Foreground-background discrimination indicated by event-related brain potentials in a new auditory multistability paradigm

Orsolya Szalárdy[1,2], István Winkler[1,3], Erich Schröger[4], Andreas Widmann[4], Alexandra Bendixen[4,5]

[1] Institute of Cognitive Neuroscience and Psychology, Research Centre for Natural Sciences, Hungarian Academy of Sciences, Hungary

[2]Department of Cognitive Science, Faculty of Natural Sciences, Budapest University of Technology and Economics, Hungary

[3]Institute of Psychology, University of Szeged, Hungary

[4]Institute of Psychology, University of Leipzig, Germany

[5]Department of Psychology, Cluster of Excellence "Hearing4all", European Medical School, Carl von Ossietzky University of Oldenburg, Germany

Running head: Auditory foreground-background discrimination

Address correspondence to:

Alexandra Bendixen

Department of Psychology, Carl von Ossietzky University of Oldenburg

Ammerländer Heerstr. 114-118, D-26111 Oldenburg, Germany

Phone: +49 441 798 4844

Fax: +49 441 798 5522

e-mail: alexandra.bendixen@uni-oldenburg.de

Abstract

For studying multistable auditory perception, we propose a paradigm that evokes integrated or segregated perception of a sound sequence, and permits decomposition of the segregated grouping into foreground and background sounds. The paradigm combines 3-tone pitch patterns with alternating timbres, resulting in a repeating 6-tone structure that can be perceived as rising based on temporal proximity, or as falling based on timbre similarity. Listeners continuously report their percept while EEG is recorded. Results show an ERP modulation starting at ~70 ms after sound onset that can be explained by whether a sound belongs to perceived foreground or background, with no additional effect of integrated vs. segregated grouping. Auditory grouping as indexed by the mismatch negativity did not correspond with reported sound grouping. The paradigm offers a new possibility for investigating effects of conscious perceptual organization on sound processing.

Keywords: auditory bistability, sound grouping, perceptual organization, background inhibition, Wessel effect, auditory streaming, P1, N1, MMN

Introduction

The phenomenon of perceptual bistability, referring to qualitative changes in perception without corresponding change in the stimulus input, has long attracted scientific interest (Leopold, Wilke, Maier, & Logothetis, 2002; Orbach, Ehrlich, & Heath, 1963; for recent overview, see Schwartz, Grimault, Hupé, Moore, & Pressnitzer, 2012). One reason is that bistability permits to investigate processes of perceptual organization as originally brought up by the Gestalt school of psychology (Köhler, 1947). Much of the early research into perceptual bistability was devoted to the visual modality (for reviews, see e.g. Blake & Logothetis, 2002; Leopold & Logothetis, 1999). However, by now it has been established that perceptual bistability can also be observed in the auditory modality and that bistable auditory and visual perceptual phenomena follow similar principles (Kondo et al., 2012; Pressnitzer & Hupé, 2006; Schwartz et al., 2012).

Recording brain responses such as event-related potentials (ERPs) in bi-/multistable stimulus configurations permits investigating processes accompanying the currently experienced perceptual organization without the confounding influence of actual stimulus changes, thus providing insights into the hidden processes of object formation (Gutschalk et al., 2005; Hill, Bishop, Yadav, & Miller, 2011; Snyder, Holder, Weintraub, Carter, & Alain, 2009; Szalárdy, Bőhm, Bendixen, & Winkler, 2013; Winkler, Takegata, & Sussman, 2005). However, the range of multistable phenomena available to auditory researchers is still rather limited (Schwartz et al., 2012). Here we present a new paradigm evoking multistable auditory perception, and apply it to the investigation of foreground/background decomposition of an auditory scene.

The issue of auditory foreground/background decomposition has been raised in several experimental and theoretical papers (Cusack, Deeks, Aikman, & Carlyon, 2004; Elhilali, Xiang, Shamma, & Simon, 2009; Sussman, Bregman, Wang, & Khan, 2005; Winkler, Denham, Mill, Bőhm, & Bendixen, 2012; Winkler, Denham, & Nelken, 2009). There is consensus that when the auditory scene is attentively segregated into two or more different streams, one of these sound streams is perceived in the foreground while the other sounds fall in the background. Somewhat less is known about the extent of processing received by the background sounds (see, however, Alain & Woods, 1993, 1994; Arnott & Alain, 2002; Nager, Teder-

Salejärvi, Kunze, & Münte, 2003; Sussman et al., 2005; Winkler, Teder-Salejärvi, Horváth, Näätänen, & Sussman, 2003). One reason for this is that the vast body of studies investigating the principles of sound organization used versions of the classical auditory streaming paradigm (Moore & Gockel, 2012). In this paradigm, a three-tone pattern of sounds (ABA) is repeated, formed by two types of sounds (A and B) that differ in some physical feature(s). This type of sequence can be heard as one coherent sound stream consisting of all sounds (ABA-ABA-ABA-; the 'Integrated' percept) or as two separate streams, one of which contains only the A sounds (A-A-A-A-A-A-) while the other stream contains only the B sounds (-B---B---B-; the 'Segregated' percept). Prolonged exposure to such sequences, where the A and B tones differ in some feature/s, leads to perception switching back and forth between the different interpretations (Bendixen et al., 2013; Bendixen, Denham, Gyimesi, & Winkler, 2010; Bőhm et al., 2013; Denham, Gyimesi, Stefanics, & Winkler, 2010, 2013; Denham & Winkler, 2006; Pressnitzer & Hupé, 2006; Roberts, Glasberg, & Moore, 2002; Szalárdy, Bendixen, Tóth, Denham, & Winkler, 2013). Typically, experimenters using the auditory streaming paradigm asked participants to mark their perception in a manner as to distinguish between the 'Integrated' and the 'Segregated' percepts, but not to further distinguish between 'Segregated – A sound appearing in the Foreground' and 'Segregated – B sound appearing in the Foreground.' In fact, it might be difficult for participants to make such a distinction because, given the relative simplicity of the stimulus configuration in the 'ABA' paradigm, it is conceivable that the representations of the A and B streams are maintained in parallel, or that rapid switching between them occurs. Although there have been attempts to instruct participants to specifically attend to either the A or the B tones during a 'Segregated' percept (Gutschalk et al., 2005), this procedure might confound percept-dependent processes with attention- and/or task-related effects.

In order to avoid such issues, we designed a stimulus paradigm in which the 'Segregated' percept made the discrimination of foreground and background easy, such that we could rely on the perceptual reports of the participants and eliminate confounds from additional processes. We adapted a design suggested by Wessel (1979), presenting a repeating three-sound pattern of rising pitch (123) combined with an alternation of timbre (A and B), resulting in a repeating six-tone A1B2A3B1A2B3 structure (cf. Figure 1a). This structure

can be perceived as *rising* in pitch – dropping back to the base level after each triplet – based on temporal proximity (A1B2A3B1A2B3…), or as *falling* in pitch (A3-A2-A1-A3-A2-A1… and B3-B2-B1-B3-B2-B1…), based on timbre similarity. The *rising* percept corresponds to 'Integration' of all the sounds, and the *falling* percept corresponds to 'Segregation' – in this case, segregation by timbre in spite of frequency proximity, as has been shown to occur with other paradigms (Cusack & Roberts, 2000; Dolležal, Beutelmann, & Klump, 2012; Grimault, Bacon, & Micheyl, 2002; Iverson, 1995; Singh, 1987; Szalárdy, Bendixen et al., 2013). Importantly, when this sequence is heard as 'Segregated', the impression is hearing one stream with its characteristic timbre in the foreground, and the remaining sounds forming the background. The distinction between foreground and background is subjectively much more salient than the perceptual difference between the two possible foreground-background configurations in the 'ABA' auditory streaming paradigm. Therefore, provided that the paradigm adapted from Wessel (1979) encourages bi-/multistable perception, it might lend itself to the study of auditory scene decomposition including a distinction between the foreground and the background percept. Therefore, in a series of pilot studies, we determined that perception indeed switches back and forth between the different possible organizations when listeners are exposed to relatively long (few minutes) sequences of the A1B2A3B1A2B3 type. Results of the pilot experiments were similar to the behavioral data obtained in the main experiment, and are therefore not reported separately here.

In the main experiment, we combined the new auditory multistability paradigm with measuring the electroencephalogram (EEG) to investigate ERP correlates of auditory scene decomposition. Our measurement approach for the ERPs was twofold. First, we directly measured whether the ERPs elicited by the tones would be modulated by the currently experienced percept. For this purpose, we compared ERPs elicited by physically identical tones when they were perceived as part of a *rising* pattern ('Integrated'), as part of a *falling* pattern in the foreground ('Segregated-Foreground'), or remained in the background while the tones of the other timbre were perceived as *falling* ('Segregated-Background'). With the same instruction, we could thus disentangle the aspects of perceptual organization ('Integrated' vs. 'Segregated')

and of foreground-background decomposition, providing one step further towards a hierarchical decomposition model of an auditory scene (Cusack et al., 2004).

Second, we added an indirect ERP measure to evaluate not only the processing of the individual tones, but also the interpretation of the tone sequence that is represented in the auditory system (Sussman, Ritter, & Vaughan, 1999; Winkler et al., 2009; Winkler, Sussman et al., 2003). This measure was based on the mismatch negativity (MMN), an ERP component indicating that the auditory system detected a regularity violation in a sequence of sounds (Näätänen, Paavilainen, Rinne, & Alho, 2007; Winkler, 2007). To determine whether the auditory system formed a representation linking a given set of sounds appearing in the stimulus sequence, one can set up some regularity connecting these sounds and then insert occasional violations of the regularity into the tone sequence. If the regularity violations elicit the MMN component, one can infer that the regularity must have been extracted (Schröger, 2007). Thus MMN can be used to indicate which sounds have been linked together at the time the deviance was encountered.

In this vein, for testing whether MMN can offer an index of the sound organization in our multistable paradigm, two additional feature regularities were introduced into the A1B2A3B1A2B3 sequence. A duration regularity was set up by assigning a characteristic duration value to each *rising* tone triplet (i.e., 'A1B2A3' or 'B1A2B3'), with the duration value randomly chosen separately for each triplet. Note that this regularity pertains to temporally adjacent tones (cf. Figure 1b). Occasionally, the regularity was violated by the third tone of the triplet having a duration value different from the first two tones. If MMN was elicited by these violations, one could infer that the auditory system formed a representation of the tone sequence based on temporal adjacency, corresponding to the *rising* percept. However, when the tones were linked by timbre similarity, the duration values were changing randomly within each perceived triplet, and thus no MMN should be elicited (cf. Figure 1b). Following the same principle, a location regularity was introduced by assigning a characteristic location value to each of the *falling* tone triplets (i.e., 'A3A2A1' and 'B3B2B1'), with the location value randomly chosen separately for each triplet. Again, this regularity was violated occasionally by the third tone of the triplet having a location value different from the initial two tones. If

MMN was elicited by these violations, one could infer that the auditory system formed a representation of the tones based on timbre similarity, corresponding to the *falling* percept.

Note that this MMN-based testing is indicative of internal representations formed and maintained by the auditory system, which do not necessarily correspond to conscious perception of the tone sequence (see e.g. Takegata et al., 2005; van Zuijen, Simoens, Paavilainen, Näätänen, & Tervaniemi, 2006). When linking the MMN-related representations to the consciously reported perceptual groupings, three alternatives are conceivable on the basis of previous studies. First, it is possible that the auditory system's representation corresponds to the consciously perceived organization of the sounds (Rahne, Böckmann, von Specht, & Sussman, 2007; Sussman, Ritter, & Vaughan, 1998; Winkler, van Zuijen, Sussman, Horváth, & Näätänen, 2006). In this case, we should see each feature regularity violation eliciting MMN only when listeners experience the corresponding percept. This result would suggest that perceptual grouping is constrained by, or alternatively acts upon, the representations available to the deviance detection process reflected by the MMN. Second, it is possible that processes producing the MMN response are bound by the physical features of the tone sequence, independent of the reported perceptual organization (Ross, Tervaniemi, & Näätänen, 1996). In this case, only the regularity carried by physically adjacent tones would be extracted, and thus MMN should be elicited only by violations of the duration regularity. This result would suggest that conscious perception is not based on the grouping mechanisms indexed by MMN in this paradigm. Finally, it is possible that the auditory system maintains all the perceptual interpretations in parallel although only one of them is consciously experienced at each point in time (Horváth, Czigler, Sussman, & Winkler, 2001; Winkler et al., 2012). In this case, we should find MMN for both feature regularity violations independent of the currently experienced percept. This result would suggest that multistable perception reflects a decision between multiple alternative groupings, each based on representations also entering the deviance-detection process indexed by the MMN response.

Materials and Methods

Participants

Twenty-two young healthy volunteers with self-reported normal hearing participated in the experiment (mean age: 25.0 years, SD: 6.2 years, range: 20-45 years; 14 females, 8 males; 21 right-handed, 1 ambidextrous). Two participants were excluded from further analysis due to difficulties with the task: one of them could not be trained to discriminate between the two timbres and therefore never started the main experiment, the other one completed the experiment but was excluded post-hoc due to poor performance in the control sequences (see below). Data of another three participants had to be excluded because their perceptual reports were very unbalanced towards segregation by timbre (the *falling* percept), leaving insufficient (fewer than 10) EEG epochs for deviant trials obtained during the *rising* percept. The mean age of the remaining seventeen participants was 24.7 years. Written informed consent was obtained from each participant according to the Declaration of Helsinki after the experimental procedure was explained to them. Participants received course credit or modest financial compensation for their contribution.

Apparatus and Stimuli

Participants were seated in a sound-attenuated and electrically shielded chamber (IAC 402-A single-walled, Industrial Acoustics Company GmbH, Niederkrüchten, Germany) at the Institute of Psychology, University of Leipzig, Germany. A computer screen was placed in front of them at a distance of ca. 100 cm, displaying a fixation cross during stimulus presentation. Four-minute sequences of tones were presented binaurally via Sennheiser HD 25-1 headphones with a mean level of 60 dB sound pressure level, calibrated with an artificial head (HEAD acoustics HMS III.0) with direction-independent equalization. Participants held a response pad in their hands and responded with their left- and right-hand thumbs as instructed (see below).

Auditory stimuli were generated with MATLAB (Mathworks, http://www.mathworks.com) with a sampling frequency of 48000 Hz. Stimuli were six different complex tones resulting from the combination of two different timbres (A and B, see below) with three different base frequency values (392 Hz, low; 415.3 Hz, middle; and 440 Hz, high; corresponding to 1 semitone difference between adjacent base frequencies). Each of the six stimuli could have any one of three tone duration values (70, 110, or 150 ms; each including a 5 ms onset and 5 ms offset ramp) and any one of three perceived locations (left/middle/right, created by interaural time difference, ITD, values of -500, 0, or +500 micro-seconds). The timbre difference was produced by the

selection and weighting of the complex tones' harmonic partials. Timbre A was composed of 8 pure sinusoidal tones for the base frequency and the 2nd to 8th harmonics, each of them starting in sine phase, and weighted by the same factor (0.125). Timbre B contained 4 pure sinusoidal tones (again starting in sine phase) at the base frequency and the 3rd, 5th, and 7th harmonics, with each harmonic being assigned a specific weighting: 0.35, 0.45, 0.15, and 0.05, respectively. This created the impression of a brighter sound for timbre B than for timbre A. These stimulus parameters were chosen on the basis of pilot studies with the goal that the three possible percepts (see Introduction and below) should appear with approximately equal probability. An audio file for demonstration can be accessed on-line at http://www.uni-leipzig.de/~biocog/bendixen/szalardy_winkler_schroeger_widmann_bendixen_2013_samplesound.mp3.

The arrangement of the timbres, base frequency values, tone durations, and perceived locations followed pre-defined patterns (Figure 1). The two different timbres were strictly alternating (ABABAB…), while the base frequency values followed a repeating three-tone *rising* pattern (low-middle-high; i.e., 123123…). This resulted in a regularly repeating 6-tone structure (A1B2A3B1A2B3…), which was presented with a uniform 200-ms onset-to-onset interval, pre-tested in our pilot experiments to result in ambiguous perception of the sequence. Variation in the duration and location features was added for the purpose of testing percept-dependent MMN elicitation (see Introduction). Pilot studies confirmed that this additional variation in the stimuli did not affect the clarity of the perceptual organizations. Each *rising* triplet ('A1B2A3' or 'B1A2B3') had a common tone duration, with duration chosen randomly for each triplet. Each *falling* triplet ('A3-A2-A1' and 'B3-B2-B1') had a common location, with location chosen randomly for each triplet. Thus the duration and the location values of the tones did not vary within the corresponding triplets (unless the regularity was violated, see below); but both varied between triplets. Both the duration and the location regularities were violated within 12.5% of the triplets. This was achieved by changing the duration/location of the triplet's third tone from the value set up for the first and second tones to one of the other two possible tone durations/locations. Each deviant triplet was preceded by at least two standard triplets. In the beginning of each stimulus block, four standard cycles (containing the full six-tone pattern) were presented. The last deviant in each block was followed by at least one standard cycle.

Additional *control* sequences were introduced to verify that participants were reliably discriminating between timbre A and timbre B, and that they were using the correct response mappings for the various percepts (see below). An unambiguous counterpart was created for each percept. The control for the *rising* percept was generated by using only the A or only the B timbre, with the arrangement otherwise identical to the above description. This resulted in sequences of the type A1A2A3A1A2A3… or B1B2B3B1B2B3…, in which unambiguous percepts of *rising* triplets are created due to the absence of the timbre variation. Control sequences for the *falling* percept were generated by omitting every other tone from the ambiguous sequences, resulting in sequences that were indeed physically *falling* (A3-A2-A1-A3-A2-A1… and B3-B2-B1-B3-B2-B1…). These sequences were not only unambiguously perceived as *falling*, but also permitted a clear distinction between timbre A and timbre B. The duration and location regularities and deviations were also included in the control sequences for the purpose of testing MMN elicitation under unambiguous stimulus conditions. Control sequences were composed by concatenating short trains (range 3 to 7 seconds, mean 5 s) of the four control patterns described above (*rising-timbre A*, *rising-timbre B*, *falling-timbre A*, *falling-timbre B*) in random order and without any breaks. Based on the pilot experiments, the durations of the short trains were chosen to match the time range during which the same percept was expected to be experienced by the listener in the *ambiguous* condition. The time interval during which the listener continuously marks experiencing the same percept is termed "perceptual phase".

Experimental Procedure

The experiment included 14 stimulus blocks overall. The first three and the final three blocks were control sequences, with a duration of four minutes per block. Each of the control blocks contained a summed length of 80 seconds of control-*rising* segments (40 seconds *rising*-timbre A, 40 seconds *rising*-timbre B), 80 seconds of control-*falling-A* segments, and 80 seconds of control-*falling-B* segments. Altogether, the six control blocks included 100 deviants for the duration regularity (based on the *rising* arrangement), and 100 deviants for the location regularity (based on the *falling* arrangement) to be used for the ERP analysis.

The 4th to 11th blocks contained four minutes of the ambiguous sequence followed by a short (15-20 seconds) control sequence, which contained each unambiguous percept exactly once in a randomized order and for a randomly selected duration. The purpose of appending these control sequences to the ambiguous sequences was to check the participants' response accuracy throughout the whole experiment; the appended sequences were only used for analyzing the behavioral responses. In the ambiguous part of the stimulus blocks, altogether 400 duration regularity deviants and 400 location regularity deviants were delivered. A balanced distribution of the three possible percepts (*rising* / *falling-timbre A* / *falling-timbre B*) would thus yield 133.3 deviants of each type encountered during each of the three possible percepts.

During each block, participants were asked to listen to the sound sequence and continuously indicate their percept by depressing the corresponding buttons on the response pad. They were told to press one of the buttons as long as they heard a *falling-A* sequence, another button as long as they heard a *falling-B* sequence, and both buttons at the same time as long as they heard a *rising* sequence. They were instructed not to press any of the buttons if they were unsure or if their percept did not fall into any of the three pre-defined categories. The assignment of the left and right buttons to timbre A and B was counterbalanced across participants. Button states were sampled every 8 ms (125 Hz sampling rate).

Prior to the experiment, the possible percepts were explained to participants with the help of auditory and visual illustrations. The experimenter made sure that participants understood the instructions, and that they were able to discriminate between the two timbres, by presenting 48-second training blocks containing only unambiguous control sequences as many times as needed. Experimental blocks were started after reliable performance was achieved in the training blocks. Between the experimental blocks, participants were given breaks as needed. The overall experiment including electrode application and removal lasted for about four hours.

EEG Recording

EEG was recorded from 34 active electrodes placed on the scalp using Ag/AgCl-electrodes on a BrainVision EEG system (Fp1, AFz, Fp2, F7, F3, Fz, F4, F8, FC5, FC1, FCz, FC2, FC6, T7, C3, Cz, C4, T8, CP5, CP1,

CP2, CP6, P7, P3, Pz, P4, P8, PO9, O1, Oz, O2, PO10 scalp locations, left and right mastoid M1 and M2) according to the international 10-20 system (Chatrian, Lettich, & Nelson, 1985; Jasper, 1958). Sampling rate was 500 Hz. FCz was used as an online reference. An additional electrode was attached to the tip of the nose, to be used for off-line re-referencing. The vertical electrooculogram (VEOG) was recorded between two electrodes attached above and below the left eye, and the horizontal electrooculogram (HEOG) was recorded between two electrodes placed laterally to the left and right outer canthi.

Data Analysis

Behavioral data. For the *ambiguous* sequences, perceptual phases (the time interval during which the participant pressed the same button or button combination) of *rising*, *falling-A* and *falling-B* percepts were extracted from the button presses, separately for each participant. Because there might be some inaccuracy in synchronizing the button press and release movements, perceptual phases shorter than 300 ms were removed from the analysis (Moreno-Bote, Shpiro, Rinzel, & Rubin, 2010). The proportion of each percept was calculated as the percentage of time that the given percept was reported relative to the overall duration of the stimulus block. For each percept, the average duration of all corresponding perceptual phases was calculated per stimulus block. When a given percept was not reported within a stimulus block, the proportion and phase duration was taken to be 0 for that percept within the corresponding block.

For the *control* sequences, the initial 1000 ms after the start of a new segment were discarded to allow for decision and response time. For the remaining time of the segment (2000 to 6000 ms depending on segment length), the proportion of time during which the participant pressed the button associated with the corresponding percept was calculated. As a sign of reliable reporting, a minimum of 85% correct was required for each percept, separately for the average of the control blocks and the average of the control segments appended to the ambiguous sequences. One participant failed to meet this criterion and was therefore excluded from further analysis.

EEG data. The continuous EEG record was re-referenced off-line to the signal recorded at the tip of the nose. EEG data were filtered with a 0.5 to 100 Hz bandpass filter; followed by an automatic eye movement

correction procedure (Gratton, Coles, & Donchin, 1983). After eye movement correction, a lowpass filter of 30 Hz was applied. Subsequently, segments were extracted from the continuous EEG separately for the two types of analyses.

Comparing the processing of the individual tones between the perceptual organizations. For comparing the ERP responses elicited by the same tones between the possible percepts, the EEG recorded in the *ambiguous* sequences was epoched into 1000-ms long segments (from -500 to +500 ms) containing five consecutive tones of the same percept, with the center tone starting at time 0. Epochs were selected so that the center tone was preceded by four other tones and followed by three other tones of the same percept. No baseline correction was applied to avoid any confounds introduced by percept-related modulations of the ERP during the baseline interval. (The results remained the same when ERP amplitudes were measured relative to a 400-ms pre-stimulus baseline, which covered the two preceding tones.) Epochs with *falling-timbre A* and *falling-timbre B* percepts were collapsed into *falling percept* and separated according to whether the center tone was perceived in the foreground or in the background. Note that in the epochs extracted for the *falling percept*, foreground and background tones alternated (while perceiving *falling-A*, the *B* tones fall in the background, and vice versa). Therefore the tone following the center tone always belonged to the opposite percept (foreground or background) than the center tone. In contrast, in the epochs for the *rising* percept, each tone was perceived in the foreground.

Epochs were rejected when the signal range throughout the epoch exceeded 100 μV at any electrode. Artifact-free epochs were averaged separately for the *rising*, *falling-foreground* and the *falling-background* percept. P1 amplitude was measured from the individual averages at Cz in the latency range of 55-85 ms following tone onset. N1 amplitude was measured from the preceding P1 peak (peak-to-peak measurement) in the latency range of 105-135 ms from tone onset to eliminate carry-over effects of the possible percept-related modulation of the preceding P1 response. P1 and N1 amplitudes were entered into repeated-measures analyses of variance (ANOVAs) with the factor Percept (*rising* vs. *falling-foreground* vs. *falling-background*). All significant effects are reported together with the partial η^2 effect size measure. Where

appropriate, Greenhouse–Geisser correction was applied, and the ε correction factors are reported. Post-hoc tests were performed with Tukey HSD.

Comparing deviance-detection responses between the perceptual organizations. For comparing the deviance-related ERP responses between the different percepts, the EEG of the *ambiguous* sequences was epoched from -100 to +300 ms relative to deviation onset. For location deviants (and the corresponding standards), deviation onset corresponds to stimulus onset. For duration deviants, deviation onset corresponds to stimulus offset for duration shortenings, and to expected stimulus offset for duration prolongations. Physically identical standards were chosen for comparison and epoched with the same temporal reference. In all cases, baseline correction was performed using the 100-ms pre-stimulus interval. Epochs with a signal range above 100 μV at any electrode were excluded from the averaging. Moreover, responses to tones immediately following a deviant were excluded, and a time interval of -300 to +400 ms around each button press or release was also excluded to avoid overlap with response-related potentials and to allow for reorganization of the conscious perceptual interpretation of the tone sequence. The remaining artifact-free epochs were averaged separately for standard and deviant tones, and for each percept. Again, epochs elicited during the *falling-timbre A* and *falling-timbre B* percepts were collapsed into one *falling* percept. Separate averages were formed for *falling-foreground* and *falling-background* tones. Deviant-minus-standard difference waveforms were created by subtracting the average ERP elicited by standard tones from the average ERP elicited by deviant tones.

The same analysis steps (without the separation by *percept*) were repeated for the EEG of the *control* sequences. Only epochs during which participants gave correct responses to the tone sequence were accepted for averaging.

For statistical analysis, data were re-referenced to the right mastoid electrode based on a priori knowledge on MMN topography, following standard recommendations to fully evaluate the component by capturing both frontal and temporal contributions (Kujala, Tervaniemi, & Schröger, 2007). The time window for analyzing MMN was set to 90-130 ms from deviation onset. MMNs for duration and location deviants were tested

separately for the control sequences (control-*rising* for testing duration MMN, control-*falling* for testing location MMN), and for the ambiguous sequences separately for each of the three percepts (*rising*, *falling-foreground*, *falling-background*). The presence of the MMN component was verified by one-tailed *t*-tests of the averaged signal of the frontocentral electrode cluster (F3, Fz, F4, C3, Cz and C4). MMN amplitudes were then compared across the four conditions by means of a repeated-measures ANOVA with the 4-level factor Condition (*control*, *ambiguous-rising*, *ambiguous-falling-foreground*, *ambiguous-falling-background*).

Results

Behavioral Measures

Each participant reported switching between the different percepts within the stimulus blocks. Only one participant did not experience all the three possible alternatives within the ambiguous sequences: this participant reported only the *rising* and *falling-A* percepts but not *falling-B* (except in the corresponding control sequences). The proportion of the *rising* percept across participants was 27.02% (S.D. = 11.70%), while *falling-A* was reported in 39.91% (S.D. = 10.00%) and *falling-B* in 32.80% (S.D. = 15.19%) of the time. The average phase duration was 4.72 s (S.D. = 2.23 s) for the *rising* percept, 8.29 s for the *falling-A* (S.D. = 5.24 s) and 7.48 for the *falling-B* (S.D. = 4.65 s) percepts. Participants did not press any of the buttons only during the remaining 0.27% of the stimulus time, suggesting that they heard one of the three predefined patterns most of the time.

ERP Measures

Comparing the processing of the individual tones between the perceptual organizations. Figure 2 shows the ERPs recorded during the *rising, falling-foreground* and the *falling-background* percept (note that foreground vs. background refers to the center tone). The ANOVA of the P1 amplitude revealed a significant main effect of Percept [$F(2,32) = 10.954$, $\varepsilon = 0.938$, $p < .001$, $\eta^2 = 0.406$]. This was caused by significantly larger P1 amplitudes during the *falling-background* percept than during the *rising* and the *falling-foreground* percept [Tukey HSD with df = 32: $p = .005$ and $p < .001$]. P1 amplitudes obtained during the *rising* and the *falling-foreground* percept did not significantly differ from each other [Tukey HSD with df = 32: $p = .537$].

The observed P1 modulation can thus be accounted for by whether a tone belonged to the perceptual foreground or background, while the overall perceptual organization (*rising* vs. *falling*) had no effect on the P1 amplitude.

The statistical analysis of the N1 amplitude measured from the preceding P1 peak revealed no main effect of Percept [$F(2,32) = 0.845$, $p = .439$].

Comparing deviance-detection responses between the perceptual organizations. MMN for duration deviants was elicited in the control-*rising* conditions [$t(16) = -1.989$, $p = .032$] as well as in all ambiguous conditions, regardless of whether the percept was *rising* [$t(16) = -2.081$, $p = .027$], *falling-foreground* [$t(16) = -3.079$, $p = .003$] or *falling-background* [$t(16) = -2.559$, $p = .011$] (cf. Figure 3, top row). The MMN amplitude did not differ between these four conditions [$F(3,48) = 0.114$, $p = .951$]. In contrast, location MMN was only elicited in the control-*falling* condition [$t(16) = -2.810$, $p = .006$] but not in any of the ambiguous conditions (*rising* [$t(16) = 0.080$, $p = .468$], *falling-foreground* [$t(16) = 0.480$, $p = .319$], *falling-background* [$t(16) = -0.927$, $p = .184$]; cf. Figure 3, bottom row). This result was corroborated by the significant effect of Condition in the ANOVA [$F(3,48) = 2.994$, $\varepsilon = 0.923$, $p = .039$, $\eta^2 = 0.158$].

Discussion

We tested the utility of a new multistable stimulus paradigm for studying perceptual sound organization, and applied it for the investigation of the foreground-background decomposition of an auditory scene. Our paradigm, which was based on Wessel's (1979) work, proved to be suitable for eliciting multistable auditory perception with prolonged exposure (4-minute stimulus sequences). By means of this paradigm, we showed that the initial sensory processing of an incoming tone (indicated by the P1 wave of the ERP) is modulated by whether this tone belongs to the currently perceived foreground or background, whereas the quality of the overall perceptual organization ('Integrated' vs. 'Segregated') did not affect the ERP in the P1 latency range. Further, we found that sound grouping reflected by the MMN component does not fully correspond to the consciously experienced perceptual organization in this paradigm.

A New Tool For Investigating Auditory Multistability

Wessel (1979) reported a stimulus configuration that is ambiguous in that the same physical input can be interpreted in different ways. We showed here that this ambiguity translates into multistability with prolonged exposure to the stimulus sequence. Behavioral reports indicated that participants experienced perceptual switching between the alternative sound organizations, similarly to that shown by previous studies using the classical 'ABA' auditory streaming paradigm (Bendixen et al., 2013; Bendixen et al., 2010; Bőhm et al., 2013; Denham et al., 2010, 2013; Denham & Winkler, 2006; Pressnitzer & Hupé, 2006; Roberts et al., 2002; Szalárdy, Bendixen et al., 2013). We thereby add a new option to the range of multistable phenomena in audition (Schwartz et al., 2012). Reporting perception is easy for participants in this paradigm, because the *rising* and *falling* pitch arrangement of the 'Integrated' and 'Segregated' sound organizations enables a clear perceptual decision. While experiencing segregation (*falling pitch*), participants can exploit the timbre difference between the sounds and thus again can clearly distinguish which sound stream they perceive in the foreground. Only two out of twenty-two participants had difficulties in performing the task (due to difficulties in discriminating the two different timbres). The remaining twenty participants reported no hesitation about their experienced percepts. This verbal report is corroborated by the very low proportion of responses during which participants were unsure of their percept.

After the experiment, participants' told that they never heard the two kinds of *falling* percepts at the same time. Instead, one of them always formed the foreground (*falling A* or *falling B*), and tones of the other timbre were perceived as interspersed events that did not form another *falling* stream in the background but remained as single-tone 'leftovers' from the foreground. This gives a much clearer foreground-background distinction than the classical 'ABA' paradigm, in which an 'A' and a 'B' stream are usually formed and experienced in parallel during the 'Segregated' percept. A possible explanation for the clear foreground-background phenomenology in the present paradigm is the overlap in the frequency regions occupied by the two different 'streams', requiring a high amount of inhibition in order to link the tones of the foreground timbre across the intervening sounds of the other timbre. In summary, the paradigm appears to be suitable for encouraging a clear distinction between the foreground and the background in addition to that of the

'Integrated' and the 'Segregated' sound organization, and thus permits to disentangle these two different aspects of decomposing an auditory scene.

Foreground-Background Discrimination Indicated By P1

We analyzed early sensory processing of incoming tones by means of comparing the P1 and N1 amplitudes between different perceptual sound organizations. P1 amplitude was affected by one aspect of auditory scene decomposition: whether the sound was a part of the foreground or the background, but not by the perceptual organization ('Integrated' vs. 'Segregated') per se. The N1 wave (measured from the preceding P1 peak) was not affected by either aspect of auditory scene decomposition. This is an important observation with regard to previous findings in the 'ABA' paradigm. These results showed percept-dependent differences in sensory processing (Gutschalk et al., 2005; Hill et al., 2011; Szalárdy, Bőhm et al., 2013; Winkler et al., 2005) but could not reveal whether these differences were due to the large-scale perceptual organization of the auditory scene, or to effects of foreground-background distinction. For instance, Szalárdy, Bőhm et al. (2013) found an enhanced P1 wave for tones perceived as 'Segregated' compared to 'Integrated', but this result is confounded by the fact that the 'Segregated' organization contains both foreground and background tones while the 'Integrated' organization contains only foreground tones. Similar results were shown by Gutschalk and colleagues (2005); these authors manipulated the formation of foreground and background during 'Segregated' percepts by instruction. However, this approach also suffers from possible confounding factors, because the listener's task is different depending on the actual perceptual organization. Using unbiased instructions (i.e., no task difference accompanying the perceived sound organization) we show here that it is the foreground-background decomposition and not the large-scale organization of the auditory scene itself that affects the early sensory processing of incoming tones.

Sound processing was affected as early as 70 ms following stimulus onset. When a sound was a part of the background, it elicited more positive amplitudes in the latency range of the P1 than when the same sound belonged to the foreground. This effect probably does not reflect a genuine P1 amplitude modulation, but a longer-lasting ERP modulation covering also the N1 latency range (see below). Studies investigating

bistability in the visual domain (e.g., binocular rivalry) have also suggested that the neural correlates of perception appear as early as in the P1 latency range. For instance, Valle-Inclan, Hackley, de Labra and Alvarez (1999) found higher ERP amplitudes from about 70 ms after stimulus onset for stimuli presented to the dominant eye compared to the suppressed eye (see also Roeber & Schröger, 2004; Roeber et al., 2008).

Such early ERP modulations are difficult to reconcile with bottom-up effects on stimulus processing, but can be explained by top-down modulations of early sensory processes (realized via efferent feedback) depending on the current perceptual organization. Top-down effects would be facilitated by the fact that stimulus onset times were fully predictable in the present paradigm. Such top-down modulations might consist in an enhanced processing of foreground and/or a suppressed processing of background stimuli. According to the baseline hypothesis proposed by Hillyard and Anllo-Vento (1998) as well as Luck and colleagues (Luck & Hillyard, 1995; Luck et al., 1994), the P1 is suppressed for unattended stimuli compared to a neutral baseline or an attended stimulus. Yet a contrasting hypothesis postulates that the P1 increases with inhibition (Klimesch, 2011). This so-called inhibition hypothesis suggests that P1 does not reflect sensory processing and cannot be explained as a sensory evoked component as it is not affected by the stimulus properties. Rather, it reflects inhibition of task-irrelevant stimuli and/or networks. Our results are in line with the inhibition hypothesis, showing larger P1 for background than for foreground tones. This is consistent with our above suggestion of selectively suppressing the tones of one timbre in order to be able to perceive the tones of the other timbre as a coherent stream in the foreground.

In contrast to the P1, the amplitude of the N1 was not affected by the percept. This result is consistent with the results of Szalárdy, Bőhm et al. (2013) who found that the N1 wave did not show percept-dependent effects in the classical ‘ABA’ streaming paradigm. However, it contrasts the results of Gutschalk et al. (2005), who found that both the P1 and the N1 were modulated by the percept in ‘ABA’ sequences. It is possible that Gutschalk et al.’s results were due to modulation of the N1 wave by attentional processes (Hillyard, Hink, Schwent, & Picton, 1973; Parasuraman, 1978). Indeed, Gutschalk et al. found that N1 amplitude increased when the ‘Segregated’ sound organization was perceived, and in that case participants were instructed to attend to one of the streams. This attentional bias might have resulted in the N1

enhancement in the 'Segregated' compared to the 'Integrated' sound organization. Without such attentional effects, the N1 wave appears to be unmodulated by the current percept of the listener (present data and Szalárdy, Bőhm et al., 2013). This conclusion refers to the N1 wave as measured from the preceding P1 peak, thereby removing the influence of the longer-lasting ERP modulation by percept that starts at the P1 latency range and extends into the N1 range.

Figure 2 is indicative of some percept-related ERP effects emerging even before the P1 latency range. Specifically, we observed a short-latency peak appearing at around 30 ms after stimulus onset, whose occurrence was not expected on the basis of previous studies. In a post-hoc analysis, we quantified the effect in the interval from 13 to 43 ms after stimulus onset and investigated whether it was affected by the current percept. The analysis revealed a main effect of Percept [$F(2,32) = 3.892$, $\varepsilon = 0.959$, $p = .031$, $\eta^2 = 0.196$]. Follow-up tests revealed a significantly higher amplitude during *rising* than during *falling-background* percepts [Tukey HSD with df = 32: $p = .047$], whereas the corresponding difference between *rising* and *falling-foreground* percepts failed to reach statistical significance [Tukey HSD with df = 32: $p = .063$]. The amplitude of the short-latency peak was not different between *falling-foreground* and *falling-background* percepts [Tukey HSD with df = 32: $p = .991$]. Although no post-hoc explanation for the occurrence of this peak and its percept-related modulation can be given at this point, it is important to consider that its modulation by the current percept differed from the pattern observed during the P1 latency range. Specifically, *falling-background* sounds elicited markedly different amplitudes than *falling-foreground* and *rising* sounds in the P1 latency range, whereas it was the sounds perceived as *rising* that differed from *falling-background* and (by tendency) from *falling-foreground* during the early short-latency peak. Therefore, the effect in the P1 latency range cannot simply reflect a carry-over effect from this earlier difference. Nevertheless, future studies should carefully investigate both the early short-latency peak and the longer-lasting ERP modulation by percept starting at the P1 latency range. Special care should be given to the possibility that either of these effects might reflect a carry-over effect from the processing of the previous sound rather than a modulation of sensory processing of the newly incoming sound.

Taken together, our results indicate that foreground-background discrimination in a segregated auditory scene affects the ERPs elicited by incoming tones at early latency ranges. These effects may be associated with the inhibition of background tones, though future studies are required to confirm this interpretation. Note that the observed ERP modulation not only provides information about sound processing mechanisms, but it can also serve as a post-hoc verification of participants performing their task in accordance with the instructions. This is because "random" pressing of the buttons would not have led to reliable ERP effects.

Percept-Independent MMN Elicitation

Our second set of ERP analyses was conducted to study the correspondence between the reported perceptual organizations and the grouping of the tones in the auditory system. Tone grouping was measured indirectly via the elicitation of the MMN component by infrequent violations of feature regularities that could only be extracted if the corresponding groups were formed. Two types of regularities and corresponding violations were inserted into the sequence, allowing us to test at the same time the existence of groupings corresponding to the 'Integrated' (*rising*) and 'Segregated' (*falling*) organizations at the level of the deviance detection processes reflected by the MMN component.

The first hypothesis for the relationship between the grouping processes underlying MMN and perception (see Introduction) was that the deviance-detection processes reflected by MMN refer only to that representation of the sound sequence which is currently consciously experienced by the listener (Rahne et al., 2007; Sussman et al., 1998). We did not find evidence for such a correspondence between the groupings in the auditory system and in conscious perception. Our results are instead compatible with the second hypothesis: MMN elicitation followed the physical arrangement of the tones (Ross et al., 1996). MMN was elicited for violations of the duration regularity, which was carried by physically adjacent tones both during *rising* and *falling* percepts, and even when the deviant tone belonged to the background. In contrast, MMN for violations of the location regularity, carried by non-adjacent tones, was not elicited in any of the ambiguous conditions, not even when the percept at the time of encountering the deviant was *falling*, thereby promoting the sound organization that should have facilitated regularity extraction.

We also found no evidence for the third hypothesis, the parallel existence of alternative groupings in the auditory system (Horváth et al., 2001; Winkler et al., 2012). On this hypothesis, MMN should have been elicited by each regularity violation, irrespective of the current percept. However, MMN was only elicited by violations of the duration regularity which corresponds to grouping the rising patterns, but not by violations of the location regularity, which corresponded to falling patterns. Therefore, multistable perception in this paradigm cannot be conceptualized as a selection from the alternative groupings initially formed by lower-level auditory processes contributing to the elicitation of MMN. Thus the current results are compatible with the conclusion of some previous studies dissociating MMN and perception (Sussman, Winkler, Huotilainen, Ritter, & Näätänen, 2002; Takegata et al., 2005; van Zuijen et al., 2006).

Importantly, location MMN was clearly elicited in the control sequences with the *falling* stimulus arrangement. Therefore, it was principally possible to extract the location regularity just as well as the duration regularity. The most plausible explanation for the absence of location MMN in the ambiguous sequences then is that MMN-related processing was bound by the physical arrangement of the tone sequence, in particular by temporal adjacency of the tones carrying the feature regularities. Note that adjacency is not a necessary constraint for regularity extraction (Bendixen, Schröger, Ritter, & Winkler, 2012) and thus it would have been possible to find location MMN even during *rising* percepts. Yet not finding location MMN even for deviants in the foreground stream during a *falling* percept is indeed surprising in view of many previous studies showing that conscious perception corresponds with MMN-related processing (Rahne et al., 2007; Sussman et al., 1998, 1999; Winkler, Kushnerenko et al., 2003; Winkler, Sussman et al., 2003; Winkler, Teder-Salejärvi et al., 2003). However, in these studies, conscious perception did not arise from multistable stimulus configurations but was driven by experimental manipulations such as changing basic auditory grouping cues (e.g., frequency or temporal proximity), adding visual cues, or giving instructions to maintain a certain percept. A previous study that investigated percept-dependent deviance detection without the confounding influence of actual stimulus or instruction changes (Winkler et al., 2005) likewise found that an early deviance detection response was not affected by conscious perception, although the response obtained in this study was too early for being an MMN component. The

current results suggest that MMN elicitation does not necessarily follow the reported perceptual organization of the participants, and thus perceptual grouping in our paradigm is likely neither constrained by nor acts upon the representations available to MMN-related processing.

Integrating The Findings On Single-Tone Processing And Deviance Detection

We showed that conscious perceptual organization affected the initial sensory processing of the tones as early as 70 ms after stimulus onset (indexed by the P1 amplitude), but did not affect deviance detection processes at around 110 ms after deviation onset (indexed by MMN elicitation). The temporal relation of these effects seems difficult to reconcile with the view that the sound groupings available to MMN-related processing represent an intermediate step in the decomposition of the auditory scene (Winkler et al., 2005). If conscious perceptual grouping occurred on the basis of the extracted feature regularities, it should not affect sound processing at earlier latency ranges than MMN-related processing. Therefore, the present data suggest that the P1 and MMN components reflect parallel and independent processing routes in this paradigm.

Indeed, if P1 does not reflect sensory processing, but rather inhibition of the currently task-irrelevant stimuli (Klimesch, 2011), then the two processes do not need to fit a single processing sequence and both can feed to conscious perception. The deviance detection indexed by MMN is largely based on sensory (bottom-up) processes (Ross et al., 1996) with only limited modulations by top-down effects (Sussman, 2007). Some of the top-down effects, such as target selection, are not reflected by the MMN (see, e.g., Näätänen, 1990). Inhibition of the processing of sounds which are currently in the background may be another top-down process not affecting deviance detection. Note that many previous studies showing that MMN elicitation is percept-dependent applied stimulus paradigms in which perceptual organization was driven by sensory (bottom-up) factors (e.g., Sussman et al., 1999; Winkler, Kushnerenko et al., 2003; Winkler, Sussman et al., 2003), whereas percept-independent MMN elicitation has been shown for more complex forms of perceptual grouping involving top-down factors (Ross et al., 1996; but see Sussman et al., 1998).

We suggest that the two systems indexed by P1 and MMN provide complementary information: Whereas the system indexed by P1 appears to be related to the stimuli which are currently suppressed, the system indexed

by MMN is related to possible groupings of the sounds primarily based on their acoustic features. Selecting groupings for perception should then occur after the processes reflected by MMN and could utilize information from both routes. However, the lack of MMN elicitation by regularity violations relevant for the grouping underlying the falling percept suggests that even taken together, these two systems do not cover all aspects of sound grouping. That is, in the current study, we found no ERP correlate for the groupings based on timbre similarity. Whether this indicates that the processing of timbre-based grouping[1] is done by a dedicated system or that the current stimulus configuration represents a larger category of grouping processes hitherto not addressed in the simplified paradigms tested with ERPs remains to be seen, as also the relation of these assumed grouping processes to the ones reflected by MMN.

Future Directions

The design of the present study was optimized towards studying the effects of the currently experienced perceptual organization on ERP correlates of sensory processing. A valuable addition might be provided by additionally taking the time since switching to the current perceptual organization (i.e., the time since the last response) into account. This would allow for investigating dynamic aspects of auditory perceptual organization. These aspects of auditory stream segregation have been investigated using the classical 'ABA' paradigm (Snyder, Carter, Lee, Hannon, & Alain, 2008; Snyder, Carter, Hannon, & Alain, 2009; Snyder, Holder et al., 2009) and could be easily transferred to our new multistability paradigm. The study of dynamics could also be extended to assess individual differences, investigating, for example, whether listeners with shorter perceptual phases show more pronounced inhibition processes.

Conclusions

We proposed a new paradigm of auditory multistability that lends itself to investigating not only auditory integration and segregation mechanisms, but also foreground-background decomposition of an auditory scene. With this paradigm, we showed that the processing of incoming sounds was affected by the foreground-background distinction as early as 70 ms after sound onset. Conscious perceptual organization

and foreground-background distinction had, however, no effect on auditory grouping as measured by MMN-related processing.

References

Alain, C., & Woods, D. L. (1993). Distractor clustering enhances detection speed and accuracy during selective listening. *Perception & Psychophysics, 54*(4), 509-514. doi: 10.3758/BF03211773

Alain, C., & Woods, D. L. (1994). Signal clustering modulates auditory cortical activity in humans. *Perception & Psychophysics, 56*(5), 501-516. doi: 10.3758/BF03206947

Arnott, S. R., & Alain, C. (2002). Stepping out of the spotlight: MMN attenuation as a function of distance from the attended location. *Neuroreport, 13*(17), 2209-2212. doi: 10.1097/01.wnr.0000045010.30898.42.cf

Bendixen, A., Bőhm, T., Szalárdy, O., Mill, R., Denham, S. L., & Winkler, I. (2013). Different roles of similarity and predictability in auditory stream segregation. *Learning and Perception, 5*(2), 37-54. doi: 10.1556/LP.5.2013.Suppl2.4

Bendixen, A., Denham, S. L., Gyimesi, K., & Winkler, I. (2010). Regular patterns stabilize auditory streams. *Journal of the Acoustical Society of America, 128*(6), 3658-3666. doi: 10.1121/1.3500695

Bendixen, A., Schröger, E., Ritter, W., & Winkler, I. (2012). Regularity extraction from non-adjacent sounds. *Frontiers in Psychology, 3*, 143. doi: 10.3389/fpsyg.2012.00143

Blake, R., & Logothetis, N. K. (2002). Visual competition. *Nature Reviews Neuroscience, 3*(1), 13-23. doi: 10.1038/nrn701

Bőhm, T. M., Shestopalova, L., Bendixen, A., Andreou, A. G., Georgiou, J., Garreau, G., . . . Winkler, I. (2013). The role of perceived source location in auditory stream segregation: Separation affects sound organization, common fate does not. *Learning and Perception, 5*(2), 55-72. doi: 10.1556/LP.5.2013.Suppl2.5

Chatrian, G. E., Lettich, E., & Nelson, P. L. (1985). Ten percent electrode system for topographic studies of spontaneous and evoked EEG activities. *American Journal of EEG Technology, 25*, 83–92.

Cusack, R., Deeks, J., Aikman, G., & Carlyon, R. P. (2004). Effects of location, frequency region, and time course of selective attention on auditory scene analysis. *Journal of Experimental Psychology: Human Perception and Performance, 30*(4), 643-656. doi: 10.1037/0096-1523.30.4.643

Cusack, R., & Roberts, B. (2000). Effects of differences in timbre on sequential grouping. *Perception & Psychophysics, 62*(5), 1112-1120. doi: 10.3758/BF03212092

Delorme, A., & Makeig, S. (2004). EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods, 134*(1), 9-21. doi: 10.1016/j.jneumeth.2003.10.009

Denham, S. L., Gyimesi, K., Stefanics, G., & Winkler, I. (2010). Stability of perceptual organisation in auditory streaming. In E. A. Lopez-Poveda, A. R. Palmer & R. Meddis (Eds.), *The Neurophysiological Bases of Auditory Perception.* (pp. 477-487). New York: Springer.

Denham, S. L., Gyimesi, K., Stefanics, G., & Winkler, I. (2013). Perceptual bi-stability in auditory streaming: How much do stimulus features matter? *Learning and Perception, 5*(2), 73-100. doi: 10.1556/LP.5.2013.Suppl2.6

Denham, S. L., & Winkler, I. (2006). The role of predictive models in the formation of auditory streams. *Journal of Physiology, Paris, 100*(1-3), 154-170. doi: 10.1016/j.jphysparis.2006.09.012

Dolležal, L.-V., Beutelmann, R., & Klump, G. M. (2012). Stream segregation in the perception of sinusoidally amplitude-modulated tones. *PLoS One, 7*(9), e43615. doi: 10.1371/journal.pone.0043615

Elhilali, M., Xiang, J. J., Shamma, S. A., & Simon, J. Z. (2009). Interaction between attention and bottom-up saliency mediates the representation of foreground and background in an auditory scene. *PLoS Biology, 7*(6), e1000129. doi: 10.1371/Journal.Pbio.1000129

Escera, C., Alho, K., Winkler, I., & Näätänen, R. (1998). Neural mechanisms of involuntary attention to acoustic novelty and change. *Journal of Cognitive Neuroscience, 10*(5), 590-604. doi: 10.1162/089892998562997

Gratton, G., Coles, M. G., & Donchin, E. (1983). A new method for off-line removal of ocular artifact. *Electroencephalography and Clinical Neurophysiology, 55*(4), 468-484. doi: 10.1016/0013-4694(83)90135-9

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*(3), 1340-1348. doi: 10.1121/1.1452740

Gutschalk, A., Micheyl, C., Melcher, J. R., Rupp, A., Scherg, M., & Oxenham, A. J. (2005). Neuromagnetic correlates of streaming in human auditory cortex. *Journal of Neuroscience, 25*(22), 5382-5388. doi: 10.1523/JNEUROSCI.0347-05.2005

Hill, K. T., Bishop, C. W., Yadav, D., & Miller, L. M. (2011). Pattern of BOLD signal in auditory cortex relates acoustic response to perceptual streaming. *BMC Neuroscience, 12*, 85. doi: 10.1186/1471-2202-12-85

Hillyard, S. A., & Anllo-Vento, L. (1998). Event-related brain potentials in the study of visual selective attention. *Proceedings of the National Academy of Sciences of the United States of America, 95*(3), 781-787. doi: 10.2307/44189

Hillyard, S. A., Hink, R. F., Schwent, V. L., & Picton, T. W. (1973). Electrical signs of selective attention in the human brain. *Science, 182*(108), 177-180. doi: 10.1126/science.182.4108.177

Horváth, J., Czigler, I., Sussman, E., & Winkler, I. (2001). Simultaneously active pre-attentive representations of local and global rules for sound sequences in the human brain. *Cognitive Brain Research, 12*(1), 131-144. doi: 10.1016/S0926-6410(01)00038-6

Iverson, P. (1995). Auditory stream segregation by musical timbre: Effects of static and dynamic acoustic attributes. *Journal of Experimental Psychology: Human Perception and Performance, 21*(4), 751-763. doi: 10.1037/0096-1523.21.4.751

Jasper, H. H. (1958). The ten-twenty electrode system of the International Federation. *Electroencephalography and Clinical Neurophysiology, 10*, 371-375.

Klimesch, W. (2011). Evoked alpha and early access to the knowledge system: the P1 inhibition timing hypothesis. *Brain Research, 1408*, 52-71. doi: 10.1016/j.brainres.2011.06.003

Kondo, H. M., Kitagawa, N., Kitamura, M. S., Koizumi, A., Nomura, M., & Kashino, M. (2012). Separability and commonality of auditory and visual bistable perception. *Cerebral Cortex, 22*(8), 1912-1922. doi: 10.1093/cercor/bhr266

Köhler, W. (1947). *Gestalt psychology: An introduction to new concepts in modern psychology.* New York: Liveright Publishing Corporation.

Kujala, T., Tervaniemi, M., & Schröger, E. (2007). The mismatch negativity in cognitive and clinical neuroscience: Theoretical and methodological considerations. *Biological Psychology, 74*(1), 1-19. doi: http://10.1016/j.biopsycho.2006.06.001

Leopold, D. A., & Logothetis, N. K. (1999). Multistable phenomena: Changing views in perception. *Trends in Cognitive Sciences, 3*(7), 254-264. doi: 10.1016/S1364-6613(99)01332-7

Leopold, D. A., Wilke, M., Maier, A., & Logothetis, N. K. (2002). Stable perception of visually ambiguous patterns. *Nature Reviews Neuroscience, 5*(6), 605-609. doi: 10.1038/nn851

Luck, S. J., & Hillyard, S. A. (1995). The role of attention in feature detection and conjunction discrimination: an electrophysiological analysis. *International Journal of Neuroscience, 80*(1-4), 281-297. doi: 10.3109/00207459508986105

Luck, S. J., Hillyard, S. A., Mouloua, M., Woldorff, M. G., Clark, V. P., & Hawkins, H. L. (1994). Effects of spatial cuing on luminance detectability: psychophysical and electrophysiological evidence for early selection. *Journal of Experimental Psychology: Human Perception and Performance, 20*(4), 887-904. doi: 10.1037/0096-1523.20.4.887

Moore, B. C. J., & Gockel, H. E. (2012). Properties of auditory stream formation. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 367*(1591), 919-931. doi: 10.1098/Rstb.2011.0355

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*(11), 1-18. doi: 10.1167/10.11.1

Näätänen, R. (1990). The role of attention in auditory information-processing as revealed by event-related potentials and other brain measures of cognitive function. *Behavioral and Brain Sciences, 13*(2), 257-258. doi: 10.1017/S0140525X00078687

Näätänen, R., Paavilainen, P., Rinne, T., & Alho, K. (2007). The mismatch negativity (MMN) in basic research of central auditory processing: A review. *Clinical Neurophysiology, 118*(12), 2544-2590. doi: 10.1016/j.clinph.2007.04.026

Nager, W., Teder-Salejärvi, W., Kunze, S., & Münte, T. F. (2003). Preattentive evaluation of multiple perceptual streams in human audition. *Neuroreport, 14*(6), 871-874. doi: 10.1097/01.wnr.0000069961.11849.3d

Orbach, J., Ehrlich, D., & Heath, H. A. (1963). Reversibility of the Necker Cube. I. An examination of the concept of "Satiation of Orientation". *Perceptual and Motor Skills, 17*, 439-458. doi: 10.2466/pms.1963.17.2.439

Parasuraman, R. (1978). Auditory evoked potentials and divided attention. *Psychophysiology, 15*(5), 460-465. doi: 10.1111/j.1469-8986.1978.tb01416.x

Pressnitzer, D., & Hupé, J.-M. (2006). Temporal dynamics of auditory and visual bistability reveal common principles of perceptual organization. *Current Biology, 16*(13), 1351-1357. doi: 10.1016/j.cub.2006.05.054

Rahne, T., Böckmann, M., von Specht, H., & Sussman, E. S. (2007). Visual cues can modulate integration and segregation of objects in auditory scene analysis. *Brain Research, 1144*, 127-135. doi: 10.1016/j.brainres.2007.01.074

Roberts, B., Glasberg, B. R., & Moore, B. C. J. (2002). Primitive stream segregation of tone sequences without differences in fundamental frequency or passband. *Journal of the Acoustical Society of America, 112*(5), 2074-2085. doi: 10.1121/1.1508784

Roeber, U., & Schröger, E. (2004). Binocular rivalry is partly resolved at early processing stages with steady and with flickering presentation: a human event-related brain potential study. *Neuroscience Letters*, *371*(1), 51-55. doi: 10.1016/j.neulet.2004.08.038

Roeber, U., Widmann, A., Trujillo-Barreto, N. J., Herrmann, C. S., O'Shea, R. P., & Schröger, E. (2008). Early correlates of visual awareness in the human brain: time and place from event-related brain potentials. *Journal of Vision*, *8*(3), 1-12. doi: 10.1167/8.3.21

Ross, J., Tervaniemi, M., & Näätänen, R. (1996). Neural mechanisms of the octave illusion: electrophysiological evidence for central origin. *Neuroreport, 8*(1), 303-306. doi: 10.1097/00001756-199612200-00060

Schröger, E. (2007). Mismatch negativity: A microphone into auditory memory. *Journal of Psychophysiology, 21*(3-4), 138-146. doi: 10.1027/0269-8803.21.34.138

Schwartz, J.-L., Grimault, N., Hupé, J.-M., Moore, B. C. J., & Pressnitzer, D. (2012). Multistability in perception: binding sensory modalities, an overview. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 367*(1591), 896-905. doi: 10.1098/rstb.2011.0254

Singh, P. G. (1987). Perceptual organization of complex tone sequences: A tradeoff between pitch and timbre? *Journal of the Acoustical Society of America, 82*(3), 886-899. doi: 10.1121/1.395287

Snyder, J. S., Carter, O. L., Hannon, E. E., & Alain, C. (2009). Adaptation reveals multiple levels of representation in auditory stream segregation. *Journal of Experimental Psychology: Human Perception and Performance, 35*(4), 1232-1244. doi: 10.1037/a0012741

Snyder, J. S., Carter, O. L., Lee, S. K., Hannon, E. E., & Alain, C. (2008). Effects of context on auditory stream segregation. *Journal of Experimental Psychology: Human Perception and Performance, 34*(4), 1007-1016. doi: 10.1037/0096-1523.34.4.1007

Snyder, J. S., Holder, W. T., Weintraub, D. M., Carter, O. L., & Alain, C. (2009). Effects of prior stimulus and prior perception on neural correlates of auditory stream segregation. *Psychophysiology, 46*(6), 1208-1215. doi: 10.1111/j.1469-8986.2009.00870.x

Sussman, E., Ritter, W., & Vaughan, H. G. (1999). An investigation of the auditory streaming effect using event-related brain potentials. *Psychophysiology, 36*(1), 22-34. doi: 10.1017/S0048577299971056

Sussman, E., Ritter, W., & Vaughan, H. G., Jr. (1998). Attention affects the organization of auditory input associated with the mismatch negativity system. *Brain Research, 789*(1), 130-138. doi: 10.1016/S0006-8993(97)01443-1

Sussman, E., Winkler, I., Huotilainen, M., Ritter, W., & Näätänen, R. (2002). Top-down effects can modify the initially stimulus-driven auditory organization. *Cognitive Brain Research, 13*(3), 393-405. doi: 10.1016/S0926-6410(01)00131-8

Sussman, E. S. (2007). A new view on the MMN and attention debate - The role of context in processing auditory events. *Journal of Psychophysiology, 21*, 164-175. doi: 10.1027/0269-8803.21.34.164

Sussman, E. S., Bregman, A. S., Wang, W. J., & Khan, F. J. (2005). Attentional modulation of electrophysiological activity in auditory cortex for unattended sounds within multistream auditory environments. *Cognitive, Affective, and Behavioral Neuroscience, 5*(1), 93-110. doi: 10.3758/CABN.5.1.93

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*(2), 149-161. doi: 10.1556/LP.5.2013.Suppl2.9

Szalárdy, O., Bőhm, T. M., Bendixen, A., & Winkler, I. (2013). Event-related potential correlates of sound organization: Early sensory and late cognitive effects. *Biological Psychology*. doi: 10.1016/j.biopsycho.2013.01.015

Takegata, R., Brattico, E., Tervaniemi, M., Varyagina, O., Näätänen, R., & Winkler, I. (2005). Preattentive representation of feature conjunctions for concurrent spatially distributed auditory objects. *Cognitive Brain Research, 25*(1), 169-179. doi: 10.1016/j.cogbrainres.2005.05.006

Tervaniemi, M., Winkler, I., & Näätänen, R. (1997). Pre-attentive categorization of sounds by timbre as revealed by event related potentials. *Neuroreport, 8*(11), 2571-2574.

Valle-Inclan, F., Hackley, S. A., de Labra, C., & Alvarez, A. (1999). Early visual processing during binocular rivalry studied with visual evoked potentials. *Neuroreport, 10*(1), 21-25.

van Zuijen, T. L., Simoens, V. L., Paavilainen, P., Näätänen, R., & Tervaniemi, M. (2006). Implicit, intuitive, and explicit knowledge of abstract regularities in a sound sequence: an event-related brain potential study. *Journal of Cognitive Neuroscience, 18*(8), 1292-1303. doi: 10.1162/jocn.2006.18.8.1292

Wessel, D. L. (1979). Timbre space as a musical control structure. *Computer Music Journal, 3*, 45-52. doi: 10.2307/3680283

Winkler, I. (2007). Interpreting the mismatch negativity. *Journal of Psychophysiology, 21*(3-4), 147-163. doi: 10.1027/0269-8803.21.34.147

Winkler, I., Denham, S., Mill, R., Bőhm, T. M., & Bendixen, A. (2012). Multistability in auditory stream segregation: a predictive coding view. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 367*(1591), 1001-1012. doi: 10.1098/Rstb.2011.0359

Winkler, I., Denham, S. L., & Nelken, I. (2009). Modeling the auditory scene: predictive regularity representations and perceptual objects. *Trends in Cognitive Sciences, 13*(12), 532-540. doi: 10.1016/j.tics.2009.09.003

Winkler, I., Kushnerenko, E., Horváth, J., Čeponienė, R., Fellman, V., Huotilainen, M., . . . Sussman, E. (2003). Newborn infants can organize the auditory world. *Proceedings of the National Academy of Sciences of the United States of America, 100*(20), 11812-11815. doi: 10.2307/3147874

Winkler, I., Sussman, E., Tervaniemi, M., Horváth, J., Ritter, W., & Näätänen, R. (2003). Preattentive auditory context effects. *Cognitive, Affective, and Behavioral Neuroscience, 3*(1), 57-77. doi: 10.3758/CABN.3.1.57

Winkler, I., Takegata, R., & Sussman, E. (2005). Event-related brain potentials reveal multiple stages in the perceptual organization of sound. *Cognitive Brain Research, 25*(1), 291-299. doi: 10.1016/j.cogbrainres.2005.06.005

Winkler, I., Teder-Salejärvi, W. A., Horváth, J., Näätänen, R., & Sussman, E. (2003). Human auditory cortex tracks task-irrelevant sound sources. *Neuroreport, 14*(16), 2053-2056. doi: 10.1097/01.wnr.0000095496.09138.6d

Winkler, I., van Zuijen, T. L., Sussman, E., Horváth, J., & Näätänen, R. (2006). Object representation in the human auditory system. *European Journal of Neuroscience, 24*(2), 625-634. doi: 10.1111/j.1460-9568.2006.04925.x

Author Notes

Acknowledgments

This work was supported by the German Academic Exchange Service (Deutscher Akademischer Austauschdienst, DAAD; Projects 50345549 and 56265741), by the Hungarian Scholarship Board (Magyar Ösztöndíj Bizottság, MÖB; Projects P-MÖB/853 and 39589), by the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG; SCH 375/20-1 to ES; DFG Cluster of Excellence 1077 "Hearing4all"), and by the Hungarian Academy of Sciences (Magyar Tudományos Akadémia, MTA; Lendület project 2012-36/2012 to IW). The experiment was realized using Cogent 2000 developed by the Cogent 2000 team at the FIL and the ICN. The EEG data were analyzed with EEGlab (Delorme & Makeig, 2004). The authors thank Susann Duwe and Nadin Greinert for assistance in data acquisition.

Contact details for reprints

Correspondence concerning this article should be addressed to Alexandra Bendixen, Department of Psychology, Cluster of excellence "Hearing4all", European Medical School, Carl von Ossietzky University of Oldenburg, Ammerländer Heerstr. 114-118, D-26111 Oldenburg, Germany. E-Mail: alexandra.bendixen@uni-oldenburg.de

Footnotes

[1]The evidence relating MMN and timbre processing is equivocal. Small timbre changes elicit MMN (Tervaniemi, Winkler, & Näätänen, 1997). However, large timbre changes, which can be regarded as evidence that the sound was produced by a different source, elicit a large negativity in the N1 range, such that MMN and N1 cannot be disentangled (e.g., Escera, Alho, Winkler, & Näätänen, 1998).

Figure Captions

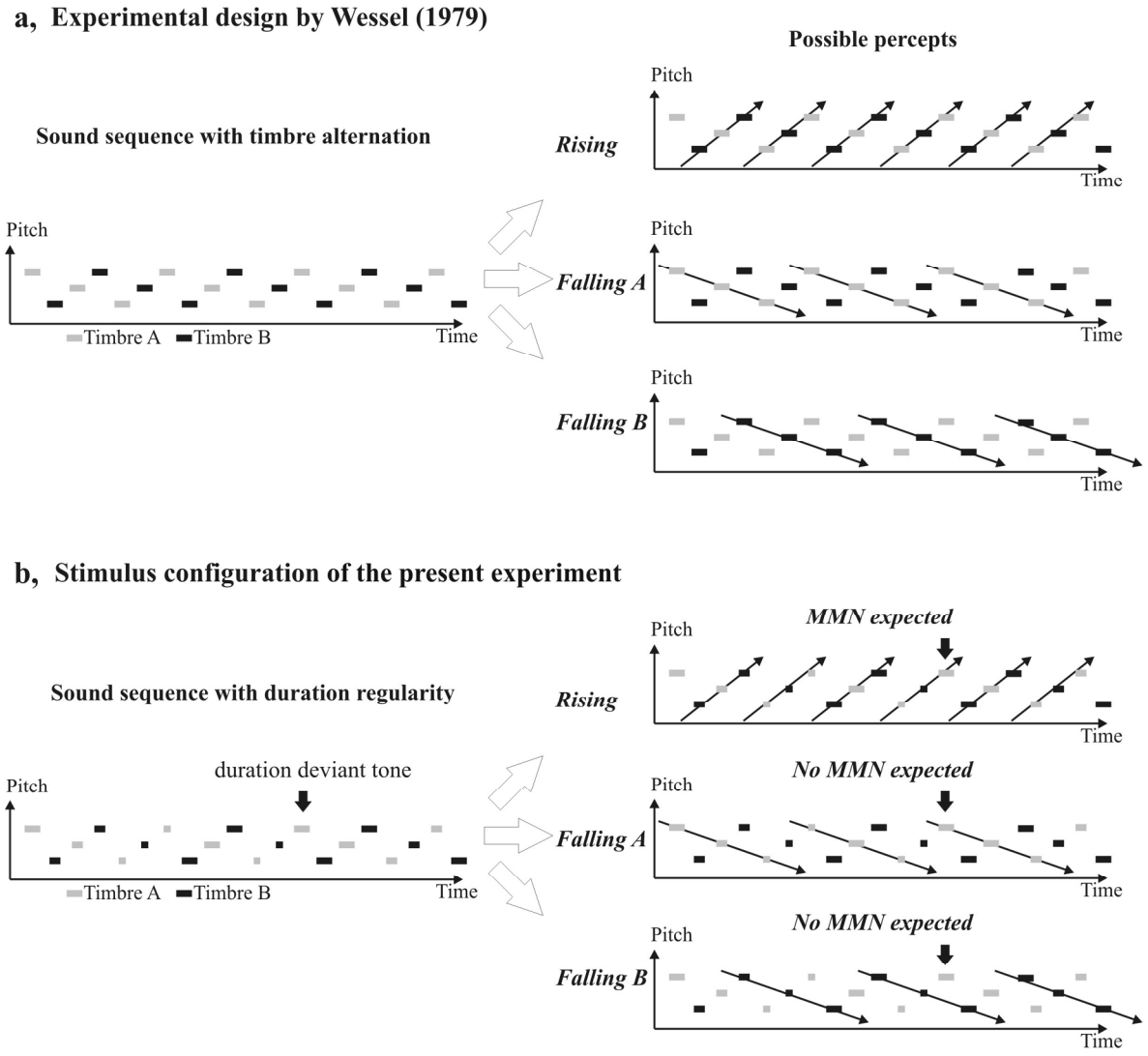

Figure 1a: Experimental design by Wessel (1979). The left side of the figure shows the sound sequence: a repeating three-tone pattern of rising frequency combined with a timbre alternation (timbre A and B). The right side of the figure demonstrates the possible percepts: *rising*, *falling A* or *falling B*. Participants continuously indicate their current perception of the sequence by depressing specified buttons. **Figure 1b**: Present paradigm, modified for testing MMN elicitation. The left side shows the same sound sequence as on Figure 1a, but each rising triplet has a separate common duration value, with duration chosen randomly for each triplet. The duration deviant tone is marked by a black arrow. The right side of the figure shows the expected results if MMN elicitation were to correspond with the conscious perceptual organization of the sounds: In this case, MMN is only expected when the sequence is perceived as *rising*. The location regularity and its violations created for the falling triplets (not depicted on the figure) followed the same principles as the duration regularity shown for the rising triplets.

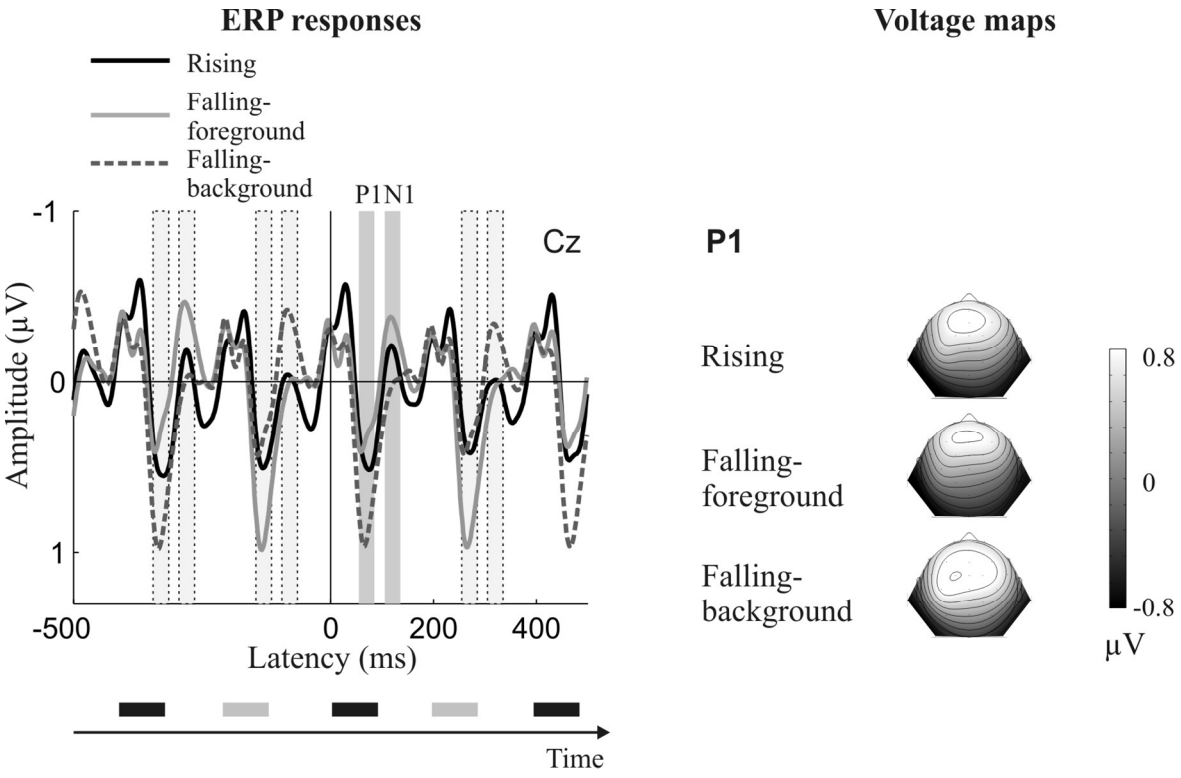

Figure 2. Left panel: Grand-average (N=17) ERP responses elicited by five consecutive tones of the same percept at the Cz electrode for the *rising* (black), *falling-foreground* (grey) and *falling-background* (dotted grey) percepts. Note that foreground vs. background refers to the center tone (starting at the 0 time point); the foreground-background structure is alternating for tones of the *falling* percept. The time intervals of the P1 and N1 are marked by grey shades. The black and grey rectangles underneath the ERP figure represent the timing of the tones, including their alternation in timbre. The dotted grey rectangles represent the P1 and N1 for two preceding tones and one following tone of the same percept. Right panel: Scalp topographies of the P1 wave elicited by the center tone, separately for the *rising* (top), *falling-foreground* (middle) and *falling-background* (bottom) percepts. Maps were spline interpolated with a smoothing factor of 10^{-7}. Calibration for the greyscale maps is shown on the right-hand side.

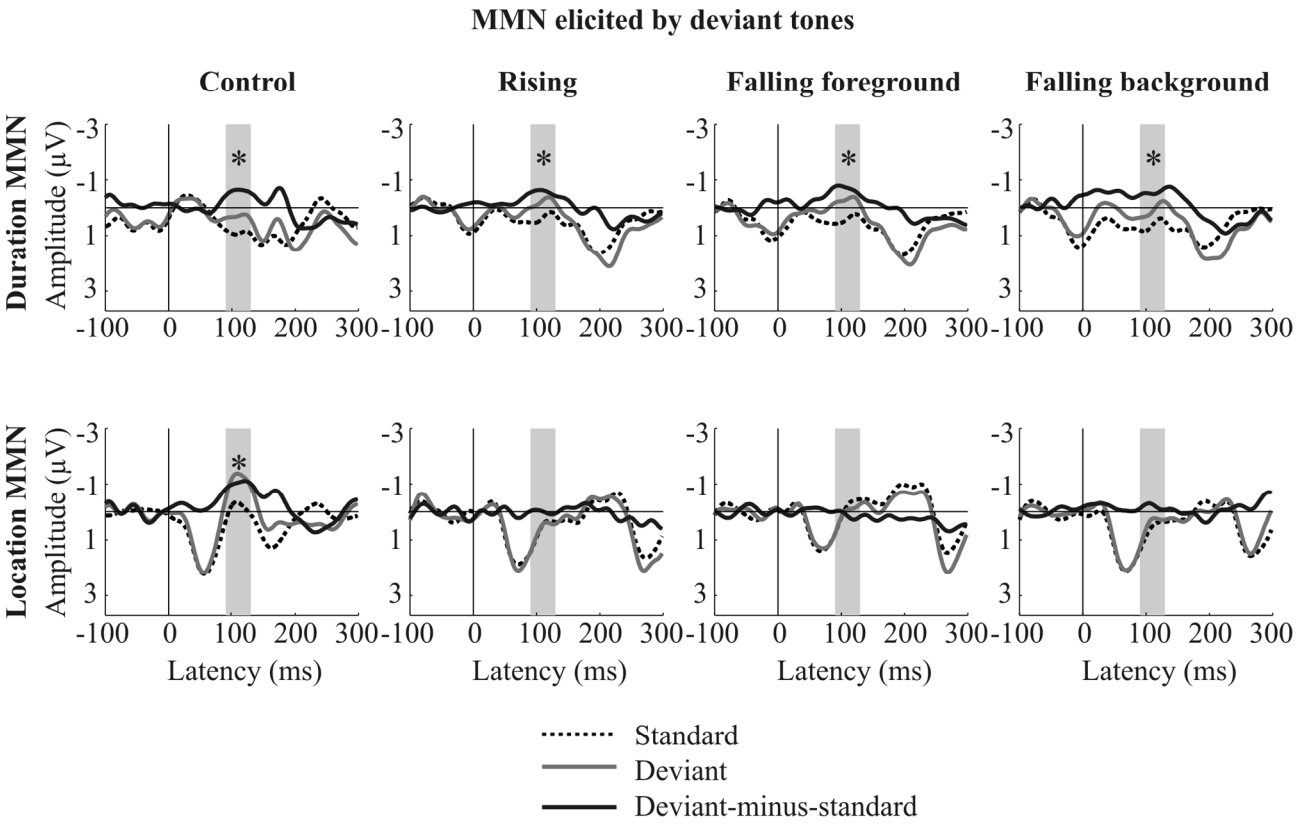

Figure 3. Grand-average (N=17) ERP responses elicited by standard (dotted black line) and deviant (solid grey line) tones, and deviant-minus-standard difference waves (solid black line) at the frontocentral electrode cluster (F3, Fz, F4, C3, Cz and C4). Top row shows ERPs for duration deviants, bottom row shows ERPs for location deviants. The different columns correspond to the different conditions and percepts: *control* sequences (extreme left), *ambiguous-rising* (left), *ambiguous-falling-foreground* (right) and *ambiguous-falling-background* (extreme right).