of the other ranks (5th to 12th) are very small, and in relation to these, the approximate position of S8 is relatively strong. A similar reasoning holds for S1.

S8 contains all spatial movement details projected in space and enhanced by grain density, but with pitch and extra volume-2 variations removed. S1 is the same as S8 but with pitch and extra volume-2 variations present. The ranks of these two stimuli then suggest that cluster 1 primarily listens for spatial movement but finds pitch and volume-2 as artificial additions or distractors. However, in general, the low values of the probabilities suggest that these listeners are not convinced that they hear human agency in any of the test stimuli.

Cluster 2 (C2)

Top three stimuli: S5, S4, and S12.

Bottom three stimuli: S9, S8, S10.

S5 contains the least movement variation where pitch and volume-2 are naturally suppressed. S4 and S12, although both similar to S1 (ranked lower), are each less dynamic in their own way: S12 is played half speed and S4 reflects the global but not smaller movement details. These top three stimuli, where S5 and S4 have high probabilities of being ranked 1st and 2nd, suggest that C2 finds faster movements as negative to their evaluations, while pitch and volume-2 variations are supportive. The bottom three stimuli sustain this assumption: stimuli S9 and S10 are the same as stimuli S5 and S4, but lack pitch and volume-2 variations, and S8 also lacks pitch and volume-2 variation. The uncertain placement of S3 which is ranked 6th, but with high variability around it (the probabilities of being ranked 4th to 7th are very similar) suggests that total removal of spatial information, yet the presence of correct pitch and volume variations, removes the ability of these listeners to make any positive or negative judgements, implying spatial information helps these listeners make a decision.

Cluster 3 (C3)

Top three stimuli: S1, S7 and S11.

Bottom three stimuli: S9, S8 and S10.

S1 is the optimised full spatial representation of the source data. Its top rank is also robust in this cluster (its probability is around 0.75). S7 is the same as S1 but with pitch variation removed, while S11 is the same as S1 played 30% slower. S9 and S8 contain significant spatial variation but lack both pitch and volume-2 variations, while S10 contains the least of all information and its bottom ranking is robust (its probability is around 0.7). We can deduce that C3 rates spatial movement highly, but also requires volume-2 variations for robust results. S3 is ranked 8th, but unlike for C2, its rank is more likely (around 0.6). This tends to indicate that pitch and volume-2 variations with real spatial movement may be more important than real spatial movements without pitch and volume-2 variations (i.e. that S3 is with high probability ranked above S8). In summary C3 appears to prefer all spatial cues that adhere to our everyday perception of spatial motion when assessing the stimuli. C3 is also the less uncertain cluster, and contains the largest group of listeners, as indicated by the η_3 parameter of Table 1.

The stability of the clustering is shown in Figure 3. The heat-plot represents the probabilities, for all the listeners (on the x-axis), for being assigned to each of the 3 clusters (y-axis) identified in Table 1. Most of the probabilities gather around one specific cluster amongst the three possibilities, indicating a reasonably stable behaviour in the cluster assignments. These results clearly identify three groups of listeners, each with different listening strategies.

The experiment was difficult as expected: 80% of the listeners reported non-transitivities in their pair comparisons and only 9 out of 46 listeners were able to stay consistent with themselves. The remaining 37 listeners produced many non-transitivities. In Figure 4 is represented the matrix of aggregated cycle co-memberships:

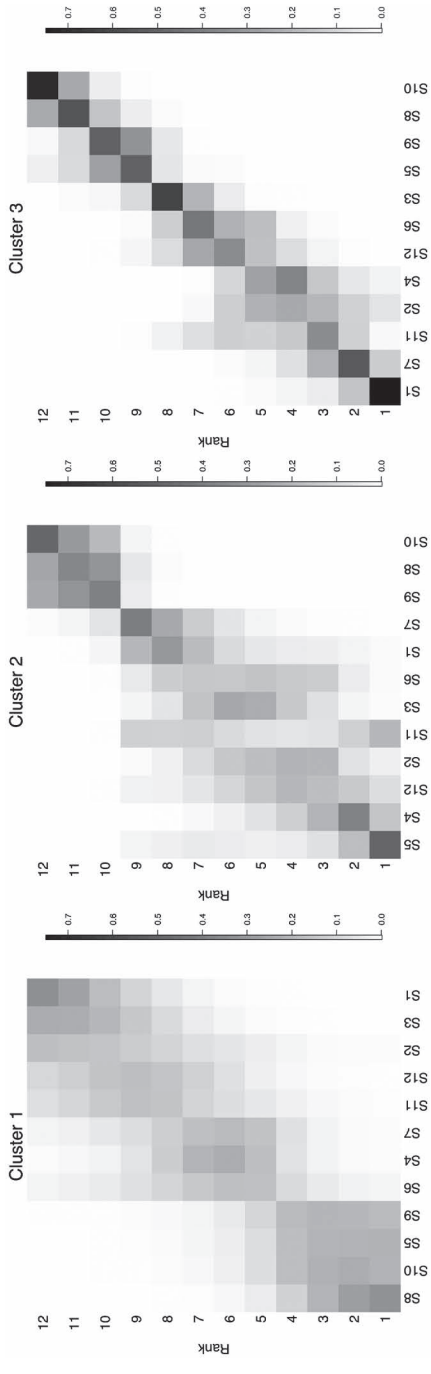

Figure 2. The heat-plot representing the estimated consensus orderings, with uncertainty. On the right of each plot, is reported the grey scale, representing the probability of each cell.

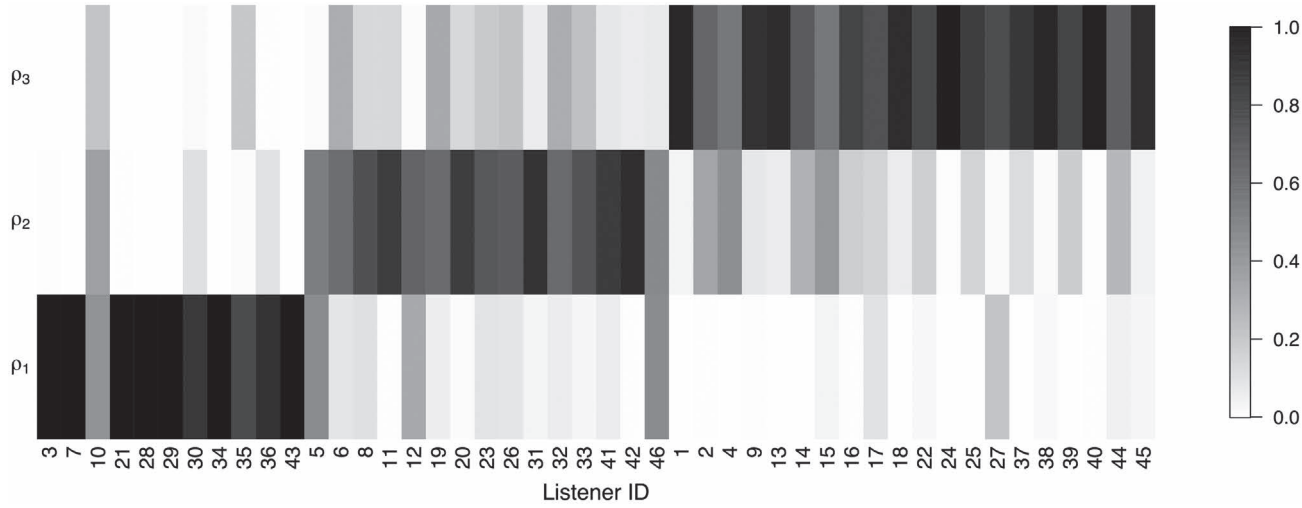

Figure 3. For all the listeners (*x*-axis), is reported the probability of being assigned to each of the three clusters (*y*-axis).

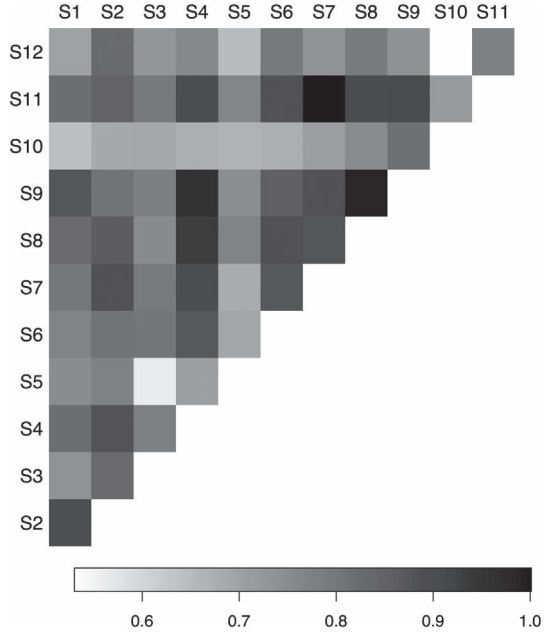

Figure 4. Heat-plot representing co-memberships of sounds in the non-transitive patterns of the data: the horizontal grey color scale under the map indicates the relationship between colors and probabilities.

each cell represents the probability that the corresponding sounds on the x-axis and on the y-axis are in a same non-transitive pattern (here called cycle), and thus the probability that they are confused by the listeners. As before, a dark colour indicates a high value, a light colour stands for a small value of the probability. This plot helps in understanding the extent to which a sound is more easily confused with another. For example S10 and S12 appear in the same cycle a small number of times (white cell), while S8 and S9 appear in the same cycle a large number of times (black cell).

4.2. Results from the questionnaires

It is interesting to investigate whether there is any relationship between listeners' musical background and the test results. Figure 5 shows the probabilities of S1 (left), S7 (middle), and S1 and S7 jointly (right) being ranked amongst the top-4 in the individual ranking of the listeners, stratified by the SAA index (that could take integer values from 0 to 3, where large values indicate higher levels of spatial audio awareness). The horizontal dashed line is the threshold in case of random assignment. There is a clear correlation between listeners' spatial audio awareness and their high ranking of test stimuli that capture the human qualities of the hypothesis (that is, S1 and S7). The same plot, when stratified by the MSI index did not exhibit a clear correlation, for any test stimuli (results not shown).

We then display, in Figure 6, the probabilities of S8 (left), S10 (middle), and S8 and S10 jointly (right), being ranked amongst the bottom-4 stimuli, stratified by the MSI index. The index is here codified either greater than 500 (that corresponds to more sophisticated listeners), or <500 (that is less sophisticated listeners).[4] We notice here a lower correlation between each listener's MSI and their evaluation of the test stimuli. However, S8 and S10, which are the test stimuli that don't capture the human qualities of the hypotheses, are more probable to be ranked in the last 4 positions by listeners with high MSI.

[4]This coding is suggested in http://marcs-survey.uws.edu.au/OMSI/omsi.php.

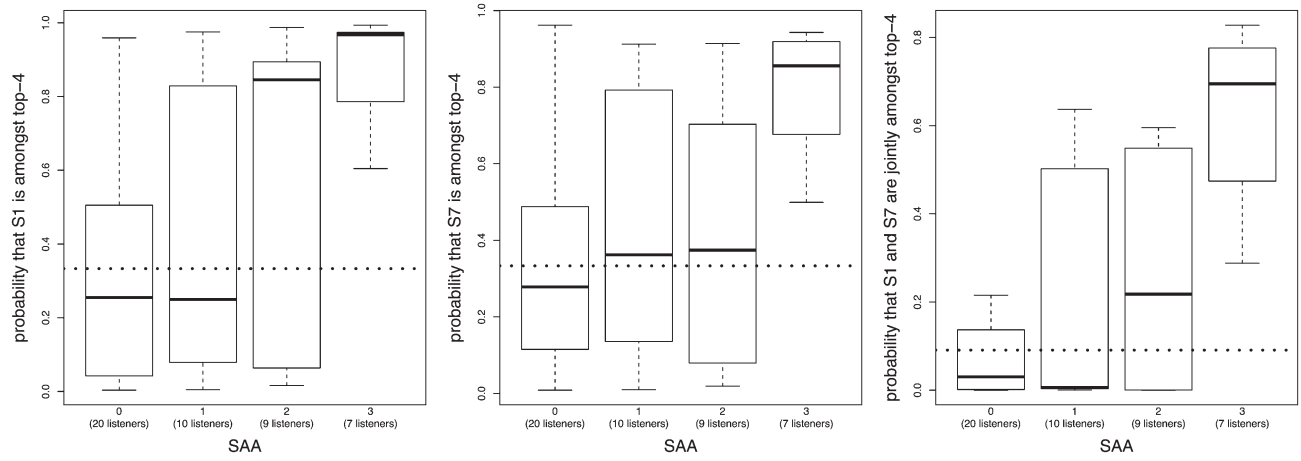

Figure 5. The probabilities of S1 (left), S7 (middle), and S1 and S7 jointly (right) being ranked among the top-4 stimuli, in the individual ranking of the listeners, stratified by the SAA index.

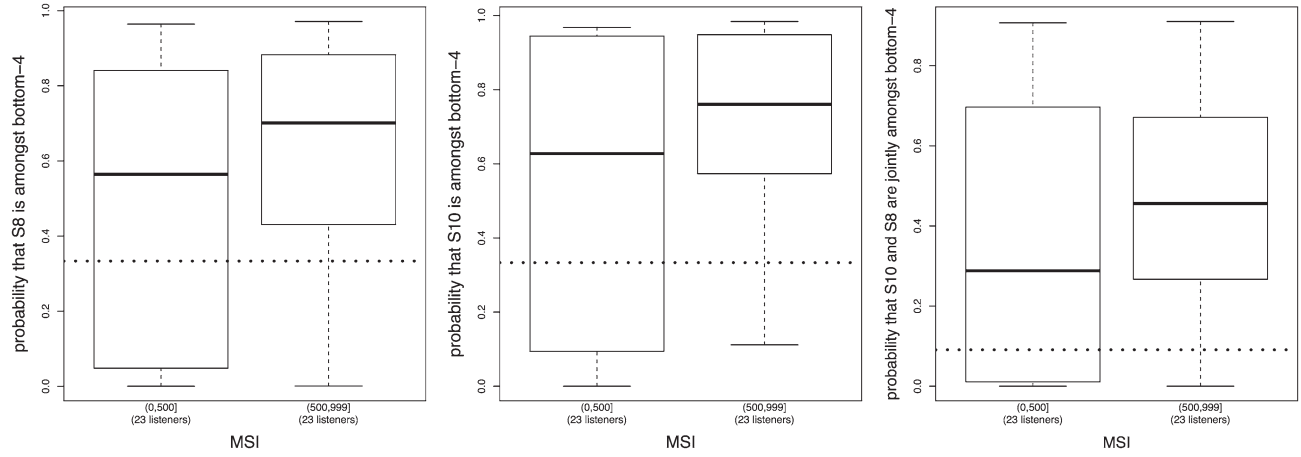

Figure 6. The probabilities of S8 (left), S10 (middle), and S8 and S10 jointly (right), being ranked among the bottom-4 stimuli, stratified by the MSI index.

5. Conclusions

We investigated how a sound's spatial behaviour influences our listening understanding, and specifically, how a sense of human agency may be evoked by the way in which sound moves through space. Sonification was used to create 12 test sounds in 3-D ambisonics. Spatial data capturing human motion was mapped to sound, reflecting changes in 3-D spatial trajectory and motion speed, where velocity of motion was further mapped to amplitude and timbral intensity changes, and vertical movement was mapped to pitch change. In addition, spatial cues correlating to source distance (changes in amplitude and low-pass filter cut off, and changes in image size) were processed when rendering the 3-D ambisonics. Amplitude variations correlating with changes in velocity were named 'volume-2', so as to differentiate from distance related amplitude variations. For most sounds, the listening location was located in the centre of the spatial motion.

Test sound labelled S1 was designed to most clearly project the source data in spatial sound by combing the sonification parameters with suitable range scaling. S1 was assessed based on the author's artistic practice and in a pilot listening study. The remaining 11 sounds reduced features in the data by suppressing spatial motion variation, by slowing down the speed of motion, or by suppressing pitch or volume-2 variations. Timbral variations were suppressed in tandem with slower movements or less dynamic velocity changes. In the sound labelled as S10, we removed all spatial features found in the human source gesture, and assumed S10 to be least successful in evoking human agency. We did not make any further assumptions as to the expected order of preference for the remaining 10 sounds. S2 and S3 were special cases: S2

was sonified similarly to S1, but with the listening location moved backwards, such that the spatial information was active only in front of the listener, as if occurring on an invisible stage; S3 played in mono directly from one loudspeaker without any ambisonics 3-D sound recreation.

Our results showed three clusters of listeners. Cluster 1 contains 11 listeners, cluster 2, 15 listeners, and cluster 3, 20 listeners, and we can conclude that each cluster contains listeners with different preferences. The spread of uncertainty for each ranking indicates that answering the question as to whether sound spatialisation can suggest human agency is far from straight forward. Our assumptions that S1 optimally captured human agency, and that S10 was least successful, is shown strongly for cluster 3. Cluster 2 also ranks S10 last, yet displays a different ranking for the top sounds, while cluster 1 shows a reversal of our assumptions, placing S1 at the bottom. However, common to cluster 1 and 3 – which together account for 31 out of the 46 listeners - is a preference for the original, complete spatial movement. We can see this in the following result: the top position in both clusters is occupied by a sound which preserved spatial movement - S1 and S8. The difference between these sounds is that pitch and volume-2 variation have been completely suppressed in S8. Specifically, cluster 3 ranks our own best sound S1 as highest, while cluster 1 ranks sound S8 highest. Conversely, cluster 1 ranks S1 as lowest, while cluster 3 ranks S8 almost at the bottom. In other words, the two clusters are in disagreement as to the role of pitch and volume-2 variations. The only common feature of S1 and S8 is the movement of the sound in space and the emulation of acoustic features designed to aid in our perception of distance. Therefore, if we are considering all listeners in clusters 1 and 3, spatial motion in itself is more salient than when the motion is enhanced by volume or pitch variations, which appear to create greater disagreement amongst listeners. We should however not forget that cluster 3 is the largest group, with almost double the number of listeners than cluster 1. Moreover, the uncertainty of the results regarding cluster 1 is much higher than those of cluster 3, which appears to group listeners that strongly agree on the final list (as indicated by the grey scale of Figure 2).

If we focus on cluster 3, we see that sound S7 is ranked in second place. S7 is the same as S1 but with pitch variation suppressed. Cluster 3 ranks sound S8 second from bottom. S8 is the same as S1 but with both pitch and volume-2 variation suppressed. Within cluster 3 we can therefore conclude that changes in velocity, effected in the sonification by volume-2 variation, are an important part of the listeners' evaluation and more salient than pitch variation.

The results for listeners in cluster 2 display a number of contradictions that imply that these listeners either find the test is too difficult, are unable to listen spatially, or cannot connect human agency to the spatial information presented in the tests. The strongest contradiction is that they rate S5 at the top and S10 at the bottom. In S5, spatial variations are flattened out. The resulting sound is less detailed in its spatial movement, and the knock-on effect is that all other sounding features, including pitch, timbre and volume variations are similarly reduced. S10 is very similar to S5. The spatial movement is identical, and although pitch and volume-2 variations are removed from in the sonification, the sounding presence of them in S5 was already very slight compared to the other stimuli.

Regarding sound S2, where the spatial scene is rendered in the front of the listening space, results are somehow uncertain. In particular, in clusters 2 and 3, S2 is ranked in the top-middle positions, but the uncertainty attached to this exact rank is high. We can however say that in cluster 3, with very low probability (almost zero), S2 is ranked in the last 5 positions, and in cluster 2 with very low probability it is ranked in the bottom 4 positions. In cluster 1, instead the result is the opposite: S2 is ranked in the bottom-middle positions, and with very low probability (almost zero) to be in the top 5 positions. These results indicate that the location of spatial activity in relation to the listener appears to affect decision making. For S2, the listener is no longer central to the motion activity, and referring back to our discussion on spatial scaling in Section 3.3.4, listeners' uncertainty could be a logical result of the spatial information appearing less present in relation to the physical body. Sounds spatialised in this way also appear perceptually to travel less distance in relation to the listening position, and this too may be important.

It was shown in the results from the spatial audio awareness questionnaire that there was a clear correlation between listeners' spatial audio awareness and their high ranking of test sounds that capture the human qualities of the hypothesis. 6 out of the 7 listeners who scored the highest mark in the questionnaire are also grouped in cluster 3, and 5 out of 9 listeners who scored the second highest mark are also in this cluster. This indicates that spatial listening which draws on all features of S1 - that is not only spatial location, but where the dynamics are enhanced by volume and pitch modulations, - is a skill that is enhanced through experience and personal interest. It is not possible to be certain as to whether this is because these listeners have been trained to hear spatially, or whether their personal interests, or occupation, have conditioned them to listen more easily to the type of abstract sounds used in the test. However, if we compare to cluster 2, no listeners scored a 3 in the questionnaire,

and rather 9 listeners scored 0, which may explain the contradictions observed in this cluster's choices.

6. Future work

The preference of cluster 2 for slower movements may suggest that listeners find the large and fast variations in pitch and volume-2 distracting, which can indicate that a future study would benefit from smaller variations in the sonification scaling range. Also, the results from clusters 1 and 3 suggest that pitch variation as an enhancement of verticality is an unnecessary addition or may have involved an incorrect sonification scaling, which may have served to make the tests trickier or the results less consistent. Nymoen et al. (2012) showed that a significantly higher amount of overall acceleration was observed for non-pitched sounds as compared to pitched sounds, where in our study, pitch modulation may have incorrectly influenced the perception of acceleration. A further study may therefore choose to remove this aspect of the sonification.

References

Barrett, N. (2016). Interactive spatial sonification of multidimensional data for composition and auditory display. *Computer Music Journal, 40*(2), 47–69.

Berg, J. (2009). The contrasting and conflicting definitions of envelopment. *Audio engineering society convention* (Vol. 126). Munich: Audio Engineering Society.

Berg, J., & Rumsey, F. (2003). Systematic evaluation of perceived spatial quality. *Audio engineering society conference: 24th international conference: Multichannel audio, The new reality*. Banff: Audio Engineering Society.

Bigand, E., & Parncutt, R. (1999). Perceiving musical tension in long chord sequences. *Psychological Research, 62*(4), 237–254.

Blauert, J. (1997). *Spatial hearing: the psychophysics of human sound localization*. Cambridge, MA: MIT Press.

Caramiaux, B., Bevilacqua, F., Bianco, T., Schnell, N., Houix, O., & Susini, P. (2014). The role of sound source perception in gestural sound description. *ACM Transactions on Applied Perception (TAP), 11*(1), Article ID: 1 (23 pages).

Carpentier, T., Barrett, N., Gottfried, R., & Noisternig, M. (2017). Holophonic sound in IRCAM's concert hall: Technological and aesthetic practices. *Computer Music Journal, 40*(4), 14–34.

Crispino, M., Vitelli, V., Barrett, N., Arjas, E., & Frigessi, A. (2017). A Bayesian Mallows approach to non-transitive pair comparison data: how human are sounds? ArXiv e-prints 1705.08805.

Daniel, J., & Moreau, S. (2004). Further study of sound field coding with higher order ambisonics. *Audio Engineering Society Convention* (Vol. 116). Berlin: Audio Engineering Society.

David, H. A. (1963). *The method of paired comparisons*, Vol. 12. London: DTIC Document.

Davies, E. (2007). *Beyond dance: Laban's legacy of movement analysis*. New York, NY: Routledge.

Eitan, Z., & Granot, R. Y. (2006). How music moves. *Music Perception: An Interdisciplinary Journal, 23*(3), 221–248.

Favrot, S., & Buchholz, J. (2012). Reproduction of nearby sound sources using higher-order ambisonics with practical loudspeaker arrays. *Acta Acustica United with Acustica, 98*(1), 48–60.

Frank, M. (2014). How to make ambisonics sound good. *Forum acusticum, (Krakow)*. European Acoustics Association.

Godøy, R. I. (2006). Gestural-sonorous objects: Embodied extensions of Schaeffer's conceptual apparatus. *Organised Sound, 11*(02), 149–157.

Godøy, R. I. (2010a). Gestural affordances of musical sound. *Musical gestures: Sound, movement, and meaning* (pp. 103–125). New York, NY: Routledge.

Godøy, R. I. (2010b). Images of sonic objects. *Organised Sound, 15*(1), 54–62.

Godøy, R. I., Haga, E., & Jensenius, A. R. (2006). Exploring music-related gestures by sound-tracing: A preliminary study. In *Proceedings of the COST287-ConGAS 2nd International Symposium on Gesture Interfaces for Multimedia Systems (GIMS2006)* (pp. 27–33).

Harrison, J. (1999). Diffusion: Theories and practices, with particular reference to the beast system. *eContact, 2*, 9.

Hermann, T., Hunt, A., & Neuhoff, J. G. (2011). *The sonification handbook*. Berlin: Logos Verlag.

Hoff, P. D. (2009). *A first course in Bayesian statistical methods*. Springer Science & Business Media.

Huber, J., Payne, J. W., & Puto, C. (1982). Adding asymmetrically dominated alternatives: Violations of regularity and the similarity hypothesis. *Journal of Consumer Research, 9*(1), 90–98.

Johnson, M. (2008). *The meaning of the body: Aesthetics of human understanding*. Chicago, IL: University of Chicago Press.

Kirsh, D. (2013). Embodied cognition and the magical future of interaction design. *ACM Transactions on Computer-Human Interaction (TOCHI), 20*(1), Article ID: 3 (30 pages).

Krumhansl, C. L. (1996). A perceptual analysis of Mozart's Piano Sonata K. 282: Segmentation, tension, and musical ideas. *Music Perception: An Interdisciplinary Journal, 13*(3), 401–432.

Leman, M. (2012). *Musical gestures and embodied cognition* (pp. 5–7). Mons: Université de Mons.

Lossius, T., Baltazar, P., & de la Hogue, T. (2009). DBAP-distance-based amplitude panning. *Proceedings of 2009 international computer music conference* (pp. 489–492), Montreal.

Mallows, C. L. (1957). Non-null ranking models. I. *Biometrika, 44*(1/2), 114–130.

Marentakis, G., & McAdams, S. (2013). Perceptual impact of gesture control of spatialization. *ACM Transactions on Applied Perception (TAP), 10*(4), Article ID: 22 (21 pages).

Marentakis, G., Zotter, F., & Frank, M. (2014). Vector-base and ambisonic amplitude panning: a comparison using pop, classical, and contemporary spatial music. *Acta Acustica united with Acustica, 100*(5), 945–955.

Nymoen, K., Torresen, J., Godøy, R. I., & Jensenius, A. R. (2012). A statistical approach to analyzing sound tracings. In S. Ystad, M. Aramaki, R. Kronland-Martinet, K. Jensen, & S. Mohanty (eds.), *Speech, Sound and Music*

processing: Embracing Research in India. (Vol. 7172, pp. 120–145). Berlin, Heidelberg: Springer.
Ollen, J. E. (2006). *A criterion-related validity test of selected indicators of musical sophistication using expert ratings* (PhD thesis). The Ohio State University.
Pedersen, M., & Alsop, R. (2012). An approach to feature extraction of human movement qualities and its application to sound synthesis. *Interactive: Proceedings of the 2012 Australiasian computer music conference*, Queensland.
Pulkki, V. (1997). Virtual sound source positioning using vector base amplitude panning. *Journal of the audio engineering society, 45*(6), 456–466.
Shestopalova, L., Bohm, T. M., Bendixen, A., Andreou, A. G., Georgiou, J., Garreau, G., ... Winkler, I. (2014). Do audio-visual motion cues promote segregation of auditory streams? *Frontiers in Neuroscience, 8*, Article ID: 64 (11 pages).
Simonson, I., & Tversky, A. (1992). Choice in context: Tradeoff contrast and extremeness aversion. *Journal of Marketing Research, 29*(3), 281–295.
Tsay, C.-J. (2013). Sight over sound in the judgment of music performance. *Proceedings of the National Academy of Sciences, 110*(36), 14580–14585.
Vines, B. W., Krumhansl, C. L., Wanderley, M. M., & Levitin, D. J. (2006). Cross-modal interactions in the perception of musical performance. *Cognition, 101*(1), 80–113.
Vitelli, V., Sørensen, Ø., Crispino, M., Frigessi, A., & Arjas, E. (in press). Probabilistic preference learning with the Mallows rank model. *Journal of Machine Learning Research*. Available at ArXiv e-prints, 1405.7945v4.
Zotter, F., Pomberger, H., & Noisternig, M. (2012). Energy-preserving ambisonic decoding. *Acta Acustica united with Acustica, 98*(1), 37–47.

Appendix 1. The instructions read to the participants

(1) First version of spoken instructions read to participants, tested in the pilot project:

This test investigates the roll of sound spatialisation in how we may understand human agency when we hear sound but cannot see how it was produced.

- You will be presented with 30 pairs of short sounds.
- Please select which, from each pair, most evokes a feeling of human causation or human physicality.

The sounds are may appear strange to you. The test will begin with a training session. Please use this time to familiarise yourself with these strange qualities.

When the test begins:

- The test number is displayed on the screen. Please write your answer on the paper. Please always make a choice even if you find it difficult to decide.
- If you need to repeat a test, please stop the operator and request the repeat.
- The lights will be dimmed. Please focus on hearing rather than vision.

Thank you for your valuable participation!

(2) Second (final) version of spoken instructions used in the tests:

This test investigates the roll of sound spatialisation in how we may associate sound that we hear, with human physical action.

- You will be presented with 30 pairs of short sounds.
- Please select which, from each pair, most evokes a feeling of human causation or human physicality.
- Please judge the sound as you experience it in your own body. Think of this as a direct human connection rather than as a human operating a machine, such as a train or car.

The sounds are made by a process called 'sonification'. We take 'silent' 3-D motion data from different sources, and turn it into sound that we can hear. The sounds may therefore appear 'strange' to you! The test will begin with a training session. Please use this time to familiarise yourself with these strange qualities.

When the test begins:

- The test number is displayed on the screen. Please write your answer on the paper. Please always make a choice even if you find it difficult to decide.
- If you need to repeat a test, please stop the operator and request the repeat.
- The lights will be dimmed. Please focus on hearing rather than vision.

Thank you for your valuable participation!
The test is anonymous.