Exp Brain Res (2014) 232:273–282
DOI 10.1007/s00221-013-3738-3

RESEARCH ARTICLE

The role of spatiotemporal and spectral cues in segregating short sound events: evidence from auditory Ternus display

Qingcui Wang · Ming Bao · Lihan Chen

Received: 1 July 2013 / Accepted: 3 October 2013 / Published online: 20 October 2013

Abstract Previous studies using auditory sequences with rapid repetition of tones revealed that spatiotemporal cues and spectral cues are important cues used to fuse or segregate sound streams. However, the perceptual grouping was partially driven by the cognitive processing of the periodicity cues of the long sequence. Here, we investigate whether perceptual groupings (spatiotemporal grouping vs. frequency grouping) could also be applicable to short auditory sequences, where auditory perceptual organization is mainly subserved by lower levels of perceptual processing. To find the answer to that question, we conducted two experiments using an auditory Ternus display. The display was composed of three speakers (A, B and C), with each speaker consecutively emitting one sound consisting of two frames (AB and BC). Experiment 1 manipulated both spatial and temporal factors. We implemented three 'within-frame intervals' (WFIs, or intervals between A and B, and between B and C), seven 'inter-frame intervals' (IFIs, or intervals between AB and BC) and two different speaker layouts (inter-distance of speakers: near or far). Experiment 2 manipulated the differentiations of frequencies between two auditory frames, in addition to the spatiotemporal cues as in Experiment 1. Listeners were required to make two alternative forced choices (2AFC) to report the perception of a given Ternus display: element motion (auditory apparent motion from sound A to B to C) or group motion (auditory apparent motion from sound 'AB' to 'BC'). The results indicate that the perceptual grouping of short auditory sequences (materialized by the perceptual decisions of the auditory Ternus display) was modulated by temporal and spectral cues, with the latter contributing more to segregating auditory events. Spatial layout plays a less role in perceptual organization. These results could be accounted for by the 'peripheral channeling' theory.

Keywords Perceptual organization · Temporal cues · Frequency · Auditory · Ternus display

Q. Wang · M. Bao (✉)
Key Laboratory of Noise and Vibration Research,
Institute of Acoustics, Chinese Academy of Sciences,
Beijing 100190, China
e-mail: baoming@mail.ioa.ac.cn

Q. Wang
e-mail: wangqingcui@mail.ioa.ac.cn

L. Chen (✉)
Department of Psychology and Key Laboratory of Machine Perception (Ministry of Education), Peking University,
Beijing 100871, China
e-mail: CLH20000@gmail.com

Introduction

Perceptual organization has a strong effect on how we hear the world (Cusack and Carlyon 2004). In natural and complicated acoustic scenes, we often hear sounds emanating from various sources. However, we have the ability to readily attend and identify both a specific, simple auditory object and more sophisticated auditory streams. This is accomplished with minimum interference from any background distracter auditory inputs. The process of separating a target auditory event or auditory stream from these distracters was first understood as the 'cocktail party problem' (Cherry 1953). Afterward, this phenomenon spawned extensive studies (Cooper and Roberts 2007; Denham and Winkler 2006; Takegata et al. 2005; Yabe et al. 2001).

The 'cocktail party problem' has been most commonly investigated through the use of complex auditory scenarios

and perceptually uncertain auditory stimuli (Cusack 2005; Pressnitzer and Hupé 2006; Szalárdy et al. 2013). Similar to the processes related to vision, there are common principles of perceptual organization within the auditory domain. The different perceptions of auditory bi-stability were mutually exclusive. Each perception features its own randomized duration distribution (Kondo et al. 2012; Pressnitzer and Hupé 2006). Among the many different types of stimulus presentations, the 'ABA—stimulus paradigm' has been one of the most popular paradigms used in the study of auditory scene analysis (ASA) (Bregman and Campbell 1971; Bregman 1990; Füllgrabe and Moore 2012). In this paradigm, a sequence of alternating tones can be perceived as either one coherent stream or two separate streams, due to differences in the features (temporal and spectral) between the A and B sounds. For example, when sequences had slow tone rates and/or small pitch differences, participants heard the sequence as one perceptual object or a single auditory stream—a single tone rising and falling over time. When the sequence was faster or the pitch gap was widened, participants heard the sequence as two streams, i.e., one tone rising and falling within the high range and one tone doing the same within the low range (Bregman and Campbell 1971; Bregman 1990).

In an ABA paradigm, the task of sound segregation is related to the separation and perceptual binding of sound over a period of time. Based on extensive psychophysical research in humans, Bregman (1990) proposed a distinction between 'primitive' and 'schema-based' processes in ASA. According to Bregman (1990), the former (primitive) is a data-driven phenomenon that consists of pre-attentive auditory processes that are both automatic and obligatory. In these 'primitive' processes, different acoustic attributes (such as frequency separation, pitch timbre and spatial location) are important cues in ASA (McCabe and Denham 1997), which all combine to trigger a bottom-up process. Among these cues, spatial location was thought to play a secondary role in the formation of the auditory streams (Darwin and Carlyon 1995; Oxenham 2000). Temporal coherence (the inter-tone interval, onset time of the auditory sequence) has the potential to facilitate one coherent auditory stream (Bee and Klump 2005; Fishman et al. 2001; Shamma et al. 2011). Spectral or tonotopic contrasts can also be used in stream segregation(Shamma and Micheyl 2010), in which frequency-to-place mapping is used as a guiding anatomical and functional principle within the auditory system (Eggermont 2001). The primitive process suggests that ASA plays an intrinsic role in the active and flexible perceptual exploration of the acoustic environment. It also holds that the perceptual decision toward the auditory stream can be very fast (Anstis and Saida 1985; Bregman 1990; Denham and Winkler 2006).

The schema-based scene analysis, on the other hand, refers to perceptual grouping processes that demand high-level, cognitive input and are influenced by the listener's attention and prior expectations based on previous learning (for instance, periodicity information as a higher level of knowledge). Hence, a schema-based analysis is a top-down process (Bregman 1990). On a given trial, listeners initially perceive one stream, and only after several seconds of buildup does the pattern of alternating tones split into two distinct streams (Anstis and Saida 1985; Bregman 1978). Most studies concerned with the buildup of the perceptual organization have one key assumption, i.e., that all sounds are considered to be part of one stream, with an initial default coherence point. The auditory system then segregates sounds into separate streams when enough evidence has been accumulated by the auditory system over several seconds (Bregman 1990). The effect of buildup in long auditory sequences usually occurs at a later and less automatic stage of processing (Snyder et al. 2006). In an ABA paradigm, segregation of sounds is likely to begin in the auditory periphery and continue at least to the primary auditory cortex for simple cues such as pure-tone frequency, but at stages as high as the secondary auditory cortex for more complex cues such as periodicity pitch (Snyder and Alain 2007).

The unfolding of the long sequence provides a greater number of complex cues, such as periodicity information, and hence diminishes the effectiveness of directly testing whether perceptual organization is being driven more by bottom-up-related auditory features (largely immune from periodicity information and in the absence of higher-level attentional and cognitive inputs). Furthermore, in the ABA-stimulus paradigm, whether a participant hears one stream or two streams is not simply a matter of the stimulus characteristics and the amount of time that has passed since the beginning of the sequence.

Rather, streaming may be a dynamic process, by which representations for different perceptual solutions compete. During the presentation, each tone potentially serves to mask a subsequent tone and act as a signal tone following a preceding masking tone (Beauvois 1998; Beauvois and Meddis 1991, 1996; Fishman et al. 2001; Hartmann and Johnson 1991; McCabe and Denham 1997). This masking phenomenon poses a risk of blurring the boundary of auditory stream segregation and makes it difficult to pinpoint a potentially early perceptual decision ('voting') toward the segregation of an auditory stream. Nevertheless, investigation of the automatic and quick nature of perceptual decisions about auditory objects is important. Stream segregation takes time to occur, which may hamper adaptation in natural settings, in situations where the rapid parsing of sounds into streams could be an important prerequisite for survival. A better understanding of how the auditory system

chooses whether perception will consist of one stream or two streams will likely inform mechanisms of perception with implications for other sensory modalities, such as vision.

Simply stated, streaming can be classified as a bi-stable perceptual phenomenon, which implies that an important aspect of streaming, in addition to the segregation and build-up processes, is how the nervous system decides at any point in time what the perceptual experience of the listener is (i.e., 'voting'). A design that uses a short auditory sequence, while at the same time maintaining the nature of perceptual uncertainty, could meet the experimental requirement of exploring the potentially rapid perceptual organization of auditory stimuli. To achieve this, in the current study, we attempt to focus more upon initial primitive stream segregation processes by using ambiguous short auditory sequences (eliciting bi-stable perception) that provide little if any periodicity information. The 'bi-stable' perception was assumed to be automatic and less subject to volitional control (Winkler et al. 2006). To address the issue of auditory perceptual organization in a short auditory sequence, we have developed a new paradigm known as the auditory Ternus display. Using this paradigm, we have also examined the roles of spatiotemporal and spectral cues in terms of segregating sound events within a single unified design.

The auditory Ternus display is analogous to visual Ternus display (Ternus 1926) (Fig. 1). In a typical Ternus display, apparent motion is produced by presenting two sequential visual frames. Here, each frame consists of two horizontal dots. When overlaid, the two frames share one common dot at the center. When the spatial configuration is fixed, observers typically report two distinct perceptions dependent on the inter-stimulus interval (ISI). These are known as 'element motion' (EM) and 'group motion' (GM). Short ISIs usually give rise to the perception of EM. In other words, the outer dots are perceived as moving, while the center dot appears to remain static or flashing. In contrast, long ISIs give rise to the perception of GM. In other words, the two dots are perceived as moving together as a group (Kramer and Yantis 1997; Pantle and Picciano 1976; Pantle and Petersik 1980). Spatial grouping (i.e., within-frame grouping) and temporal grouping (i.e., across-frame grouping), facilitating GM and EM, respectively, have been the dominant theories underlying the mutually exclusive perception of Ternus motion (Aydın et al. 2011; He and Ooi 1999; Kramer and Yantis 1997; Petersik and Rice 2008; Scott-Samuel and Hess 2001; Wallace and Scott-Samuel 2007). We took the visual elements for auditory units ('white noise' or 'tone'), played by three speakers, with nearly the same configurations (except for minor temporal disparities in the within-frame elements, see "Methods") to compose the auditory Ternus display.

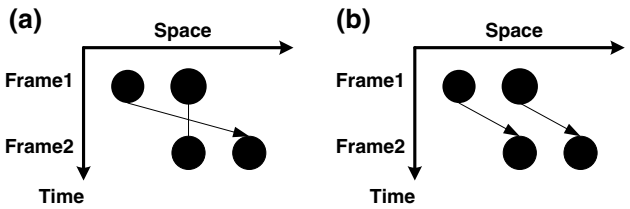

Fig. 1 The Ternus display. Two possible motion perceptions: **a** EM for short ISIs with the *middle disk* perceived to remain static, while the *outer disk* is perceived as moving from one side to the other. **b** GM for long ISIs, with two disks perceived to be moving together as a group

Therefore, the auditory Ternus display can be viewed as a simplified demonstration shown through a presentation of a short sequence of tones. Here, characteristics such as inter-tone intervals and other spectral features, such as frequency differentiation, can be manipulated in order to investigate the roles of spatiotemporal and spectral cues in segregating sound events. Furthermore, the observations of perceptual groupings in an auditory Ternus display (if existing robustly) would extend the general governing laws of perceptual grouping, as shown in the visual and tactile Ternus displays (Chen et al. 2010; Harrar and Harris 2007).

Specifically, we conducted two experiments. Experiment 1 examined the perception of auditory apparent motion as a function of the inter-frame interval (IFI). In addition, we varied the spatial distance of the speakers to examine the influence of spatial location. Experiment 2 investigated the role of auditory frequencies in modulating the perception of auditory Ternus motion.

Experiment 1

Experiment 1 was carried out mainly to inspect the role of temporal grouping in segregating sound events in short auditory sequences. Two sub-experiments were implemented in Experiment 1: Experiment 1a (center-to-center distance for speakers 45 cm) and Experiment 1b (center-to-center distance for speakers 25 cm), in which all the stimulus configurations were identical, except for the spatial layout of the speakers. Hence, we can also observe the effect of the spatial location.

Methods

Participants

Fourteen undergraduate and graduate students (four females, aged between 20 and 30; average age 24.9 years) participated in Experiment 1a. Twelve undergraduate and

graduate students (four females, aged between 21 and 25; average age 23.1 years) participated in Experiment 1b. All participants reported having normal hearing and were naïve to the purposes of the study. The experiment was performed in compliance with all institutional guidelines set by the Academic Affairs Committee, Department of Psychology at Peking University.

Apparatus and Stimuli

Three hamburger mini-speakers (DK-601, diameter 3.6 cm) were placed horizontally on a desk (see Fig. 2a). The center-to-center distances between the speakers were set at 45 cm in Experiment 1a and 25 cm in Experiment 1b. A monitor was placed behind the speakers. A normal PC—interfaced with a sound card (RME Fireface UFX)—was used for all stimuli presentation, instruction presentation (with 17-inch CRT monitor) and data collection (by key-press). The computer program used to control the experiment was developed with Matlab (Mathworks Inc.) and the Psychophysics Toolbox (Brainard 1997; Pelli 1997). The test cabin was semi-anechoic. No light was present, except that which was emitted by the monitor. The viewing distance was set at 70 cm.

The auditory stimuli consisted of four sequentially presented, identical 50-ms burst of white noise (65 dB) to generate auditory apparent motion. The initial noise was provided by the first (flanker) speaker. The second and third noises were generated by the middle speaker. The fourth noise was emitted from the third (flanker) speaker. The first two and final two sounds were treated as two frames. The IFI was the interval between the offset of the second tone in the first frame and the onset of the first tone in the second frame (see Fig. 2b). The IFI was chosen between 50, 80, 110, 140, 170, 200 and 230 ms on a trial-by-trial basis. This was similar to the settings in a visual or tactile display (Chen et al. 2010; Harrar and Harris 2007; Shi et al. 2010). A small, within-frame interval (WFI, 5, 10 and 20 ms) was manipulated to avoid participants' hearing only one sound stemming from the middle position between the two speakers, rather than two successive sounds, i.e., the precedence effect Litovsky et al. (1999). To prevent the abrupt onset and offset of the sounds, the auditory stream was preceded and followed by empty intervals of 50 ms, in addition to the 5-ms ramp time (Fig. 2).

Design and procedures

Prior to the experiment, participants were shown demonstrations of EM (auditory apparent motion from sound A to B to C) and GM (auditory apparent motion from sound 'AB' to 'BC') (with all three WFIs and the smallest and largest IFI conditions included). They then practiced by performing a series of trials. All participants reported clear discriminations between EM and GM with a correct response rate above 90 %. A 3 (WFI 5, 10 and 20 ms) × 7 (IFI 50, 80, 110, 140, 170, 200 and 230 ms) block design was adopted. Each configuration was presented 40 times, with the directions of apparent motion (left or right) balanced evenly across the 40 trials. Therefore, each experiment had a total of 840 trials, all of which were divided into 10 blocks. A typical trial went as follows: Participants were instructed to keep their eyes on the monitor and pay attention to the auditory stimuli. The inter-trial interval (ITI) was randomly selected at between 500 and 700 ms on a trial-by-trial basis. After the auditory stream (Fig. 2b) finished, with a random pause of 300–500 ms, participants were presented with a question mark, which

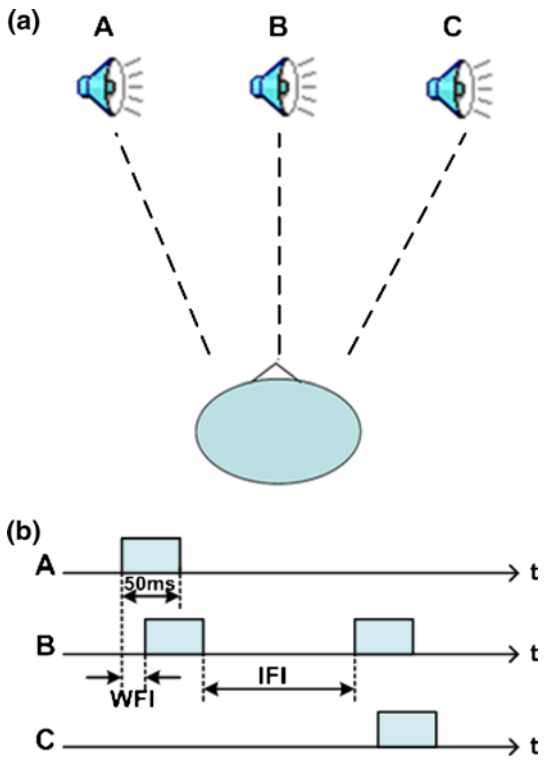

Fig. 2 Experimental setup and temporal correspondence of motion streams used in Experiment 1 and Experiment 2. **a** Three speakers were placed *horizontally* with a center-to-center distance of 45 cm in Experiment 1a and 25 cm in Experiment 1b. Participants sat in front of the middle speaker with a viewing distance of 70 cm and were asked to judge their perception of auditory Ternus motion ('EM' or 'GM'). **b** Temporal correspondences of three sounds: here, sound A, B and C composed two auditory frames ('AB' and 'BC') with a within-frame interval (WFI, interval between A and B, and between B and C) of 5, 10 or 20 ms for Experiment 1, and 5 or 20 ms for Experiment 2. The inter-frame interval (IFI, the interval between the two frames 'AB' and 'BC') was selected from between 50–230 ms for Experiment 1 and 30–210 ms for Experiment 2. The frequencies used for the sounds in Experiment 2 were 800 Hz for a standard frame and 820, 860 or 1,000 Hz for a comparative frame. The direction of the auditory motion (*left* or *right*) was randomized and counterbalanced

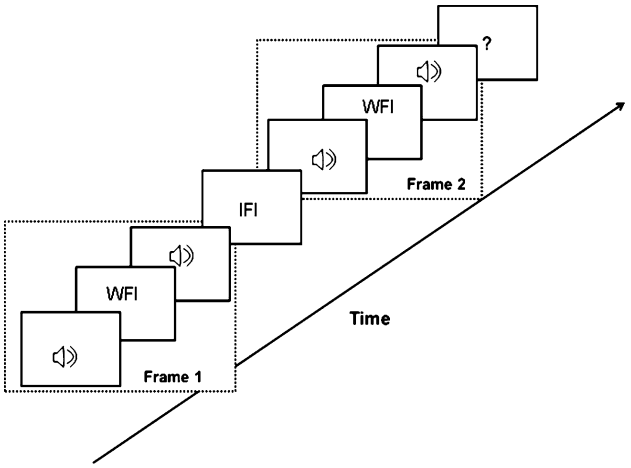

Fig. 3 Schematic illustration of the events presented on one trial in both Experiment 1 and Experiment 2. The auditory stimuli were composed of two sound frames (white noise or tone) separated by a within-frame interval (WFI) and an inter-frame interval (IFI). After the auditory stimuli, a question mark was presented to prompt participants to make a choice between two options (*left* or *right* keypress)

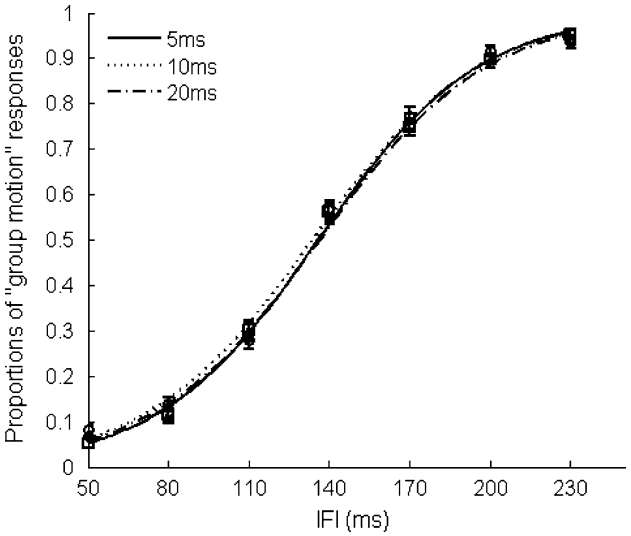

Fig. 4 The average psychometric curves for all participants under the three within-frame interval conditions in Experiment 1a: the curves represent the proportions of group motion responses as a function of the IFI between the two auditory frames. The *solid curve* shows the proportion of group motion for a 5-ms within-frame interval condition. The *dash curve* represents the proportion of group motion for a 10-ms within-frame interval condition. The *dash–dot curve* illustrates the proportion of group motion for a 20-ms within-frame interval condition. The *error bars* represent the associated standard errors

was an indication that they should respond by making a choice (left or right keypress). Their choice would indicate whether they had perceived either element or GM. For half of the participants, a left key corresponded to 'GM' and a right key to 'EM.' The opposite setup was established for the other half of the participants (Fig. 3).

Results and discussion

For each condition, the transition threshold is the point at which EM and GM were reported with equal frequency. Transition threshold is also referred to as the point of subjective equality (PSE), which is calculated by estimating the 50 % performance point on the fitted logistic function. The just noticeable difference (JND) is the difference between the two motion perceptions obtained from the psychometric curve by estimating the IFI difference between 50 and 75 % of the GM responses (Treutwein and Strasburger 1999). Figure 4 shows the average results from all participants in Experiment 1a. As summarized in Fig. 5, the PSE values were 136.0 (SE 4.8) ms, 133.4 (SE 5.6) ms and 136.6 (SE 5.5) ms when WFIs of 5, 10 and 20 ms, respectively, were applied in Experiment 1a. In Experiment 1b, the PSE values were 133.2 (SE 3.9) ms, 128.8 (SE 2.8) ms and 129.8 (SE 3.6) ms under conditions using WFIs of 5, 10 and 20 ms, respectively. As shown in Fig. 6, the corresponding JNDs were 30.7(SE 3.3) ms, 32 (SE 3.9) ms and 31.5 (SE 2.9) ms in Experiment 1a, and 29.3 (SE 2.9) ms, 27.2 (SE 3.2) ms and 29.0 (SE 2.8) ms in Experiment 1b. A repeated-measures analysis of variance

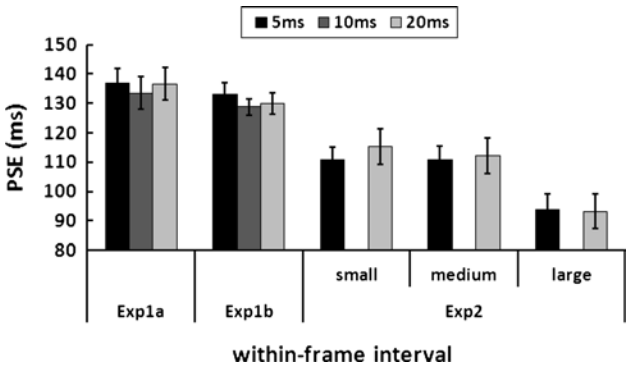

Fig. 5 The mean PSEs for discriminating 'EM' and 'GM' under three within-frame intervals (5, 10 and 20 ms in Experiment 1) and two within-frame intervals (5 and 20 ms in Experiment 2): The *black bars* represent the PSEs for a within-frame interval of 5 ms. The *dark gray bars* represent a within-frame interval of 10 ms, and the *light gray bars* represent a within-frame interval of 20 ms. 'Small' signifies two frames with a very small difference in frequency (800 vs. 820 Hz). 'Medium' represents middle-level frequency differences (800 vs. 860 Hz), and 'large' represents a larger frequency disparity (800 vs. 1,000 Hz). The *error bars* represent standard errors

(ANOVA) of the estimated PSEs—with WFIs of 5, 10 and 20 ms—as factors revealed the main effect of the WFI to be insignificant, $F(2,26) = 0.378$, $p = 0.689$ (Experiment 1a), $F(2,22) = 1.196$, $p = 0.321$ (Experiment 1b).

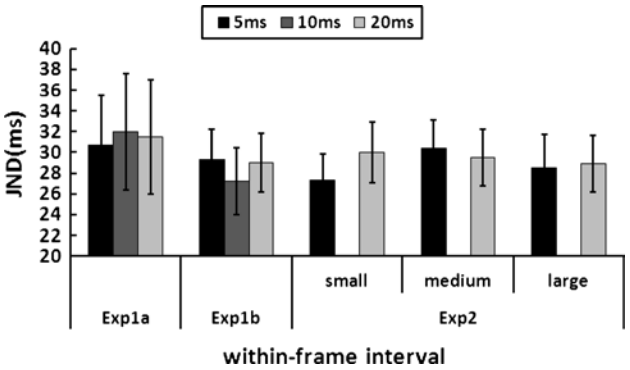

Fig. 6 The mean JNDs for discriminating 'EM' and 'GM' under three within-frame intervals (5, 10 and 20 ms) in Experiment 1a and Experiment 1b, and two within-frame intervals (5 and 20 ms) in Experiment 2. The connotations of the labels are the same as in Fig. 5

The ANOVA of the estimated JNDs, with WFIs of 5, 10 and 20 ms, also revealed the main effect to be insignificant—$F(2,26) = 0.324$, $p = 0.726$ (Experiment 1a), $F(2,22) = 0.321$, $p = 0.729$ (Experiment 1b).

A repeated-measures ANOVA toward the percentages of GM perception (using WFI and IFI as the two within-participant independent factors) revealed a significant main effect of IFI in both Experiment 1a—$F(6,78) = 230.875$, $p < 0.001$ and Experiment 1b—$F(6,66) = 500.326$, $p < 0.001$. Nevertheless, no significant main effect for the WFI was observed—$F(2,26) = 0.257$, $p = 0.775$ (Experiment 1a), $F(2,22) = 1.148$, $p = 0.336$ (Experiment 1b). Furthermore, no significant effect on the interaction between the WFI and IFI was observed—$F(12,156) = 0.387$, $p = 0.967$ (Experiment 1a), $F(2,26) = 1.362$, $p = 0.192$ (Experiment 1b).

We then performed a cross-experiment analysis to discover the effects of the spatial layout, if any. A Univariate ANOVA was carried out for PSE with WFI and spatial layout (45 cm in Experiment 1a and 25 cm in Experiment 1b) as dependent factors. Importantly, the analysis results revealed no significant effect of spatial layout, $F(1,72) = 1.557$, $p = 0.216$. The effect of WFI was insignificant, $F(2,72) = 0.227$, $p = 0.759$, and no significant interaction between spatial layout and WFI was found, $F(2,78) = 0.091$, $p = 0.913$. The cross-experiment analysis for JND likewise yielded no statistical differences.

The results showed that, similar to visual and tactile Ternus, the perception of auditory Ternus motion was mainly modulated by the IFIs. The perception of 'GM' was dominant under longer IFIs conditions. The distinction between 'EM' and 'GM' in auditory Ternus was generally based on the principles of temporal grouping. Here, the longer IFIs made the temporal boundary of two auditory frames ('AB' and 'BC') distinctive. A longer IFI also enhanced the perceived separation of the two grouped auditory events. This led to a dominant perception of GM. In a precedence effect (Hartung and Trahiotis 2001), the lagging sound might fuse to the leading sound when both are in short temporal separations (less than 10 ms). Here, we adopted three within-frame delays (WFIs of 5, 10 and 20 ms as brief gaps) and found that the WFI imposed no discernible influence on the participants' ability to discriminate and perceive apparent motion. The spatial grouping within the current setting had no modulating effect on perceptual classification in short auditory sequences. In general, we managed to replicate the Ternus motion in the auditory domain by using similar IFI settings, as in both the visual and tactile domains.

Experiment 2

Experiment 2 was designed to investigate the role of spectral cues in separating sound events in a short auditory sequence. The same auditory Ternus setting from Experiment 1a was employed in Experiment 2, except for the fact that the IFI range was set from between 30 and 210 ms, because we observed near-ceiling effects when IFI = 230 ms (Experiment 1).

Methods

Participants

Thirteen undergraduate and graduate students (four females, aged between 19 and 29; average age 23.5 years) from Peking University participated in Experiment 2. All of them reported having normal hearing and were naïve to the purposes of the study.

Stimuli and procedure

The stimuli and experimental settings in Experiment 2 generally remained the same as in Experiment 1a. However, the following changes were made: The auditory stimuli was a white noise (as in Experiment 1) as a carrier, with the addition of pure tones with different frequencies, and the IFI range was set from 30 to 210 ms. Of the two auditory frames, one frame contained two tones with a standard frequency of 800 Hz. The other frame consisted of a tone pair with comparative frequencies of 820, 860 and 1,000 Hz. The frequencies were selected in order to simulate the conditions present when the peripheral tonotopic channels were (1) partly overlapped (800 vs. 820 Hz, difficult to separate two frames), (2) medially separated (800 vs. 860 Hz,

relative easy to separate the two frames) and (3) distinctly separated (800 vs. 1,000 Hz, easy to separate two frames).[1]

The amplitude of each tone was set according to the equal-loudness level. This was done because the WFI has little influence on the perception of apparent motion (as concluded from Experiment 1). Only two WFIs (5 and 20 ms) were employed in Experiment 2, in order to reduce the number of trials. In addition, due to the fact that the vast majority of participants had made virtually 100 % GM judgments for long IFIs in the previous experiments, the range of IFIs was adjusted to from 30 to 210 ms, with increased step sizes of 30 ms.

A 2 (WFI) × 7 (IFI) × 3 (frequency separation: low, medium and high) block design was adopted. There were still 840 trials throughout the experiment, which were divided into 5 blocks. The presentation order of the standard frame (auditory pair of 800 Hz) and the comparative frame (auditory pair of 820, 860 and 1,000 Hz), and the directions of apparent motion (left or right) were fully randomized and balanced. The participants received the same amount of practice as in Experiment 1, in order to assure a clear distinction between EM and GM. In the following formal experiment, participants were asked to concentrate on discriminating their perceptions of apparent motion, rather than the pitch differences between the two frames. The data collection method was the same as in Experiment 1.

Results and discussion

A repeated-measures ANOVA toward the percentages of GM perception (using WFI and IFI as the two within-participant independent factors) revealed a significant main effect of IFI, $F(6,72) = 330.67$, $p < 0.001$.

A 2 × 3 ANOVA was conducted with WFI (5 vs. 20 ms) and frequency separation (small vs. medium vs. large) as within-subject independent factors and PSE as dependent factor revealed nonsignificance of the main effect of WFI, $F(1,12) = 0.249$, $p = 0.627$. The effect of frequency separation was significant, $F(2,24) = 13.378$, $p < 0.001$. For both WFI conditions, Bonferroni-corrected comparisons showed that the PSE at a higher frequency (1,000 Hz) was lower than the PSE at a lower frequency (820 Hz) ($p < 0.05$) and lower than the PSE at the medial frequency (860 Hz) ($p < 0.01$). However, the interaction between the WFI and frequency was insignificant, $F(2,24) = 1.341$, $p = 0.280$. Therefore, an obvious decrease in the PSE was observed in the frequency separation between two auditory frames. This was especially true under higher frequency conditions. However, the within-experimental analysis showed that the facilitation of separating the auditory events by using spectral cues was limited to only a certain frequency range (such as approximately 1,000 Hz in the current setting).

ANOVA toward JNDs with WFI, frequency separation as two within-subject independent factors, revealed the insignificance of the main effect of the WFI, $F(1,12) = 0.554$, $p = 0.471$. The main effect of frequency separation was not significant, $F(2,24) = 0.521$, $p = 0.600$. The interaction of WFI and frequency separation was also not significant, $F(2,24) = 0.736$, $p = 0.490$.

Cross-experimental analysis

We separate the data for 5 and 20 ms WFIs in Experiment 2. We then performed a cross-experiment analysis. A Univariate ANOVA was carried out for PSE with a WFI (5 and 20 ms, the data for different frequencies were averaged in Experiment 2) and experiments (Experiment 1a and Experiment 2) as dependent factors. The analysis revealed that the PSEs (mean 106 ± 3.6) in Experiment 2 were significantly reduced compared to those (136.3 ± 3.5) in Experiment 1a, $F(1,54) = 36.44$, $p < 0.001$. The averaged PSEs were 120.6 ± 3.5 (ms) for a 5 ms WFI and 121.8 ± 3.5 (ms) for a 20 ms WFI, $F(1,54) = 0.058$, $p = 0.811$. The interaction between the WFI and experiments was insignificant, $F(1,54) = 0.012$, $p = 0.911$. Likewise, we performed a cross-experimental analysis of the JNDs. The main effect of WFI was not significant, $F(2,24) = 0.119$, $p = 0.732$. The main effect of experiment (Experiment 1a vs. Experiment 2) was not significant, $F(2,24) = 0.751$, $p = 0.390$. The interaction between the WFI and experiments was insignificant, $F(1,54) = 0.004$, $p = 0.950$. Therefore, with the same temporal configurations, the larger separations of auditory frequencies led to a significant segregation of short auditory sequences, as observed with the more dominant perception of 'GM' in the auditory Ternus display.

General discussion

A large body of research has detailed the fact that stimulus factors assist in auditory grouping and auditory segregation, i.e., 'ASA' (Bregman 1990). Auditory grouping occurs on the basis of frequency similarities and spectral

[1] In order to confirm that the frequencies selected justified our research purposes, we asked 10 participants to do a pitch discrimination task. Two frames of pure tones were presented, and the frequencies of the reference frame were kept at 800 Hz, while the comparative frame had a frequency selected from 660, 700, 740, 780, 820 Hz, 860, 900 and 940 Hz. ANOVA with the frequency as the single independent factor showed a significant frequency effect, $F(2,16) = 20.201$, $p < 0.001$. Bonferroni-corrected pairwise comparisons for 820, 860 and 940 Hz conditions confirmed that performance was better at 940 Hz (90.4 %) than at 860 Hz (83.8 %) ($p < 0.05$), better at 860 Hz (83.8 %) than at 820 Hz (74.3 %) ($p < 0.01$) and better at 940 Hz (90.4 %) than at 820 Hz (74.3 %) ($p < 0.01$).

continuity (Bregman and Campbell 1971; Bregman 1990). Segregation is also aided when auditory objects differ in their spectral content or temporal structure, such as occurs with repetition rate (Perrott 1984; Stellmack 1994). Theories have been proposed to account for how temporal cues and spectral cues contribute to auditory streaming. An influential theory, known as the 'Peripheral Channeling Hypothesis,' suggests that streaming is primarily based on stimulus processing occurring in the auditory periphery (Beauvois and Meddis 1996; Hartmann and Johnson 1991; Van Noorden 1975).

The 'peripheral channeling' theory supposes that stream segregation happens when stimuli excite distinct—or not highly overlapped—cochlear filters or peripheral tonotopic channels (Hartmann and Johnson 1991; van Noorden 1975). Therefore, the theory contends that consecutive sounds will be perceptually grouped into a single stream when they activate these strongly overlapped, central auditory neurons. However, separate streams will be perceived if they correspond to the separated tonotopic channels within the auditory system (Carlyon 2004), thus providing higher levels of the nervous system with clear evidence of clearly distinguishable sound sources (but see Vliegen and Oxenham 1999 and Grimault et al. 2002, the stream segregation perception was elicited even with the same cochlear channels).

By using an auditory Ternus display, we replicated the visual Ternus-like apparent motion in the auditory domain. The auditory Ternus used here incorporates a simple and clear-cut perceptual decision task, which can be used to index stream segregation without the subject having to make any explicit judgments relating to their streaming perceptions. In the study, we illustrated the spatiotemporal cues and spectral cues used to separate auditory events. In Experiment 1a, the stimuli had no different spectral features, except for the IFI between the elementary auditory stimuli. This provided a good starting point from which to test the role of temporal cues in auditory segregation and grouping. For temporal cues, the localization/perceptual grouping of sound was generally achieved by combining the information from the two ears in the form of inter-aural time differences (ITDs) and inter-aural level differences (ILDs) (Blauert 1997). With the symmetrical layout of the two flanker speakers (regardless of whether the inter-distance was near or far), the ITDs should be equal in both 'near' and 'far' conditions, and the spatial factor plays little role in auditory grouping (Lakatos and Shepard 1997). Comparatively, the temporal distance (IFIs) between auditory Ternus frames was more important in discerning the perception of apparent motion. The significant main effect of the IFIs was manifested in the role of temporal intervals on the perception of auditory apparent motion. The larger IFIs gave rise to more observable separation between auditory events that led to the dominant perception of 'GM.'

Experiment 2 (with both temporal cues and spectral cues) revealed similar results in terms of the influence of IFIs. The longer the IFIs between two auditory frames, the more dominant the perception of 'GM' became. More importantly, cross-experimental analysis demonstrated the significant modulation effect of frequency cues. The PSEs were generally reduced through the introduction of frequency cues, and the trend was most obvious at higher frequencies. This modulation effect was probably due to differentiation in *auditory critical bandwidth* and *frequency-to-place mapping*.

Critical bandwidth is the frequency bandwidth of the auditory filter created by the cochlea. Here, a second tone will interfere with the perception of the first tone by means of auditory masking. This occurs when auditory frequencies are of a similar range (Moore and Glasberg 1987). With the three comparative frequencies given (820, 860 and 1,000 Hz), the critical bandwidth around 800 Hz (standard stimuli) was between 740 and 860 Hz (Fastl and Zwicker 1999). This covered the lower and medium frequencies we selected for Experiment 2. In addition, evidence has now established frequency selectivity within the auditory periphery (Forrest and Formby, 1996; Hartmann and Johnson 1991; Heinz et al. 1996). When two marker frequencies (of the auditory Ternus frames) are the same, or even when they are very similar, the markers stimulate the same region of the cochlear partition. In turn, this leads to responses from the same auditory nerve fibers (and hence the corresponding 'EM'). When the two markers have larger differences in their frequencies, they are separated in the cochlea, so that they maximally stimulate different places along the cochlear partition. This leads to different populations of auditory nerve fibers responding to each frequency, and hence to the corresponding 'GM' (Oxenham 2000; Vliegen et al. 1999).

The perceptions of GM and EM observed in auditory modality were also found in visual and tactile modalities (Chen et al. 2010; Harrar and Harris 2007; Shi et al. 2010). Previous studies have rigorously investigated the perceptual grouping between auditory and tactile modalities (Chen et al. 2011; Spence et al. 2007), auditory and visual modalities (Sanabria et al. 2005a, 2005b; Shi et al. 2010), and visual and tactile modalities (Harrar and Harris 2007). They have suggested a supra-modal perceptual grouping among different sensory modalities. The current results strengthen this 'supra-modal' view of perceptual groupings.

It should be noted that the tone sequence used in the auditory Ternus display lasted less than one second. This is quite different from the classical alternating tone sequences, which are always repeated for several seconds or even for minutes, and the buildup of the strength of

stream segregation, which takes from 5 to10 s. The quick perceptual decisions for short tone sequences were seemingly at odds with Bregman's proposal that auditory scene analyses start at the same coherent position and are subsequently segregated into separate streams after a sufficient number of cues are collected. The current findings therefore suggest that bi-stable perception could be both an active exploration of the sensory environment and a fundamental aspect of sensory cognition, which supports flexible decision making (Kim et al. 2006). Considerable studies have been conducted to explore neural mechanisms' mediating of auditory stream segregation (Gutschalk et al. 2005; Micheyl et al. 2007; Rauschecker 2005). Recent research has indicated an important role for both primary (A1) and non-primary auditory cortexes, and one study has suggested a role for the intra-parietal sulcus (Cusack 2005). Using an ABA-stimulus paradigm, Cusack (2005) found that regions in the intra-parietal sulcus (IPS) showed greater activity when two streams were perceived rather than one stream. Indeed, the auditory system contains several subcortical nuclei, which are generally believed to establish basic feature encoding even before perceptual organization starts at the cortical level (Griffiths and Warren 2002; Nelken 2004). Using an ABA-stimulus paradigm, Pressnitzer et al. (2008) found that ASA starts much earlier in the auditory pathways, by recording single units from one peripheral structure of the mammalian auditory brainstem, the cochlear nucleus. Peripheral responses were similar to cortical responses and displayed all of the functional properties required for streaming. During the presentation of long auditory sequences, adaptation in peripheral auditory neurons may also be influenced by the descending feedback from upper processing stages, including the auditory cortex. However, at present, the explorations of neural substrates that correspond to the roles of temporal and spectral cues (differential frequencies) and the temporal courses for perceptual grouping in short auditory sequences are lacking and await future investigations. (Getzmann and Lewald 2012; Getzmann 2011; Hall et al. 2002).

Acknowledgments This study was supported by grants from the Natural Science Foundation of China (11174316, 31200760), National High Technology Research and Development Program of China (863 Program) (2012AA011602) and Strategic Priority Research Program of the Chinese Academy of Sciences (XDA06020201).

References

Anstis S, Saida S (1985) Adaptation to auditory streaming of frequency modulated tones. J Exp Psychol Hum Percept Perform 11(3):257–271

Aydın M, Herzog MH, Öğmen H (2011) Attention modulates spatiotemporal grouping. Vis Res 51:435–446

Beauvois M (1998) The effect of tone duration on auditory stream formation. Percept Psychophys 60:852–861

Beauvois MW, Meddis R (1991) A computer model of auditory stream segregation. Q J Exp Psychol A 43:517–541

Beauvois MW, Meddis R (1996) Computer simulation of auditory stream segregation in alternating-tone sequences. J Acoust Soc Am 99:2270–2280

Bee MA, Klump GM (2005) Auditory stream segregation in the songbird forebrain: effects of time intervals on responses to interleaved tone sequences. Brain Behav Evol 66:197–214

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

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

Bregman AS (1978) Auditory streaming is cumulative. J Exp Psychol Hum Percept Perform 4:380–387

Bregman AS (1990) Auditory scene analysis. MIT Press, Cambridge

Bregman AS, Campbell J (1971) Primary auditory stream segregation and perception of order in rapid sequences of tones. J Exp Psychol 89:244–249

Carlyon RP (2004) How the brain separates sounds. Trends Cogn Sci 8:465–471

Chen L, Shi Z, Müller HJ (2010) Influences of intra- and crossmodal grouping on visual and tactile Ternus apparent motion. Brain Res 1:152–162

Chen L, Shi Z, Müller HJ (2011) Interaction of perceptual grouping and crossmodal temporal capture in tactile apparent-motion. PLoS ONE 6:e17130

Cherry EC (1953) Some experiments on the recognition of speech, with one and with two ears. J Acoust Soc Am 25:975–979

Cooper HR, Roberts B (2007) Auditory stream segregation of tone sequences in cochlear implant listeners. Hear Res 225:11–24

Cusack R (2005) The intraparietal sulcus and perceptual organization. J Cogn Neurosci 17:641–651

Cusack R, Carlyon RP (2004) Auditory perceptual organization inside and outside the laboratory. In: Neuhoff JG (ed) Ecological psychoacoustics. Emerald Group Publishing Limited, Bingley, pp 15–48

Darwin CJ, Carlyon RP (1995) Auditory grouping. In: Moore BCJ (ed) Hearing, vol 6. Academic, Orlando, pp 387–424

Denham SL, Winkler I (2006) The role of predictive models in the formation of auditory streams. J Physiol Paris 100:154–170

Eggermont JJ (2001) Between sound and perception: reviewing the search for a neural code. Hear Res 157:1–42

Fastl H, Zwicker E (1999) Psychoacoustics: facts and models. Springer, New York, Incorporated

Fishman YI, Reser DH, Arezzo JC, Steinschneider M (2001) Neural correlates of auditory stream segregation in primary auditory cortex of the awake monkey. Hear Res 151:167–187

Forrest TG, Formby C (1996) Temporal gap detection thresholds in sinusoidal markers simulated with a single-channel envelope detector. Aud Neurosci 3:21–33

Füllgrabe C, Moore BCJ (2012) Objective and subjective measures of pure-tone stream segregation based on interaural time differences. Hear Res 291:24–33

Getzmann S (2011) Auditory motion perception: onset position and motion direction are encoded in discrete processing stage. Eur J Neurosci 33:1339–1350

Getzmann S, Lewald J (2012) Cortical processing of changes in sound location: smooth motion versus discontinuous displacement. Brain Res 1466:119–127

Griffiths TD, Warren JD (2002) The planum temporale as a computational hub. Trends Neurosci 25:348–353

Grimault N, Bacon SP, Micheyl C (2002) Auditory stream segregation on the basis of amplitude-modulation rate. J Acoust Soc Am 111:1340–1348

Gutschalk A, Micheyl C, Melcher JR, Rupp A, Scherg M, Oxenham AJ (2005) Neuromagnetic correlates of streaming in human auditory cortex. J Neurosci 25:5382–5388

Hall DA, Johnsrude IS, Haggard MP, Palmer AR, Akeroyd MA, Summerfield AQ (2002) Spectral and temporal processing in human auditory cortex. Cereb Cortex 12:140–149

Harrar V, Harris LR (2007) Multimodal Ternus: visual, tactile, and visuo-tactile grouping in apparent motion. Perception 36:1455–1464

Hartmann WM, Johnson D (1991) Stream segregation and peripheral channeling. Music Percep 9:155–183

Hartung K, Trahiotis C (2001) Peripheral auditory processing and investigations of the "precedence effect" which utilize successive transient stimuli[J]. J Acoust Soc Am 110:1505

He ZJ, Ooi TL (1999) Perceptual organization of apparent motion in the Ternus display. Perception 28:877–892

Heinz MG, Goldstein MH, Formby C (1996) Temporal gap detection thresholds in sinusoidal markers simulated with a multi-channel, multi-resolution cochlear model. Aud Neurosci 3:35–56

Kim YJ, Grabowecky M, Suzuki S (2006) Stochastic resonance in binocular rivalry. Vis Res 46:392–406

Kondo HM, Kitagawa N, Kitamura MS, Koizumi A, Nomura M, Kashino M (2012) Separability and commonality of auditory and visual bistable perception. Cereb Cortex 22(8):1915–1922

Kramer P, Yantis S (1997) Perceptual grouping in space and time: evidence from the Ternus display. Percept Psychophys 59:87–99

Lakatos S, Shepard RN (1997) Constraints common to apparent motion in visual, tactile, and auditory space. J Exp Psychol Hum Percept Perform 23:1050–1060

Litovsky RY, Colburn HS, Yost WA, Guzman SJ (1999) The precedence effect. J Acoust Soc Am 106:1633–1654

McCabe SL, Denham MJ (1997) A model of auditory streaming. J Acoust Soc Am 101:1611–1621

Micheyl C, Carlyon RP, Gutschalk A, Melcher JR, Oxenham AJ, Rauschecker JP, Tian B, Wilson EC (2007) The role of auditory cortex in the formation of auditory streams. Hear Res 229:116–131

Moore BCJ, Glasberg BR (1987) Formulae describing frequency selectivity as a function of frequency and level, and their use in calculating excitation patterns. Hear Res 28:209–225

Nelken I (2004) Processing of complex stimuli and natural scenes in the auditory cortex. Curr Opin Neurobiol 14:474–480

Oxenham AJ (2000) Influence of spatial and temporal coding on auditory gap detection. J Acoust Soc Am 107:2215–2223

Pantle AJ, Petersik JT (1980) Effects of spatial parameters on the perceptual organization of a bistable motion display. Percept Psychophys 27:307–312

Pantle AJ, Picciano L (1976) A multistable movement display: evidence for two separate motion systems in human vision. Science 193(4252):500–502

Pelli DG (1997) The VideoToolbox software for visual psychophysics: transforming numbers into movies. Spat Vis 10(4):437–442

Perrott DR (1984) Concurrent minimum audible angle: a reexamination of the concept of auditory spatial acuity. J Acoust Soc Am 75:1201–1206

Petersik JT, Rice CM (2008) Spatial correspondence and relation correspondence: grouping factors that influence perception of the Ternus display. Perception 37:725–739

Pressnitzer D, Hupé J-M (2006) Temporal dynamics of auditory and visual bistability reveal common principles of perceptual organization. Curr Biol 16:1351–1357

Pressnitzer D, Sayles M, Micheyl C, Winter IM (2008) Perceptual organization of sound begins in the auditory periphery. Curr Biol 18:1124–1128

Rauschecker JP (2005) Neural encoding and retrieval of sound sequences. Ann N Y Acad Sci 1060:125–135

Sanabria D, Soto-Faraco S, Chan J, Spence C (2005a) Intramodal perceptual grouping modulates multisensory integration: evidence from the crossmodal dynamic capture task. Neurosci Lett 377:59–64

Sanabria D, Soto-Faraco S, Spence C (2005b) Assessing the effect of visual and tactile distractors on the perception of auditory apparent motion. Exp Brain Res 166:548–558

Scott-Samuel NE, Hess RF (2001) What does the Ternus display tell us about motion processing in human vision? Perception 30:1179–1188

Shamma SA, Micheyl C (2010) Behind the scenes of auditory perception. Curr Opin Neurobiol 20:361–366

Shamma SA, Elhilali M, Micheyl C (2011) Temporal coherence and attention in auditory scene analysis. Trends in Neurosci 34:114–123

Shi Z, Chen L, Müller HJ (2010) Auditory temporal modulation of the visual Ternus effect: the influence of time interval. Exp Brain Bes 203:723–735

Snyder JS, Alain C (2007) Toward a neurophysiological theory of auditory stream segregation. Psychol Bull 133:780–799

Snyder JS, Alain C, Picton TW (2006) Effects of attention on neuroelectric correlates of auditory stream segregation. J Cogn Neurosci 18:1–13

Spence C, Sanabria D, Soto-Faraco S (2007) Intersensory Gestalten and crossmodal scene perception. In: Noguchi K (ed) Psychology of beauty and Kansei: new horizons of Gestalt perception. Fuzanbo Int, Tokyo, pp 519–579

Stellmack MA (1994) The reduction of binaural interference by the temporal nonoverlap of components. J Acoust Soc Am 96:1465–1470

Szalárdy O, Bőhm TM, Bendixen A, Winkler I (2013) Event-related potential correlates of sound organization: early sensory and late cognitive effects. Biol Psychol 93:97–104

Takegata R, Roggia SM, Winkler I (2005) Effects of temporal grouping on the memory representation of inter-tone relationships. Biol Psychol 68:41–60

Ternus J (1926) Experimentelle Untersuchugen über phänomenable Identität. Psychologishce Forschung 7:81–136

Treutwein B, Strasburger H (1999) Fitting the psychometric function. Percept Psychophys 61:87–106

van Noorden LPAS (1975) Temporal coherence in the perception of tone sequences. University of Technology, Eindhoven

Vliegen J, Oxenham AJ (1999) Sequential stream segregation in the absence of spectral cues. J Acoust Soc Am 105:339–346

Vliegen J, Moore BCJ, Oxenham AJ (1999) The role of spectral and periodicity cues in auditory stream segregation, measured using a temporal discrimination task. J Acoust Soc Am 106:938–945

Wallace JM, Scott-Samuel NE (2007) Spatial versus temporal grouping in a modified Ternus display. Vis Res 47:2353–2366

Winkler I, Van Zuijen TL, Sussman E, Horváth J, Näätänen R (2006) Object representation in the human auditory system. E J Neurosci 24:625–634

Yabe H, Winkler I, Czigler I, Koyama S, Kakigi R, Sutoh T, Hiruma T, Kaneko S (2001) Organizing sound sequences in the human brain: the interplay of auditory streaming and temporal integration. Brain Res 897:222–227