THE ROLE OF PERCEIVED SOURCE LOCATION IN AUDITORY STREAM SEGREGATION: SEPARATION AFFECTS SOUND ORGANIZATION, COMMON FATE DOES NOT

TAMÁS M. BŐHM[1,2,], LIDIA SHESTOPALOVA[3], ALEXANDRA BENDIXEN[4,5], ANDREAS G. ANDREOU[6,7], JULIUS GEORGIOU[7], GUILLAME GARREAU[7], PHILIPPE POULIQUEN[6,7], ANDREW CASSIDY[6], SUSAN L. DENHAM[8] and ISTVÁN WINKLER[1,9]

[1]Institute of Cognitive Neuroscience and Psychology, Research Centre for Natural Sciences, Hungarian Academy of Sciences, H-1394 Budapest, P.O. Box 398, Hungary

[2]Department of Telecommunications and Media Informatics, Budapest University of Technology and Economics, H-1117 Budapest, Magyar tudósok krt. 2., Hungary

[3]Pavlov Institute of Physiology, Russian Academy of Sciences, Makarova nab. 6, 100034 St. Petersburg, Russia

[4]Institute of Psychology, University of Leipzig, Neumarkt 9–19, D-04109 Leipzig, Germany

[5]Institute of Psychology, Carl von Ossietzky University of Oldenburg, Ammerländer Heerstr. 114–118, D-26129 Oldenburg, Germany

[6]Department of Electrical and Computer Engineering, Johns Hopkins University, 3400 North Charles Str., Baltimore MD 21218, USA

[7]Department of Electrical and Computer Engineering, University of Cyprus, P.O. Box 20537, 1678, Nicosia, Cyprus

[8]Cognition Institute and School of Psychology, University of Plymouth, Drake Circus, Plymouth PL4 8AA, UK

[9]Institute of Psychology, University of Szeged, H-6722 Szeged, Petőfi S. sgt. 30–34, Hungary

ABSTRACT

The human auditory system is capable of grouping sounds originating from different sound sources into coherent auditory streams, a process termed auditory stream segregation. Several cues can influence auditory stream segregation, but the full set of cues and the way in which they are integrated is still unknown. In the current study, we tested whether auditory motion can serve as a cue for segregating sequences of tones. Our hypothesis was that, following the principle of common fate, sounds emitted by sources moving together in space along similar trajectories will be more likely to be grouped into a single auditory stream, while sounds emitted by independently moving sources will more often be heard as two streams. Stimuli were derived from sound recordings in which the sound source motion was induced by walking humans. Although the results showed a clear effect of spatial separation, auditory motion had a negligible influence on stream segregation. Hence, auditory motion may not be used as a primitive cue in auditory stream segregation.

Keywords: auditory motion, auditory stream segregation, auditory scene analysis, bistable perception, auditory streaming paradigm, natural movement, localization, auditory perception

INTRODUCTION

In most everyday situations, we find ourselves surrounded by objects producing sounds that often overlap each other in time, spectral content, or both. Therefore, in order to perceive the acoustic environment in terms of objects, the auditory system needs to disentangle the mixture of signals reaching the ears. Although sound segregation is a computationally challenging task, from experience we know that the human auditory system can perform it efficiently. The general problem of deriving an organized representation of the environment from the auditory input has been termed "auditory scene analysis" by Bregman (1990). Within this framework, a wide range of cues have been investigated with respect to their influence on the process of organizing sounds into coherent sequences, termed "auditory streams". For example, timbre proved to be an efficient cue for sorting sounds by their source (e.g., Smith et al., 1982; for studies testing various other auditory features, see Vliegen & Oxenham, 1999; Grimault et al., 2002; Roberts et al., 2002).

The majority of these studies employed the "auditory streaming paradigm" (van Noorden, 1975) delivering to listeners a repeating triplet composed of two kinds of sounds which differ from each other in some acoustic feature(s). This stimulus configuration can be heard either as a single auditory stream (termed the 'integrated' percept) or as two concurrent streams ('segregated' percept). Perceptual similarity between the two types of sounds and the presentation rate strongly affect how short sequences of this type are perceived (Moore & Gockel, 2002). Under prolonged exposure, participants report spontaneous perceptual switches among the alternative sound organizations (Pressnitzer & Hupé, 2006; Denham & Winkler, 2006; Denham et al., 2009). These perceptual switches occur even when the stimulus configuration strongly promotes one alternative organization over another (Denham et al., 2013).

Both van Noorden (1975) and Judd (1977) reported that sequences of tones presented in an alternating fashion to the two ears were perceived as segregated into two streams (see Bregman, 1990). Further, Denham and colleagues (2009) as well as Szalárdy et al. (2013) showed that location difference, simulated by imposing an inter-aural intensity difference (IID) on the tones in the repeating triplets, promoted segregation of the tones.

In the current study, we assessed the influence of real-space separations and spatial motion on auditory stream segregation. Thus we tested whether the Gestalt principle of 'common fate' (Köhler, 1947) provides a cue acting separately from the 'similarity-principle' based cue (another Gestalt principle that is known to be utilized in auditory stream segregation, as shown in the previously referred studies employing van Noorden's 1975 paradigm). Specifically, we examined whether sound sources moving together are more likely to be perceived as a single auditory stream than sources that are also co-located but remain stationary; and, conversely, we tested whether sound sources that move on separate trajectories are more likely to be experienced as two streams than spatially separated stationary sources.

Studies showing that the auditory system detects violations of spatial motion-based regularities suggest that the auditory system tracks the trajectories of sound sources. These studies showed that sounds with unexpected apparent movements embedded in a sequence of stationary sounds elicit the mismatch negativity (MMN) event-related potential component (Altman et al., 2005; Altman et al., 2010). Detecting the violations of sequential regularities has been linked with the formation and maintenance of auditory objects (streams) (Winkler et al., 2009;

Winkler, 2007). Further support for the possibility that motion-based cues may be utilized in auditory streaming comes from a theory of motion detection (Perrott & Marlborough, 1989) that assumes the existence of specialized motion-sensitive mechanisms in the auditory system. If this was the case, then sound source velocities are directly perceived sound features (Middlebrooks & Green, 1991). This makes auditory motion a readily available cue to be utilized in auditory stream segregation. Although the competing view of auditory motion detection suggests that sound source velocities are derived indirectly from the change in location over time (Middlebrooks & Green, 1991; Grantham, 1995), this assumption could also allow access to motion-related auditory features.

The experiment reported here is novel in that the stimuli have been produced by natural movement of sound sources, human walking. Two human assistants carried speakers emitting the test tones. Although this paradigm allows less accurate control of the spatial locations and trajectories than motion induced by finely controlled motors or simulated motion (two techniques frequently used in psychophysical and physiological experiments on auditory motion), we expect the motion patterns to be a closer match to normal perceptual experience outside the lab. In the experiment, we varied the motion pattern (stationary vs. two different kinds of spatial motion) and the co-location (mean spatial separation, i.e., moving side by side or on separate trajectories) of the two sound sources. Listeners were instructed to continuously mark whether they heard a single or two separate sound streams. Based on previous studies (Denham et al., 2009; Szalárdy et al., 2013), we expect that larger mean spatial separation promotes the perception of two streams. Regarding the effects of the motion cues, we hypothesized that:

1. When the two sound sources follow identical or joint trajectories, the proportion of perceiving a single auditory stream will increase compared to that reported for two stationary sound sources of similar spatial separation.
2. When two sound sources follow separate trajectories, the proportion of perceiving two auditory streams will increase compared to that reported for two stationary sounds of comparable mean spatial separation.

MATERIALS AND METHODS

Participants

Nineteen university students (12 females; 18 right-handed; 18–23 years, mean 21.1 years) participated in the experiment receiving modest financial compensation. Participants were screened in advance for normal hearing: thresholds below 25 dB HL, measured at 250 Hz, 500 Hz, 1000 Hz, 2000 Hz and 4000 Hz, and a maximum threshold difference between the two ears of 5 dB at 250 Hz and 500 Hz and 10 dB at 1000 Hz and 2000 Hz were enforced. Participants gave written informed consent before the experiment, which was approved by the Ethical Committee of the Institute of Cognitive Neuroscience and Psychology, Research Centre for Natural Sciences, MTA.

Apparatus and stimuli

The stimuli were based on audio recordings collected as part of a multi-modal data collection

(Georgiou et al., 2011). Alternating sequences of two sinusoid tones of equal amplitude delivered with 25 ms inter-stimulus interval, one with a frequency of 400 Hz and the other one three semitones higher, at 475.7 Hz, were recorded. Tone duration was 100 ms, including onset and offset ramps of 10 ms long raised cosines.

The tone sequences were played by Anthony Gallo Acoustics A'Diva Ti compact speakers mounted on construction helmets worn and carried around by two assistants. The speaker drivers were facing upwards to provide equal sound emission in horizontal directions. The sound signals were generated on an IBM PC and transmitted to the speakers by FM radio units.

Table 1. Experimental conditions

Condition	Motion	Co-location	Stimulus type	Spatial cues
1	Stationary	Identical	Recorded	Binaural
2	Circular	Identical	Recorded	Binaural
3	Random	Identical	Recorded	Binaural
4	Stationary	Joint	Recorded	Binaural
5	Circular	Joint	Recorded	Binaural
6	Random	Joint	Recorded	Binaural
7	Stationary	Separate	Recorded	Binaural
8	Circular	Separate	Recorded	Binaural
9	Random	Separate	Recorded	Binaural
10	Circular	Semi-fixed	Recorded	Binaural
11	Random	Semi-fixed	Recorded	Binaural
12	Stationary	Identical	Time-invariant	Binaural
13	Stationary	Separate	Time-invariant	Binaural
14	Stationary	Identical	Synthetic	Binaural
15	Stationary	Separate	Synthetic	Binaural
16	Circular	Separate	Recorded	Diotic

Sounds were recorded in a reduced-echo sound-isolated chamber with a Head Acoustics HSU III.2 head microphone, placed at the center of the chamber, and digitized with a National Instruments 4462 data acquisition card of an IBM PC at a sampling rate of 200 kHz and 16 bits resolution. In order to reduce noise inherent in on-site sound recordings, we post-processed the audio signals by removing the DC offset, bandpass filtering in the 200–700 Hz range with a sixth-order Butterworth filter, and applying a set of IIR notch filters within that frequency range (with no notches close to the tone frequencies)[1].

The two assistants, each wearing a helmet with the loudspeaker on top, performed a number of different scenarios during the recording session (for a full description of the recording session and methods, see Georgiou et al., 2011). Each experimental condition was based on the recording of a separate scenario. The three kinds of sound **Motion** tested were *Stationary* (based on scenarios with one or both assistants standing still about 80 cm in front of the head microphone, the location of the stationary assistant unless otherwise noted), *Circular* (assistants were walking along a circle with a radius of about 80 cm around the head microphone), and *Random* (assistants were freely wandering around the chamber). The **Co-location** of the two sound sources could either be characterized by *Identical* (based on recordings with one assistant standing or moving, his helmet-loudspeaker delivering the full alternating sequence), *Joint* (the two assistants moved together hand-in-hand, i.e. the distance between their trajectories being

roughly constant, with one loudspeaker delivering the high, the other the low tones – this was the mode of stimulus delivery for all scenarios with two assistants), *Semi-fixed* (scenarios with one assistant standing still at about 75 degrees to the right at a distance of about 170 cm while the other was moving on one of the trajectories), and S*eparate* trajectories (when the two assistants were standing separately – one assistant at 80 cm in front of the head microphone and the other at about 75 degrees to the right at a distance of about 170 cm – or moving independently of each other: in opposite directions for the *Circular* and independently of each other in the *Random* motion condition). The three **Motion** conditions were fully crossed with the four **Co-location** conditions, except that *Stationary* was not combined with *Semi-fixed*, as that would be identical to the *Stationary–Separate* combination. The resulting 11 conditions are summarized in *Table 1* (conditions 1–11) and illustrated in *Figure 1*.

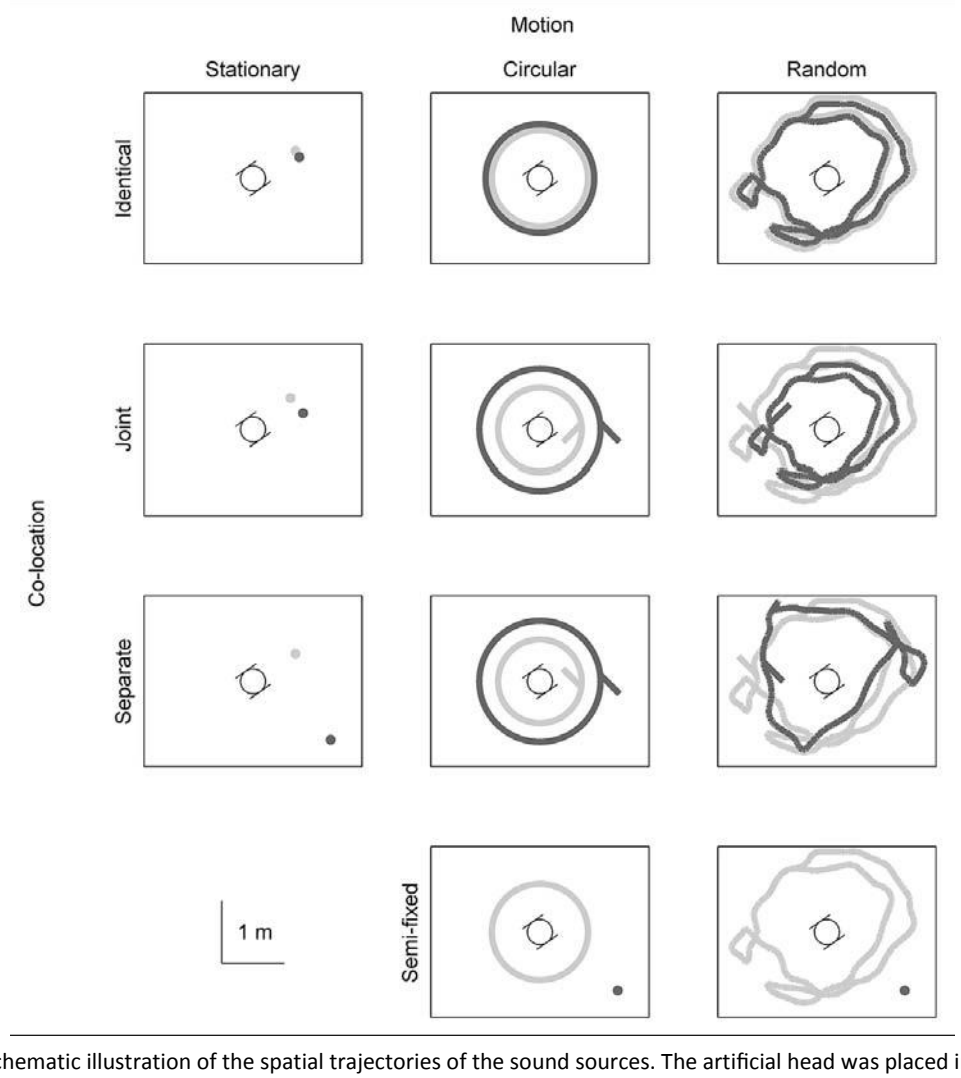

Figure 1. Schematic illustration of the spatial trajectories of the sound sources. The artificial head was placed in the center of the sound-isolated chamber, facing the top right corner. Thus listeners perceived the stimuli as if sitting in this position. The two trajectories are drawn in light and dark grey. For *Joint* and *Separate* motion, arrows indicate the direction of movement for each trajectory. Note that the trajectories are illustrations only and are not calculated from the recordings. Only the *Recorded Binaural* conditions are shown (conditions 1–11 in *Table 1*).

For creating each condition, we selected a contiguous section from the audio recording of the corresponding scenario and extended it to 4 minutes (the length of each stimulus block) by looping. We chose the longest possible segment of the signal that had good recording quality throughout and could be looped. We avoided introducing discontinuities into the spatial trajectories by selecting sections in which the initial and final estimated ITDs and their first derivatives roughly matched (separately for the high and the low tones). The endpoints of the sections were always placed in the middle of the silent interval before a 400 Hz tone. Note that the alternating tone sequence was chosen for the recordings because it provided more potential section endpoints than the repetitive triplet pattern typically used in the auditory streaming paradigm (van Noorden, 1975). In a pilot study we found that using identical stimulus parameters (Δf and Δt), the two patterns produce similar distributions of the reported percepts. While looping the audio segments to block length, a cross-fading procedure was applied in the 1 ms vicinity of the concatenation points to prevent audible clicks. The length of the segments chosen ranged between 13 and 43 seconds. The signals were then re-sampled at 192 kHz, so that the sampling interval was still smaller than the just noticeable difference for ITD (Blauert, 1997). Finally, the intensity of the sound sequence was normalized.

In order to mask any residual transient noise (e.g. footsteps) in the recordings, Brown noise segments lasting for 90 minutes were created separately for each participant and filtered with the same bandpass filter that was applied to the recordings. This noise was then played throughout the entire experimental session, including both the stimulus blocks and breaks between them. The masking threshold was determined to be ca. 5 dB in an informal test before the experiment. Therefore this signal-to-noise ratio was used for presenting the noise to each participant.

Synthetic and manipulated versions of the stimulus sequences were created for adding five control conditions (conditions 12–16 in *Table 1*). These conditions were presented to test the effects of unintended sounds and recording distortions, such as noises in the background, small spatial displacements of the loudspeakers (e.g., due to unavoidable head movements while the assistants were standing still), residual sounds from the ones produced by the assistants walking, and other unintentional cues appearing in the stimulus set on how listeners perceived the stimuli.

To this end, two *Synthetic* tone sequences were generated with the same parameters (frequencies, ramps and timing) as the recorded sounds, one with both high and low tones apparently arriving from the same spatial location and the other with the high tones apparently arriving from one and low tones from a different location. This was achieved by imposing on the synthetic alternating tone sequence the ITD (inter-aural time difference) and IID tracks extracted once from the *Stationary–Identical* and once from the *Stationary–Separate* tone sequence. Thus, these two synthetic sequences contained no noise or sound artifacts potentially present in the natural recordings. The stimulus sequences were then looped identically to the corresponding recorded segments.

However, the above sequences still contained the binaural cues produced by small movements of the assistants while standing still. Therefore, for the *Time-invariant* conditions, a single cycle of the repeating high-low tone sequence (one tone pair) was extracted and looped for 4 minutes, separately for the *Stationary–Identical* and the *Stationary–Separate* condition. Displacement of the assistants was assumed to be negligible within the period of 125 ms.

The *Diotic* control condition served to assess whether listeners utilized monaural cues, such as sound amplitude, which can be regarded as a cue for distance from the recording

microphone. Therefore, in the *Diotic* condition, the same audio signal was delivered to both ears, the original left or right channel of the *Circular–Separate* condition, balanced across participants.

In *Table 1,* the control conditions (12–16) are distinguished from the main test conditions by the variables **Stimulus type** and **Spatial cues**. **Stimulus type** can be *Recorded* (all main test sequences and the *Diotic* sequence), *Time-invariant,* or *Synthetic,* whereas **Spatial cues** can be *Binaural* (all but the *Diotic* sequence) and *Diotic*.

Procedure

Each experimental condition was administered in one 4-minute stimulus block during the experimental session. The order of the 16 conditions was randomized separately for each participant. There were short breaks lasting about 30 s between successive stimulus blocks and participants could choose to have a 5-minute break after the 8th block or as needed. The stimuli were played by an IBM PC with an Audiotrak Prodigy HD2 sound card, amplified by a custom-made mixer-amplifier and delivered by Etymotic Research ER-2 insert earphones. The insert earphones made sure that participants heard the sounds as if standing where the artificial head was located at the time when the sounds were recorded (i.e., spatial location cues related to head-related transfer functions were adequately reproduced). The masking noise was played from a separate computer continuously throughout the experiment and mixed together with the test sounds during the stimulus blocks. Participants were instructed to ignore the soft noise as a peculiarity of our equipment. The master sound level was set to 40 dB above the hearing threshold of the participant, as determined immediately before the experiment in a simplified staircase measurement using the frequencies of the test tones. Participants were seated in the same anechoic chamber in which the sound recordings had been carried out.

Participants were instructed to continuously report the perceived sound organization throughout the entire stimulus block using two response keys, one key held in each hand. They were to depress one key when they perceived both high and low tones as part of a single repeating pattern (termed the ‘integrated’ percept). The other was to be depressed when they heard tones of the same pitch forming separate repeating patterns (the ‘segregated’ percept). When they heard both types of patterns concurrently, they were to keep both keys depressed (the ‘both’ percept). During times when the participant did not perceive any repeating sound pattern, he/she was instructed to release both keys (the ‘neither’ percept). The instructions emphasized that the appropriate key combination was to be held as long as the participant perceived the corresponding sound organization but to be changed immediately with a change in the perceived organization. Participants were also told that there is no correct or incorrect way to perceive any of the stimulus sequences. Thus they should not try to force to hear the sounds in one or another way. Rather, they should report what they actually hear. A description of the interpretation of the percepts reported by depressing both keys at the same time can be found in Denham et al. (in this issue). The assignment of the keys was counterbalanced across participants. The ‘integrated’ and ‘segregated’ percepts were explained and illustrated (both with sound examples and visual illustrations) to the participants before the experiment, and the experimenter made sure they understood the task (for further details about the instructions, see Denham et al., 2013). Because there is no single prototype for the “both” percept, listeners were not trained specifically on it, but were only told to use it when they experienced both an

integrated and a segregated pattern, in parallel.

Data recording and analysis

The state of the two response keys was sampled at a 250 Hz rate, and the data was analyzed similarly to the procedure used in Denham et al. (2013). For each perceptual phase (i.e., the time interval between two consecutive perceptual switches), the logarithm of its duration in milliseconds and the reported percept was extracted. Perceptual phases shorter than 300 ms were excluded from the analysis as these presumably originate from inaccurate timing of button presses and releases, rather than from two separate perceptual switches quickly following each other (Moreno-Bote et al., 2010). Based on this data, we calculated the mean proportions of each percept (i.e., the percent of time experiencing a given percept within the stimulus block) and mean perceptual phase durations, separately for each participant, perceptual organization (the 'integrated', 'segregated' and 'both' percepts), and condition. 'Neither' responses were not analyzed as they appear in only 3.7% of the block time in all conditions and are typically shorter than 1s. Two repeated measures analyses of variance (ANOVAs) were carried out on each of the six data sets (proportion vs. phase duration × 'integrated' vs. 'segregated' vs. 'both' percept) with **Motion** and **Co-location** as the dependent factors. In one ANOVA, **Motion** was represented by the conditions *Stationary*, *Circular,* and *Random* with **Co-location** being *Identical* and *Separate* (conditions 1–3 and 7–9 in *Table 1*). In the other ANOVA, **Motion** could be *Circular* or *Random* with **Co-location** being *Semi-fixed*, *Joint* or *Separate* (conditions 5–6, 8–9, and 10–11 in *Table 1*). The *Stationary–Joint* condition (condition 4 in *Table 1*) was excluded from the analyses because, after completing the experiment, we realized that in fact, this was also a *Stationary–Separate* type of condition, since the azimuth difference between the two speakers when the two assistants were standing shoulder-to-shoulder was roughly 30 degrees, well above the just noticeable difference (Blauert, 1997).

The same six variables (mean time-proportions and average phase durations of the 'integrated', 'segregated' and 'both' percepts) were analyzed in two ANOVAs, each to test for some possible extraneous effects. In one ANOVA, responses from the *Diotic* control condition were compared with the responses from the corresponding *Binaural* condition (conditions 8 and 16 in *Table 1*). In the other ANOVA, the factors were **Stimulus type** (*Recorded* vs. *Synthetic* vs. *Time-invariant*) and **Co-location** (*Identical* vs. *Separate*) for the stationary sound sources (conditions 1, 7, 12–13, and 14–15 in *Table 1*).

Where applicable, degrees of freedom were adjusted with the Greenhouse–Geisser correction factor (ε). These and the partial η^2 effect sizes are reported for significant effects. Post hoc comparisons were performed using Tukey's HSD tests. All analyses were carried out at the 95% confidence level.

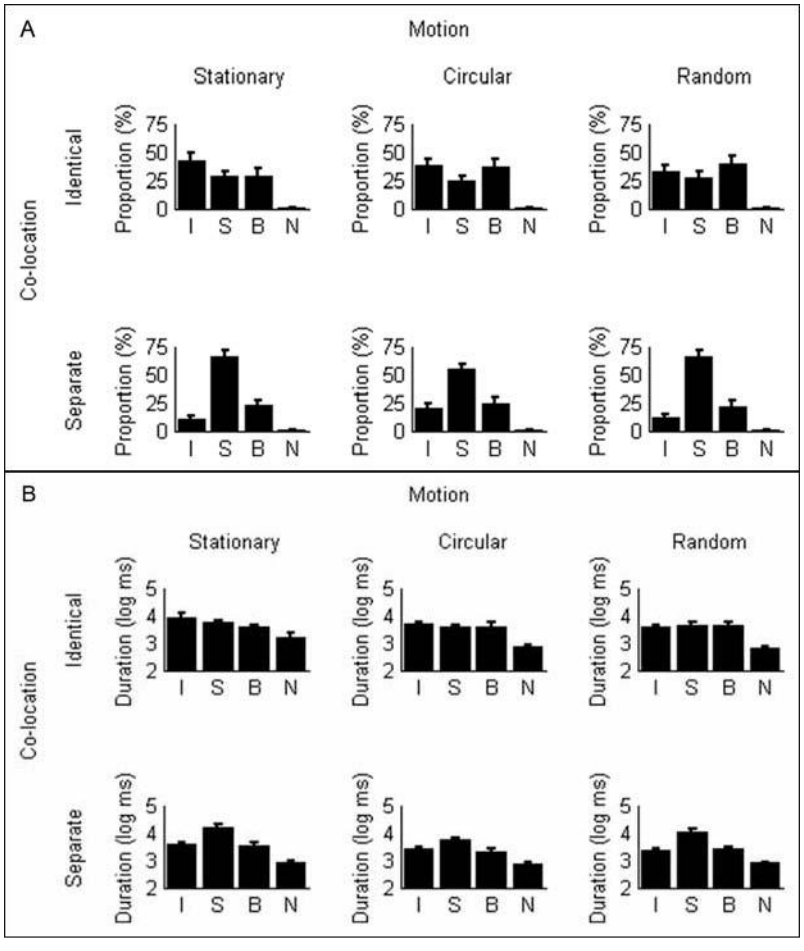

Figure 2. Group-averaged (*N* = 19) percentage of time (A) and average phase durations (B) of the four response alternatives (I: 'integrated', S: 'segregated', B: 'both', N: 'neither') are shown for the *Stationary*, *Circular*, and *Random* **Motion** conditions, separately for *Identical* and *Separate* **Co-locations**. Error bars show the standard error of the means.

RESULTS

Effects of motion and mean spatial separation

Figure 2 shows the group-averaged time percentages and average phase durations of the four perceptual alternatives reported by the listeners ('integrated', 'segregated', 'both', and 'neither') for the *Stationary*, *Circular*, and *Random* **Motion** conditions, separately for *Identical* and *Separate* **Co-locations**. In the six ANOVAs with these factors, the main effect of Motion was limited to the durations of 'integrated' and 'segregated' phases (the results of the analysis is summarized in *Table 2*). Post-hoc tests revealed that for both kinds of percepts, the significant main effect reflected shorter phase durations when the sound source was moving compared with that obtained for stationary sound sources ($df = 36$, $p < 0.05$ and $p < 0.001$ for the 'integrated' phase duration difference between the *Stationary* and the *Circular* and between the *Stationary* and the *Random* conditions, respectively; $df = 36$, $p < 0.001$ and $p < 0.05$ for the 'segregated' phase duration difference between the *Stationary* and the *Circular* and between the *Circular* and the *Random* condition, respectively).

Table 2. Significant effects obtained in the ANOVA for the *Stationary*, *Circular*, and *Random* **Motion** conditions with *Identical* and *Separate* **Co-locations**. "Measure" indicates the measure compared: Int-Prop = proportion of 'integrated' phases; SegProp = proportion of 'segregated' phases; BothProp = proportion of 'both' phases; IntDur = average duration of all 'integrated' phases; SegDur = average duration of all 'segregated' phases; BothDur = average duration of all 'both' phases. "Factors" are the ANOVA factors. Degrees of freedom (*df*), *F* values (*F*), significance levels (*p*), Greenhouse–Geisser correction factors (G–G, where applicable), and η^2 effect sizes are shown. Significant effects are typed in boldface.

Measure	**Factor**	**df**	**F**	**p**	**G-G**	**η^2**
IntProp	Motion	2,36	1.954	0.161	0.917	
	Co-location	**1,18**	**40.809**	**< 0.001**		**0.766**
	Motion × Co-location	2,36	1.850	0.182	0.761	
SegProp	Motion	2,36	3.480	0.052	0.827	
	Co-location	**1,18**	**55.750**	**< 0.001**		**0.756**
	Motion × Co-location	2,36	0.844	0.407	0.721	
BothProp	Motion	2,36	1.196	0.313	0.937	
	Co-location	**1,18**	**16.345**	**< 0.001**		**0.476**
	Motion × Co-location	2,36	1.746	0.193	0.895	
IntDur	**Motion**	**2,36**	**11.008**	**< 0.001**	**0.764**	**0.379**
	Co-location	**1,18**	**23.108**	**< 0.001**		**0.562**
	Motion × Co-location	2,36	0.566	0.554	0.888	
SegDur	**Motion**	**2,36**	**11.900**	**< 0.001**	**0.921**	**0.398**
	Co-location	**1,18**	**32.719**	**< 0.001**		**0.645**
	Motion × Co-location	2,36	2.178	0.133	0.912	
BothDur	Motion	2,36	1.602	0.217	0.951	
	Co-location	**1,18**	**7.152**	**< 0.05**		**0.398**
	Motion × Co-location	2,36	2.487	0.124	0.609	

Co-location also showed a main effect with participants reporting segregation for a higher proportion of time when listening to tone sequence with *Separate* than with *Identical* trajectories. At the same time, the proportion of the 'integrated' and 'both' responses decreased. In accordance, 'segregated' percepts were longer and 'integrated' and 'both' percepts were shorter on average for *Separate* than for *Identical* trajectories. No significant interaction was obtained between **Motion** and **Co-location**.

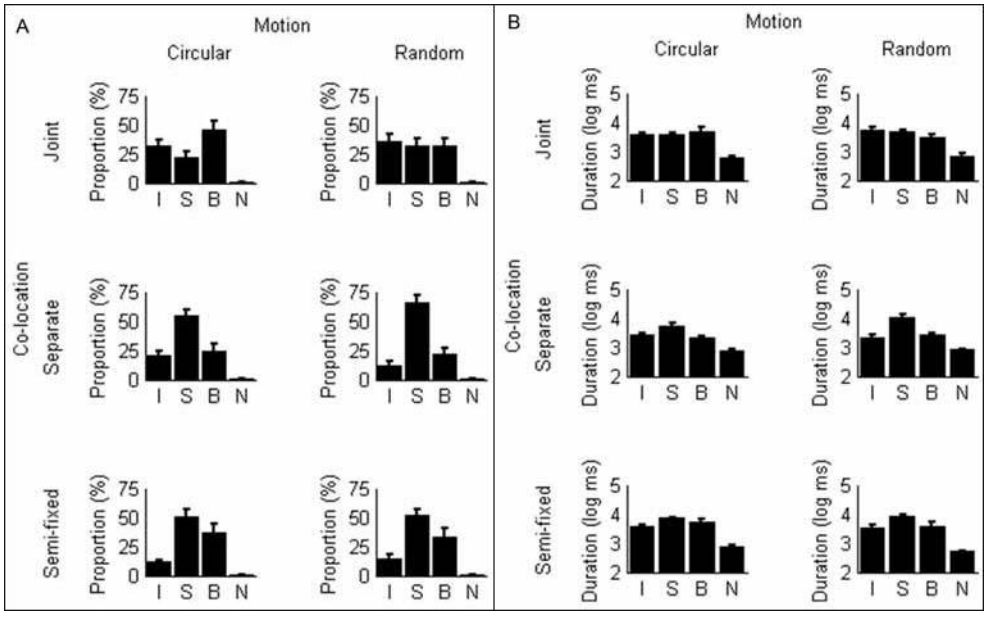

Figure 3. Group-averaged (*N* = 19) percentage of time (A) and average phase durations (B) of the four response alternatives (I: 'integrated', S: 'segregated', B: 'both', N: 'neither') are shown for the *Circular*, and *Random* **Motion** conditions, separately for *Semi-fixed*, *Joint* and *Separate* **Co-locations**. Error bars show the standard error of the means.

Figure 3 plots the group-averaged time percentages and average phase durations of the four perceptual alternatives reported by the listeners for the *Circular* and *Random* **Motion** conditions, separately for *Semi-fixed*, *Joint,* and *Separate* **Co-locations**. **Motion** had a significant main effect both on the proportion and duration of the 'segregated' phases, which were significantly larger, whereas the proportion of 'both' responses was lower for the *Random* compared with the *Circular* **Motion** condition (see *Table 3* for the details of the analysis). No other **Motion** main effects were significant.

The main effect of **Co-location** was found to be significant in all six ANOVAs. According to post-hoc analyses, tone sequences with *Joint* and *Separate* trajectories were perceived differently, with the former resulting in higher proportions and longer phase durations for 'integrated' and 'both' percepts ($df = 36$, $p < 0.001$ and $p < 0.01$ for the 'integrated' proportions and phase durations, respectively; and $p < 0.01$ and $p < 0.05$ for the 'both' proportions and phase durations, respectively) and *Separate* trajectories resulting in lower proportions and shorter phase durations for 'segregated' percepts ($p < 0.001$ and $p < 0.01$, respectively). In the *Semi-fixed* conditions, integration was perceived for a lower percentage of time, and 'segregated' for a higher percentage with longer perceptual phases than in the *Joint* conditions ($df = 36$, $p < 0.001$, $p < 0.001$, and $p < 0.05$, respectively). Finally, 'both' responses occurred with a higher proportion and were longer for the *Semifixed* than for the *Separate* trajectories ($df = 36$, $p < 0.05$ and $p < 0.01$, respectively). There were no significant interactions between **Motion** and **Co-location**, except for the phase duration of 'both' responses.

Table 3. Significant effects obtained in the ANOVA for the *Circular*, and *Random* **Motion** conditions with *Semi-fixed*, *Joint* and *Separate* **Co-locations**. "Measure" indicates the measure compared: IntProp = proportion of 'integrated' phases; SegProp = proportion of 'segregated' phases; BothProp = proportion of 'both' phases; IntDur = average duration of all 'integrated' phases; SegDur = average duration of all 'segregated' phases; BothDur = average duration of all 'both' phases. "Factors" are the ANOVA factors. Degrees of freedom (*df*), *F* values (*F*), significance levels (*p*), Greenhouse–Geisser correction factors (G–G, where applicable), and η^2 effect sizes are shown. Significant effects are typed in boldface

Measure	**Factor**	**df**	**F**	**p**	**G-G**	**η^2**
IntProp	Motion	1,18	0.008	0.929		
	Co-location	**2,36**	**23.360**	**< 0.001**	**0.696**	**0.565**
	Motion × Co-location	2,36	2.677	0.100	0.736	
SegProp	**Motion**	**1,18**	**5.480**	**< 0.05**		**0.233**
	Co-location	**2,36**	**22.991**	**< 0.001**	**0.874**	**0.561**
	Motion × Co-location	2,36	1.555	0.229	0.808	
BothProp	**Motion**	**1,18**	**4.427**	**< 0.05**		**0.197**
	Co-location	**2,36**	**7.577**	**< 0.01**	**0.926**	**0.296**
	Motion × Co-location	2,36	1.546	0.230	0.852	
IntDur	Motion	1,18	0.022	0.883		
	Co-location	**2,36**	**7.210**	**< 0.01**	**0.853**	**0.286**
	Motion × Co-location	2,36	2.804	0.089	0.784	
SegDur	**Motion**	**1,18**	**16.313**	**< 0.001**		**0.475**
	Co-location	**2,36**	**6.933**	**< 0.01**	**0.779**	**0.278**
	Motion × Co-location	2,36	3.509	0.054	0.774	
BothDur	Motion	1,18	1.214	0.285		
	Co-location	**2,36**	**7.922**	**< 0.01**	**0.946**	**0.306**
	Motion × Co-location	**2,36**	**3.448**	**< 0.05**	**0.969**	**0.161**

Control analyses

Figure 4 compares the proportions and average durations of the four percepts obtained in the *Diotic* and the corresponding *Binaural* condition. By restricting participants' localization of the tones to monaural cues, we elicited an increase in the proportion and phase duration of 'both' responses, accompanied by a decrease in the proportion of the 'segregated' percept ($F(1,18) = 6.620$, $p < 0.05$, $\eta^2 = 0.269$; $F(1,18) = 10.479$, $p < 0.01$, $\eta^2 = 0.368$, and $F(1,18) = 5.541$, $p < 0.05$, $\eta^2 = 0.235$, respectively).

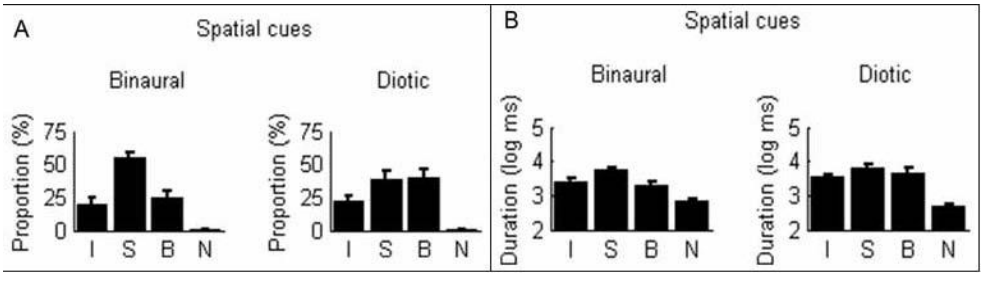

Figure 4. Group-averaged (*N* = 19) percentage of time (A) and average phase durations (B) of the four response alternatives (I: 'integrated', S: 'segregated', B: 'both', N: 'neither') are shown for conditions with *Binaural* and *Diotic* **Spatial cues**. Error bars show the standard error of the means.

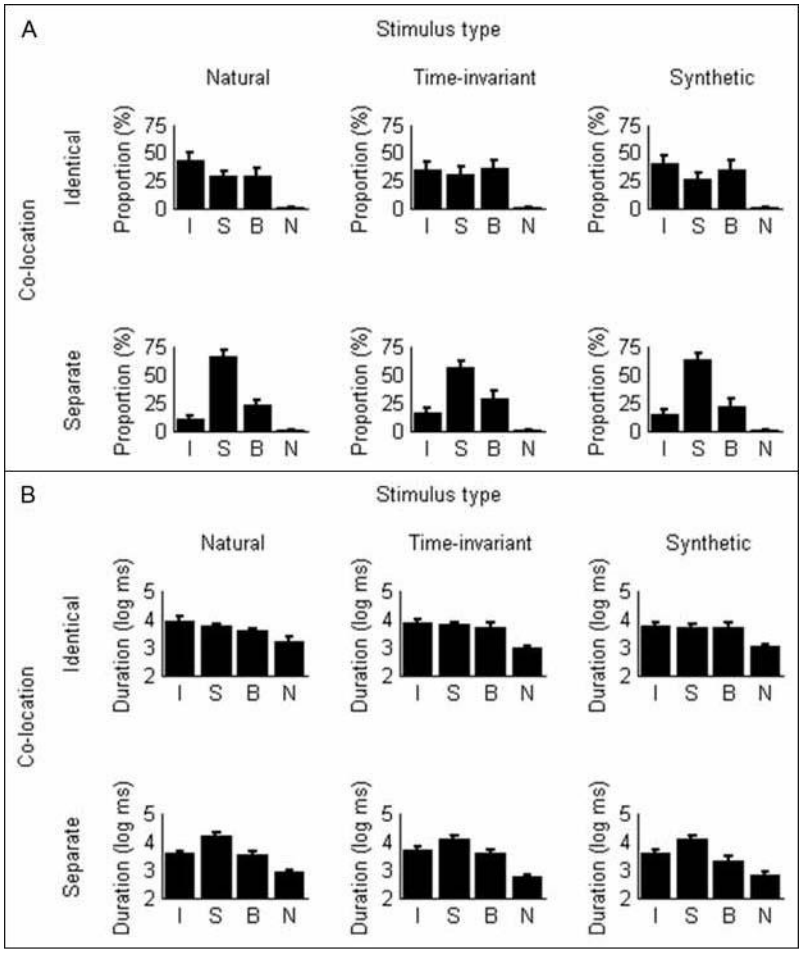

Figure 5. Group-averaged (*N* = 19) percentage of time (A) and average phase durations (B) of the four response alternatives (I: 'integrated', S: 'segregated', B: 'both', N: 'neither') are shown for the *Recorded*, *Time-invariant*, and *Synthetic* **Stimulus type** conditions. Error bars show the standard error of the means.

The results obtained with three different **Stimulus types** (*Recorded*, *Time-invariant* and *Synthetic*) for stationary sound sources with *Identical* and *Separate* **Co-location** are plotted in *Figure 5*. In all six analyses, we found a significant main effect of **Co-location** ($F(1,18) = 32.691$, $p < 0.001$, $\eta^2 = 0.645$ for 'integrated' proportions; $F(1,18) = 40.828$, $p < 0.001$, $\eta^2 = 0.694$ for 'segregated' proportions; $F(1,18) = 4.915$, $p < 0.05$, $\eta^2 = 0.214$ for 'both' proportions; $F(1,18) = 21.405$, $p < 0.001$, $\eta^2 = 0.543$ for 'integrated' durations; $F(1,18) = 26.963$, $p < 0.001$, $\eta^2 = 0.600$ for 'segregated' durations; and $F(1,18) = 6.277$, $p < 0.05$, $\eta^2 = 0.258$ for 'both' durations). Stimulus type did not have a significant effect in any of the analyses, and there were no significant interactions between the two factors.

DISCUSSION

We investigated whether sound source motion is used as a primary cue in auditory stream segregation. We hypothesized that sounds emitted by two sources moving on a common trajectory would be more likely to be grouped together, and sounds emitted by two sources moving on separate (independent) trajectories would be more likely to be perceived as

segregated. However, we found no clear effect of sound source motion on the perceptual organization of the test sequences. Based on the common-fate principle, we expected to fi nd significant interactions between the presence or absence of motion and the co-location (average spatial separation) of the sound sources. Specifically, we expected to fi nd differences in perceptual organization between sound sources moving on common trajectories (*Identical* and *Joint* conditions) in comparison with stationary sources that are spatially identically located or close to each other, as well as between sound sources moving on separate trajectories (*Separate* and *Semi-fixed* conditions) in comparison with stationary, spatially separated sound sources. However, no significant interaction was obtained between the **Motion** and **Co-location** factors in all but one analysis (the average duration of 'both' percepts). The data suggests that the auditory system did not utilize auditory motion cues for segregating streams for the sequences delivered in our experiment.

Thus our data do not support the hypothesis that auditory motion can serve as a primary cue in auditory stream segregation. The results provide no evidence that the effects of auditory motion can be described as a case of the Gestalt principle of common fate (Köhler, 1947). In contrast, spatial separation – the lack of 'similarity', in terms of the Gestalt principles – strongly facilitated the segregation of the sounds into separate streams. It is possible that we overestimated the reliability of the cues provided by joint and separate trajectories. These cues may not serve as an efficient heuristic if we often encounter multiple objects moving together. This may in fact be the case when we hear chatting people passing by or moving cars with a number of spatially joint sound sources. Alternatively, one could argue that despite the motion cues being recorded from a real-life scene, the sounds themselves were not ecologically valid in that they were discontinuous and contained only a single frequency. Both of these features could have made the extraction of auditory motion cues more difficult. However, according to informal reports, participants clearly heard the sound sources moving around in space (in the appropriate conditions). Thus it is not very likely that their auditory system could not extract these cues.

The effect of auditory motion showed up in decreasing the average phase durations of both the 'integrated' and the 'segregated' percepts compared to stationary sound sources without corresponding effects on the proportions of these perceptual organizations. Shorter average phase durations correspond to faster switching for moving compared with stationary sound sources. One may speculate that changes in the spatial location of the sound sources reduce the efficacy of the cues in stabilizing perception by forcing the system to reevaluate the continuity of its perceptual objects. This may have been exacerbated by the discrete sounds delivered in the current experiment. Thus one would expect to find smaller or no increase in perceptual switching with continuous sounds.

Spatial separation promoted the 'segregated' percept, as was observed in previous studies for stationary sound sources (Denham et al., 2009; Szalárdy et al., 2013). We extended these previous observations to sound sources located in real space and to moving sound sources. Participants also reported fewer and shorter 'both' percepts in these cases. The **Co-location** of concurrent sound sources determines their mean spatial separation: when the two sound sources moved on separate trajectories, they were spatially clearly separated during most of the stimulus block (except for the relatively short time intervals during which the two sources crossed each other's paths), and they were less separated or not separated at all in conditions with joint and identical trajectories.

The effect of location differences was substantially reduced for the diotic stimuli. When

binaural cues were removed, the proportion of ‘segregated’ percepts decreased and ’both’ percepts became more frequent and longer on average. Thus the proportions of the reported percepts were similar to those for identically located sound sources. Therefore, binaural location cues are likely to underlie the ubiquitous main effect of spatial separation (see above).

Our results cannot be accounted for by noises or other acoustic artifacts remaining in the recorded stimulus sequences due to the recording environment, or by small fl uctuations in the nominally stationary sound source locations over time, because exchanging the recorded stimulus sequences for stationary sound sources for sequences created by looping a single cycle or for a synthesized sequence with the same binaural cues did not cause significant changes in the perceptual organization.

One of the two competing theories of auditory motion detection (Grantham, 1995; Middlebrooks & Green, 1991) suggests the existence of specific motion-sensitive mechanisms. This theory would imply for the current results that an available potential cue of auditory stream segregation is ignored by the auditory system. Though our results do not provide conclusive evidence in favor of either theory, the “snapshot” mechanism (which proposes that properties of sound source motion are inferred from subsequent “snapshots” of the perceived sound location) is more consistent with our findings, as it suggests that motion information becomes available only after the source location “snapshots” have been processed. This would also be compatible with the finding of increased switching between alternative organizations when the sound sources are moving.

ACKNOWLEDGEMENTS

This work was supported by the European Commission's Seventh Framework Programme (ICT-FP7-231168), the Lendület project awarded to IW by the Hungarian Academy of Sciences (contract number LP2012-36/2012), the German Academic Exchange Service (Deutscher Akademischer Austauschdienst, DAAD, Project 50345549), the Russian Foundation for Fundamental Research (Project 11-04-00008-a), the Russian Academy of Sciences – Hungarian Academy of Sciences (Exchange Cooperation Agreement 2010-2012), and the Hungarian Scholarship Board (Magyar Ösztöndíj Bizottság, MÖB, Project P-MÖB/853). The authors thank Zsuzsanna D'Albini for collecting the perceptual data.

NOTE

[1] We found a 100 Hz artifact in the recorded audio signal, with harmonics up to quite high frequencies. This steady narrow-band signal was substantially softer than the actual tones. The purpose of the IIR notch filters was to further reduce the amplitude of this artifact within the passband of the Butterworth fi lter. Notches were placed at 100 Hz, 200 Hz, 300 Hz, 600 Hz and 700 Hz. The notch filters closest to the stimulus frequencies (at 300 Hz and 600 Hz) had a bandwidth of 0.1 Hz, while the bandwidth of the rest of the notch filters was 1 Hz. According to informal listening tests, the artifact became inaudible in the notch-filtered sound signals.

REFERENCES

Altman, J. A., Vaitulevich, S. P., Shestopalova, L. B., Petropavlovskaia, E. A. (2010): How does mismatch negativity reflect auditory motion? *Hearing Research, 268,* 194–201.

Altman, J. A., Vaitulevich, S. P., Shestopalova, L. B., Varfolomeev, A. L. (2005). Mismatch negativity evoked by stationary and moving auditory images of different azimuthal positions. *Neuroscience Letters, 384,* 330–335.

Blauert, J. (1997). *Spatial Hearing*. Cambridge, Massachusetts: MIT Press.

Bregman, A. S. (1990). *Auditory Scene Analysis*. Cambridge, Massachusetts: MIT Press.

Denham, S. L., Gyimesi, K., Stefanics, G., Winker, I. (2013). Perceptual bi-stability in auditory streaming: How much do stimulus features matter? *Learning and Perception, 5*(Suppl. 2), 73–100. (this issue)

Denham, S. L., Gyimesi, K., Stefanics, G., Winkler, I. (2009). Stability of perceptual organisation in auditory streaming. In: Lopez-Poveda, E.A., Palmer, A. R., Meddis, R. (eds.), *The Neurophysiological Bases of Auditory Perception* (pp. 477–488). Springer.

Denham, S. L., Winkler, I. (2006). The role of predictive models in the formation of auditory streams. *Journal of Physiology–Paris, 100,* 154–170.

Georgiou, J., Pouliquen, P., Cassidy, A., Garreau, G., Andreou, C., Stuarts, G. et al. (2011). A multimodal-corpus data collection system for cognitive acoustic scene analysis. In *45th Annual Conference on Information Sciences and Systems (CISS 2011)* (pp. 1–6).

Grantham, D. W. (1995). Spatial hearing and related phenomena. In B.C.J.Moore (Ed.), *Hearing* (pp. 297–346). San Diego: Academic Press.

Grimault, N., Bacon, S. P., Micheyl, C. (2002). Auditory stream segregation on the basis of amplitude-modulation rate. *Journal of the Acoustical Society of America, 111,* 1340–1348.

Judd, T. (1977). An explanation of Deutsch's scale illusion. Unpublished manuscript. Department of Psychology, Cornell University.

Köhler, W. (1947). *Gestalt Psychology*. (2 ed.) New York: Liveright.

Middlebrooks, J. C., Green, D. M. (1991). Sound Localization by Human Listeners. *Annual Review of Psychology, 42,* 135–159.

Moore, B. C. J., Gockel, H. (2002). Factors influencing sequential stream segregation. *Acta Acustica United with Acustica, 88,* 320–333.

Moreno-Bote, R., Shpiro, A., Rinzel, J., Rubin, N. (2010). Alternation rate in perceptual bistability is maximal at and symmetric around equi-dominance. *Journal of Vision, 10*.

Perrott, D. R., Marlborough, K. (1989). Minimum Audible Movement Angle – Marking the End-Points of the Path Traveled by A Moving Sound Source. *Journal of the Acoustical Society of America, 85,* 1773–1775.

Pressnitzer, D., Hupé, J. M. (2006). Temporal dynamics of auditory and visual bistability reveal common principles of perceptual organization. *Current Biology, 16,* 1351–1357.

Roberts, B., Glasberg, B. R., Moore, B. C. (2002). Primitive stream segregation of tone sequences without differences in fundamental frequency or passband. *Journal of the Acoustical Society of America, 112,* 2074–2085.

Smith, J., Hausfeld, S., Power, R. P., Gorta, A. (1982). Ambiguous musical figures and auditory streaming. *Percept.Psychophys., 32,* 454–464.

Szalárdy, O., Bendixen, A., Tóth, D., Denham, S. L., Winkler, I. (2013). Modulation frequency acts as a primary cue for auditory stream segregation. *Learning and Perception, 5*(Suppl. 2) 149–161 (this issue)

van Noorden, L. P. A. S. (1975). *Temporal coherence in the perception of tone sequences.* Ph.D. Eindhoven University of Technology, Leiden, The Netherlands. Vliegen, J., Oxenham, A. J. (1999). Sequential stream segregation in the absence of spectral cues. *Journal of the Acoustical Society of America, 105,* 339–346.

Winkler, I. (2007). Interpreting the mismatch negativity. *Journal of Psychophysiology, 21,* 147–163.

Winkler, I., Denham, S. L., Nelken, I. (2009). Modeling the auditory scene: predictive regularity representations and perceptual objects. *Trends in Cognitive Sciences, 13,* 532–540.