Budapest University of Technology and Economics

PhD School in Psychology – Cognitive Science

Zsuzsanna Kocsis

EEG signatures of separating concurrent sounds

PhD thesis

Supervisor:

Dr. István Winkler

Budapest, 2017

Contents

Contents i

Acknowledgements iii

Glossary of abbreviations iv

Abstract v

Kivonat vi

1. Introduction 1

    1.1. Auditory scene analysis 4

    1.2. Auditory object perception 20

    1.3. EEG correlates of concurrent sound segregation 28

    1.4. Auditory figure-ground segregation 39

2. Synopsis and rationale of theses 44

    2.1. Thesis I: Auditory figure-ground segregation 45

    2.2. Thesis II: Effects of multiple congruent cues on concurrent sound segregation 45

    2.3. Thesis III: Theta oscillations accompanying concurrent sound segregation 46

    2.4. Thesis IV: Two and three concurrent sound objects 47

3. Studies 48

    3.1. Study I: EEG signatures accompanying auditory figure-ground segregation 48

    3.2. Study II: Effects of multiple congruent cues on concurrent sound segregation during passive and active listening: An event-related potential (ERP) study 60

    3.3. Study III: Theta oscillations accompanying concurrent auditory stream segregation 74

    3.4. Study IV: Promoting the perception of two and three concurrent sound objects: an event-related potential study 85

4. General discussion 98

5. Conclusions and further directions 106

References 109

Acknowledgements

I am very grateful to my supervisor, Dr. István Winkler, who introduced me to auditory research and always has been a great source of inspiration, and who provided me with relentless help and advice. I am also very grateful to Dr. Alexandra Bendixen for her professional support. She has taught me most of what I know about EEG analysis and has helped me in most of my experimental work.

I wish to thank all my co-authors who contributed to the papers appearing in this thesis (in alphabetical order): Dr. Claude Alain, Dr. Gábor P. Háden, Dr. Barbara G. Shinn-Cunningham, Dr. Orsolya Szalárdy, Ágnes Szerafin, Dr. Brigitta Tóth, and Gábor Urbán.

Special thanks to Zsuzsanna D'Albini, Emese Várkonyi, Zsófia Zavecz, Csenge Török, and Yu He, who helped me with the data acquisition.

The experimental work reported in this thesis was supported by the Lendület project (LP2012-36/2012) awarded to István Winkler, and by the Erasmus Mundus Student Exchange Network in Auditory Cognitive Neuroscience awarded to myself.

I would also like to thank my family and friends for their love and support.

Finally, I would like to thank András Tasi for his love, help and support.

Glossary of abbreviations

ABR	Auditory brainstem response
AM	Amplitude modulation
ANOVA	Analysis of variance
ARN	Awareness related negativity
EEG	Electroencephalography/Electroencephalogram
ERP	Event-related brain potential
ERSP	Event-related spectral perturbation
FM	Frequency modulation
fMRI	Functional magnetic resonance imaging
ILD	Interaural level difference
ITD	Interaural time difference
MEG	Magnetoencephalography/Magnetoencephalogram
MMN	Mismatch negativity
ORN	Object-related negativity
TRF	Temporal response function

Abstract

The human auditory system continuously receives mixtures of sounds. This mixture needs to be parsed in order to arrive at information useful for behavioural goals, such as communication (termed auditory scene analysis). The most widely accepted theoretical framework of auditory scene analysis suggests that based on the type of cues they utilize, there are two main types of processes parsing the incoming sound: sequential and simultaneous sound segregation. Sequential sound segregation links together sounds that are separated over time, while simultaneous sound segregation groups sounds that occur at the same time. The object-related negativity (ORN) component of the event-related brain potentials (ERP) is elicited by the presence of more than one concurrent sound. When the manipulated sounds are task-relevant they also elicit the P400 ERP component.

This thesis is based on four electrophysiological studies investigating signatures of simultaneous sound segregation. The overall pattern of results suggests that the ORN reflects the global assessment of bottom-up grouping mechanisms regarding the presence of multiple concurrent objects, while P400 represents the outcome of the perceptual decision about the number of concurrent sound objects. These interpretations are compatible with the assumption of a grouping and a subsequent competition stage in auditory scene analysis. However, the results can also be explained by temporal coherence models suggesting that sound sources are segregated by the co-modulation of sets of sound features. In the latter case, ORN would reflect temporal coherence detection, whereas P400 would reflect the outcome of top-down biasing of feature selection.

Kivonat

A mindennapi életben folyamatosan hangok egyvelegével szembesülünk. A hallórendszerünknek szét kell tudnia választania ezeket a hangokat, hogy a viselkedéses célok számára (például a kommunikáció szempontjából) hasznos információt állítson elő (ezt nevezzük hallási jelenetelemzésnek). A legelfogadottabb hallási jelenetelemzési elméleti rendszer szerint két fő hangszétválasztó folyamat létezik az alapján, hogy milyen jelzőmozzanatokat használnak: a szekvenciális és a szimultán hangszétválasztás. A szekvenciális szétválasztási folyamat során azokat a hangokat csoportosítjuk egybe, amelyek időben egymást követik, míg a szimultán szétválasztási folyamat az időben egyszerre megjelenő hangcsoportokat különítik el. A tárgyhoz kötött negativitás (TKN) olyan elektromos agyi potenciál (EAP), mely akkor jelenik meg, ha egyszerre egynél több hallási tárgy van jelen a hallási jelenetben. Amennyiben a manipulált hangok relevánsak a feladat számára szignifikáns P400 komponens is megjelenik.

A tézis négy elektrofiziológiai tanulmányra épül, mely a szimultán hangszétválasztás agyi jeleit vizsgálja. Az eredmények azt mutatják, hogy a TKN az olyan „bottom-up" csoportosító mechanizmusok átfogó értékelését tükrözi, mely azt jelzi, hogy a hallási jelenetben egyszerre több mint egy hang van jelen, míg a P400 a jelen lévő tárgyak számáról való perceptuális döntés eredményét mutatja. Ezek az eredmények összeegyeztethetőek azzal a feltételezéssel, miszerint elsőként egy csoportosítási, majd egy erre épülő versengési fázis különíthető el a hallási jelenetelemzésben. Azonban az eredményeket az idői koherencia modell alapján is lehetséges magyarázni, mely szerint a hangforrások a hang tulajdonságainak közös modulációja alapján különíthetők el. Az utóbbi esetben a TKN az idői koherencia detekcióját, míg a P400 a tulajdonságok kiválasztásának „top-down" befolyásolását jelzi.

1. Introduction

In an everyday environment, we are constantly facing a myriad of sounds that are coming from different sound sources. Our auditory system has to be capable of making sense of the rich incoming stimuli and parse them in order to perceive the world in terms of meaningful objects. For example, imagine a party with loud music where several people are talking, glasses are clinking while you are having a conversation with a friend. This situation has been termed as the cocktail party effect (Cherry, 1953). In such a case, you need to be able to focus your attention on the talker's voice and filter out all the irrelevant stimuli in order to maintain the conversation. Creating auditory objects and streams has been investigated within the framework of auditory scene analysis introduced by Bregman (1990).

Bregman's theoretical framework states that there are several cues that can guide the parsing processes. These cues can be divided into two main groups (see, also Carlyon, 2004; Haykin and Chen, 2005; Snyder and Alain, 2007): 1) grouping together sound elements along time (horizontal, or sequential sound organization – grouping elements of the incoming sound mixture on the basis of their temporal/sequential relationship), 2) grouping together sound elements that occur at the same time, but differ in their spectral contents (vertical, or simultaneous, or concurrent sound organization – grouping elements of the incoming sound mixture on the basis of spectral structure). It is useful to distinguish these two types of grouping processes, because they are based on different acoustic cues. In everyday situations, the two types of grouping processes jointly separate auditory streams (coherent sound sequences likely originating from the same or multiple synchronized sources, such as several instruments carrying the same tune at the same time) from the rest of the acoustic background. The auditory stream appearing in the foreground is also termed „figure", and this auditory function is called "figure-ground segregation". Whereas grouping processes (both simultaneous and sequential) subserve the segregation of auditory streams and the detection

of sound patterns, figure-ground segregation goes one step further: it specifies one stream (separated on the basis of one or both types of grouping processes) as being selected for processing and the rest of the sound, which may be left undistinguished. Further, auditory streams can be regarded as perceptual objects, as they can enter cognitive operations, such as being selected, remembered, transposed, etc. (cf. Kubovy and Van Valkenburg, 2001; Griffiths and Warren, 2004; Winkler, Denham and Nelken, 2009).

It has been shown that many instantaneously available cues can be used by the concurrent sound segregation process, such as inharmonicity, source location separation, onset asynchrony, differences in amplitude and frequency modulations, etc. Combinations of some of these cues have also been tested. Concurrent sound segregation has been shown to have a direct event-related potential (ERP) component correlate: the object-related negativity (ORN) has been shown to follow the listener's perception when two concurrent sounds are perceived as compared to one (Alain et al., 2001). When listeners are instructed to distinguish stimuli including multiple concurrent sounds from single sound ORN is followed by another ERP component, the P400 (Alain et al., 2001).

The thesis focuses on the electrophysiological correlates of concurrent sound segregation with the following main questions: Does auditory figure-ground segregation evoke the electrophysiological correlates of concurrent sound segregation? How do the different combinations of cues affect the ORN component? Do ORN and/or P400 sum together the outputs of independent detectors for each cue, or do they reflect the overall readout of the auditory system's assessment of the likelihood of the presence of multiple concurrent sounds? In order to provide answers to these questions the electroencephalogram was measured while participants listened to auditory stimuli composed of multiple concurrent sounds. In *Study I*, we employed a recently developed stimulus paradigm for investigating the electrophysiological correlates of auditory figure-ground segregation. In *Study II*, the three

most well-known cues of concurrent sound segregation and their combinations were studied to investigate the relationship between ORNs (and P400s) elicited by auditory stimuli carrying multiple congruent cues of concurrent sound segregation and sounds carrying only a subset of these cues. In *Study III*, the event-related spectral perturbations of the same three cues were investigated. Finally, in *Study IV*, we tried to develop a stimulus paradigm promoting the perception of three auditory objects, and investigated the effects of the cues and their combinations on the ORN component.

The structure of the thesis is the following: *Section 1.1* begins with an overview of sequential and concurrent sound segregation describing the basic principles, theories and findings. *Section 1.2* gives a review about auditory object perception. This is followed by *Section 1.3* in which the electrophysiological correlates of concurrent sound segregation are discussed. In this section, the basic findings about ORN are listed, the P400 component is introduced, and a short introduction into neural oscillations is given. *Section 1.4* describes the auditory figure-ground segregation, and the previous findings. The *Synopsis and rationale of the theses* section summarizes the goals of the studies, followed by the articles in their published form (*Studies* section). This is followed by the *General discussion* which summarizes the findings of the papers placing them within the literature and in the *Conclusions and further directions* section future research directions are discussed.

Section 1.1. Auditory scene analysis

The problem of auditory scene analysis is best described by the analogy given by Bregman (1990) in his seminal book, *Auditory scene analysis (pp. 5-6)*: "Imagine that you are on the edge of a lake and a friend challenges you to play a game. The game is this: Your friend digs two narrow channels up from the side of the lake. Each is a few feet long and a few inches wide and they are spaced a few feet apart. Halfway up each one, your friend stretches a handkerchief and fastens it to the sides of the channel. As waves reach the side of the lake they travel up the channels and cause the two handkerchiefs to go into motion. You are allowed to look only at the handkerchiefs and from their motions to answer a series of questions: How many boats are there on the lake and where are they? Which is the most powerful one? Which one is closer? Is the wind blowing? Has any large object been dropped suddenly into the lake?"

This problem appears to be nigh impossible to solve, showing the difficulties of inverse solutions problems, but the analogy is really close to what our auditory system has to deal with. Here, the lake is the air surrounding us, the two channels are the two ear canals, and the handkerchiefs are the ear drums. The motion of the handkerchiefs is the vibrations caused by the atmospheric pressure changes arriving to the two ear drums, which is the only actual external information the auditory system has about the incoming sounds. However, we "know" a lot about the nature of sounds, in general (partly encoded in the structure of the human auditory system, partly learned during our life) and possibly also about the characteristics of the actual incoming sounds. This pre-existing knowledge is used during auditory scene analysis. Still, from this scarce information, the auditory system appears to be able to answer many questions about the auditory scene with remarkable reliability.

Bregman's theoretical framework (1990) comprises two processing stages. In the first stage, the sound input is grouped in parallel by different heuristic algorithms which embody the Gestalt principles of perception (Köhler, 1947). In this stage, possible sound groupings (which can also be termed putative auditory objects) are created. In the second stage of processing, the previously created groupings compete with each other and the winner of the competition appears in our perception. According to Bregman (1990) in the first processing stage, based on whether the knowledge employed is innate or learned, the sound input can be grouped in two different ways: by primitive and schema-based processes. Primitive mechanisms typically group sounds on the basis of simple stimulus properties, while schema-based mechanisms group parts of the input that match some previously learned relationships. The primitive mechanisms do not necessarily involve specific experience, whereas the schema-based mechanisms rely greatly on learned mechanisms, and thus depend on the listener's experience.

Based on the cues utilized, two types of processes have been distinguished by Bregman (1990): sequential and spectral grouping. Sequential grouping (also referred to as horizontal grouping) is the process of grouping together sound elements that have been emitted at different times by the same source while segregating them from sound elements generated by other sources. This process uses the time course of spectral changes as cues for grouping/segregation. When concurrent spectral components are grouped into a single sound, this fusion is referred to as spectral (simultaneous or vertical) grouping. This process groups together sound elements with different spectral contents that overlap in time, which are likely to have originated from the same source. The resulting complex is perceived as a single sound.

These processes - similarly to visual perception - operate in accordance with the rules of the Gestalt theory (presented in the following section). Although the Gestalt principles

provided the foundation for Bregman's auditory scene analysis theory, there are some differences in the views of the Gestaltists and Bregman: The members of the Gestalt school of psychology suggested that the grouping principles they have described originate from the laws of physics, describing a trend towards minimizing some property similar to the energy minimum principle. In contrast, Bregman regarded the Gestalt principles as heuristic algorithms that were selected through evolution as they contributed to the veridical perception of the world. Perhaps common grounds between these two views can be found when looking at the memory capacity needed to store sensory information. Recent findings (Peterson and Berryhill, 2013; Halberda, Simons and Wetherhold, submitted) suggest that processing algorithms based on Gestalt principles facilitate visual perception and affect the visual working memory (vWM), apparently increasing its capacity. These results support the notion that Gestalt principles may represent a minimization principle, i.e., minimization of the storage capacity required for describing sensory data (cf. Gordon, 2004, *pp. 45-51*).

1.1.1. Gestalt principles

The Gestalt school of psychology focused on the perception of the "unified whole" (e.g., Koffka, 1935; Köhler, 1947) contrasting the ideas of the structuralist notion of explaining perception (e.g., Titchener, 1901, referred by Gordon, 2004). It was based on the observation that one's perception often cannot be broken down to simple sensations. In other words: "The whole is different from the sum of the parts acting in isolation" (Gordon, 2004, *pp. 18*). The central idea behind Gestalt psychology is that by self-organizing rules, our perceptual system forms and manages complex perceptual entities (now termed perceptual objects). The Gestalt principles of grouping provide the algorithms by which object representations are extracted from the sensory input.

Principle of exclusive allocation. A form, a shape, or a silhouette is naturally perceived as figure while the area that surrounds it is perceived as ground or background. Figures are often perceived as more salient and have an object-like character whereas the ground may appear as an indistinct background. The principle of exclusive allocation can best be explained by Rubin's face-vase illusion (Rubin, 1915, Figure 1). One can interpret the picture as two symmetric profiles (the figure) in front of a white background or a vase (the figure) over a black background. When we perceive the picture in terms of two faces in the foreground, the edge between the faces is allocated to the face. In contrast, when we see the vase in the foreground, the edge belongs to it. Based on this and similar phenomena, Gestalt scientists suggested that a sensory element can only belong to one object at a time.

Figure 1. Rubin's face-vase illusion (Rubin, 1915), it can either be perceived as two symmetrical profile outlines facing each other or as a vase in front of a black background.

Similarity. Similar items tend to be grouped together. In audition, sounds of similar timbre or other characteristics are likely grouped together, such as in a musical piece, the

sound of the violins are typically grouped together and segregated from those of the harp, even if they are played in the same register.

Proximity. Sensory elements appearing close to each other tend to be grouped together. This principle is difficult to grasp in the auditory modality, as proximity can mean temporal proximity or spatial proximity of the sound sources. However, the spatial resolution of hearing is rather poor as compared to vision and the effects of spatial proximity are typically weaker than those of other cues – although, in some cases, the grouping may be helped by spatial proximity (Bőhm et al., 2013). Altogether, it is not entirely clear whether or not the proximity principle plays a crucial role in audition.

Closure. Our visual system tends to close broken contours, such as we experience a broken circle alike to a full circle. An auditory example can be experienced in the continuity illusion: when a loud wide-spectrum sound masks a softer tone we will not hear the soft tone. If, however, the softer tone can be heard both before and after the masker it will be perceived as if it continued throughout the mask, even if it is not physically present, only the mask contains the frequency of the soft tone. This phenomenon can also be viewed as an example of the "*old plus new heuristic*" (Bregman, 1990). The auditory system prefers the continuation of previously discovered sound groups over assuming that a new sound group has emerged. This means that only those sound elements are left to form a new group, which cannot be regarded as the continuation of a previously formed group. In the illusion, the auditory system represents the softer tone as being continued behind the masker, as opposed to treating the latter part of the soft tone as a new sound group.

Common fate. The common fate principle refers to correlated changes in features, for example temporal coherence (i.e., common onset), coherent correlations in modulations, or spatial trajectory. The principle is that if different parts of the spectrum change together they

most likely belong to the same sound, and therefore, they are segregated from other elements that change in a different way.

Good continuation. Oriented units or groups tend to be integrated into perceptual units if they are aligned. When there is an intersection in a figure between several objects we tend to perceive each of them as a single, uninterrupted object. Sharp and abrupt changes are not favoured by the grouping mechanisms.

Past experience principle. Our visual system tends to group together those elements that were often perceived together in the past experience of the observer. The Gestaltists believed this principle to be of secondary importance as the stimulus-based principles typically dominate over it.

1.1.2. An overview of the sequential sound segregation

Auditory stream segregation occurs when links are formed between temporally separate parts of the sound input and the sequence of linked sounds form a coherent sequence of sound (termed "stream") which is then experienced as a single pattern or melody. For example, when one listens to someone talk and a series of footsteps are heard during it, the footsteps are going to be connected and separate from the speech stream.

The most widely used paradigm for studying auditory stream segregation is the 'ABA-' paradigm, studied in length by Leo van Noorden (1975). Here, subjects hear a tone sequence of a repeating 'low-high-low-silence' pattern (low tone = A, high tone = B) and their task is to mark whether they hear a galloping ABA-silence sequence (termed the integrated percept) or the two frequencies split into separate streams and they either hear only the high (-B--) or the low (A-A-) tones repeating in the foreground (termed the segregated percept). This effect can

be influenced by parameters such as frequency separation of the two tones: the bigger the separation the more likely the subjects are to perceive the two frequencies as segregated. The presentation rate also impacts the perception: the slower the rate the more integrated percepts are reported, while at higher presentation rates people are more likely to hear the segregated alternative. These perceptual alternatives can be explained in terms of the rate of change. Figure 2 shows a diagram depicting the effects of tone repetition rate and pitch separation in the 'ABA-' paradigm on the likelihood of hearing one or two streams during the repetition of the pattern.

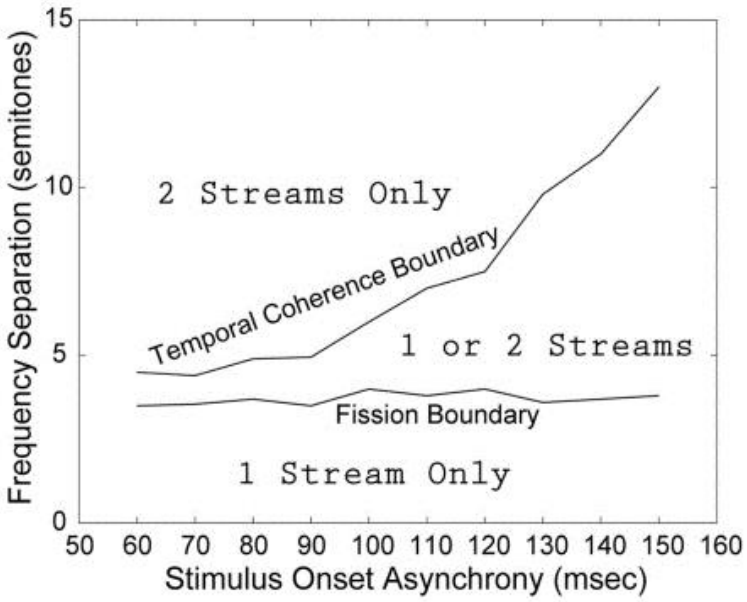

Figure 2. A reproduction of van Noorden's (1975) streaming diagram depicting perceptual regions as a function of the combination of tone onset to onset interval (termed "stimulus onset asynchrony", the inverse of the tone repetition rate) and pitch separation between the high and low tones. The fission boundary separates the regions in which only one stream can be heard and where the perception can be voluntarily switched between 1 or 2 streams, whereas above the temporal coherence boundary, listeners mainly hear two separate streams.

There are several cues that can induce sequential segregation such as the spatial location of the tones, spectral frequency, fundamental frequency, timbre, and amplitude difference, etc. (Moore & Gockel, 2002).

Because the studies included in the thesis focused on concurrent sound segregation, sequential sound segregation is not discussed in detail. The reader is referred to some reviews of sequential sound segregation (cf., Denham & Winkler 2015; Ciocca, 2008; Snyder & Alain, 2007).

1.1.3. An overview of concurrent sound segregation

A basic assumption of sound segregation is that before the properties of an auditory event can be established some decisions must be made with regards to which components of the input belongs to the event. Concurrent or horizontal/simultaneous sound segregation is the process of partitioning the set of concurrent sounds into distinct subsets and then placing them into separate streams where they are used to characterize the properties of distinct sound sources (Bregman, 1990). The question is how this partitioning is made? The basic principle is that if a group of components stem from the same source or physical event they will have characteristic relationships between them that are not likely to have occurred by chance, such as common timing or a harmonious frequency relationship. When concurrent sound segregation works on the basis of such physical properties it is considered a primitive grouping mechanism. However, for example, if one grows up listening to Western music, major and minor chords are learnt and thus the tones comprising them are automatically grouped. Thus there are also schema-based versions of concurrent grouping. Here I will list those relationships that are likely to occur between unrelated events, therefore reflecting the presence of multiple physical sources.

Yost (1991) described many physical variables that are likely to affect the segregation of auditory images: spectral separation, different intensity, inharmonicity, spatial separation, temporal separation, different temporal onsets and offsets, incoherent slow temporal modulation. In a real acoustic environment, many of these cues are simultaneously present and strengthen each other so that they contribute to the identification of individual sound sources.

Many early experiments were aimed at investigating judgements about the phenomenal experience of the number of sound sources that participants heard when listening to various test sounds, which will be listed in the following sections.

1.1.3.1. Inharmonicity

The most well-studied cue is the (in)harmonicity (or frequency periodicity) of the complex sounds. Periodic sounds (or harmonic sounds; e.g., such as produced by some musical instruments, e.g., piano) have frequency spectra which contain energy only at the fundamental frequency (corresponding to the repetition rate of the waveform) and at harmonics of that fundamental which are positive integer multiples of its frequency. The mistuned harmonic paradigm was first studied extensively by Moore, Glasberg, and Peters (1986). They presented stimuli half of which were harmonic complex tones and in the other half they mistuned the frequency of one harmonic by either adding or subtracting a certain percentage of its original frequency value. They asked listeners to judge whether they heard one or two sounds. They found that up to the sixth harmonic 2% of mistuning was enough for listeners to report two sounds, whereas for higher harmonics performance was not reliable. Hartmann and colleagues (1990) created another pitch-matching task where participants had to adjust the pitch of a pure tone to the mistuned harmonic's pitch and their results indicated that segregation increased over the range of 0.5-4%, irrespective of sign. They also

emphasized the role of spectral resolution: listeners lost the ability to segregate mistuned harmonics at higher frequencies where synchronous neural firing diminishes.

Also, decreasing signal duration was found to reduce the likelihood of reporting hearing the lower mistuned harmonic as a separate object. The results of subsequent studies generally confirmed this observation, but also showed that absolute frequency plays an important role, as the listeners' ability to match the frequency of a mistuned harmonic deteriorates with increasing frequency (Roberts and Brunstrom, 2001).

1.1.3.2. Harmonic sieve

The idea of a central harmonic template of pitch perception was first introduced by Goldstein and colleagues (Goldstein, 1973; Gerson and Goldstein, 1978) based on their findings of how a low pitch is derived from the spectrum of the complex by a central pattern-recognition mechanism, and this account has been extended to cases where more than one sound is heard concurrently (Duifhuis, Willems and Sluyter, 1982). These models propose that a complex tone's pitch corresponds to the fundamental frequency (*f0*) of a harmonic sieve that fits best to the distribution of resolved components. Moore and colleagues (1985, 1986) further proved that frequency components outside regions that are defined by the *f0* are excluded from pitch computation and are perceived as separate sound objects.

Roberts and Brunstrom (1998, 2001) provided further evidence for the harmonic template and proposed that the auditory system responds to a common pattern of equal spacing between components but it is only sensitive to this pattern over a limited range.

The general idea that spectral grouping depends on harmonicity per se has been questioned by Roberts and colleagues (Roberts and Bregman, 1991; Roberts and Bailey, 1993) who found that the main factor was regular spectral spacing. They based this conclusion on the study where they asked participants to hear out odd or even harmonics from

a complex tone that otherwise only contained odd harmonics. Participants performed generally better when they had to find the even harmonic and they reasoned that an even harmonic popped out because it violated the regular spectral pattern formed by the odd-harmonic base.

It is also likely that this harmonic template can be extended to other features, such as onset asynchrony, or frequency and amplitude modulation, however no direct investigation has been conducted in the matter.

1.1.3.3. Onset asynchrony

Components that start and stop at the same time are more likely to have come from the same source and tend to be grouped together. Several studies addressed this issue. For example, Bregman and Pinker (1978) played a sequence of a pair of more or less synchronous tones (A and B) and a preceding pure tone (C) the frequency of which was varied so that it could be close to the higher frequency tone (A) in the tone pair. They tested whether A, and B would fuse together if their onset is at the same time, or C will be grouped sequentially with A if they are close enough in frequency. The results showed a tendency of the pure tones to fuse together if their onsets are in synchrony, or very close to each other and the tendency of sequential organization to take over if the frequencies are close enough. The two sound organizations compete in order to determine the best perceptual decision of the input.

There is also ample evidence that onset asynchrony is an especially powerful cue for sound segregation, however for different tone durations different onset asynchrony is needed to make judgements about them (e.g., which tone came first). Pastore, Harris and Kaplan (1982) found that for 300 ms tones 15-20 ms is needed, whereas when the tone is only 100 ms long 10 ms is sufficient to make a reliable judgement.

Micheyl and colleagues (2013) used a paradigm where repeating target tones can be heard out from a multi-tone background when the targets form a separate stream (based on the paradigm of Kidd et al., 1994). Depending on the condition, the target tones were either temporally coherent or incoherent with the background and they were either harmonically or inharmonically related to the background. They found that harmonicity-based grouping was not sufficient to overcome segregation based on temporal incoherence; however, harmonicity can facilitate the segregation of target tones embedded in masker tones even when the targets and the background are temporally coherent.

1.1.3.4. Location difference

It has been long assumed that the components of a sound should be perceptually segregated according to their lateral position relative to the listener, because all the components of a single sound tend to originate from the same place. Some evidence, however, supports that two pure tones of different frequencies at different locations can be heard as simultaneous sounds, but they will only fuse if their frequency is very close (Perrott and Barry, 1969).

Localizing a sound in the azimuthal space depends on two main cues: interaural level difference (ILD, or interaural intensity difference - IID) and interaural time difference (ITD), for complex sounds the dominant cue is the ITD. There is some evidence that a difference in ITD imposed on one component of a complex can cause it to be separated from the others. For example, Kubovy, Cutting and McGuire (1974) mixed together eight sinusoids whose frequencies were the same as of a diatonic scale; when played diotically it is almost impossible to hear out sinusoids, but when the phase of one sinusoid is shifted between the two ears the sinusoid pops out in a different location. This can also be interpreted as the effect of the lack common fate instead of spatial proximity.

It also seems that lateralization is more effective when the experimental situation allows one to pay attention for an extended time to one direction rather than another. An example of this is a study by Deutsch (1979) who found that it was much harder to identify a tune when notes alternated between the ears than when they were all presented to the same ear.

Nevertheless, segregation based on location cues is very weak. It has been found that a single, resolved harmonic can be segregated from a vowel based on changing its onset time or tuning, but it cannot be segregated based on changing the ITD (Culling and Summerfield, 1995). Further, Buell and Hafter (1991) examined harmonic relation and ITD cues for low-frequency stimuli and they found that for localization purposes the interaural information is used. However, when the task is grouping the complex acoustic stimuli in space, harmonic structure is more influential than the commonality of spatial position.

1.1.3.5. Frequency and amplitude modulation

It has been a long-standing view in the history of modulation research that both coherent amplitude (AM) and frequency-modulation (FM) can help to bind together frequency components, but FM is not very suitable to be used to segregate sounds differentially on the basis of different FM rates or phase. It is also a basic fact that coherent FM is found naturally in sounds that share harmonicity (Darwin, Ciocca and Sandell, 1994).

Békésy (1963) found that with the help of simultaneous amplitude changes our auditory system helps us segregate or fuse sounds that come not only from different parts of the spectrum, but also from different source locations. If the listener is presented with different tones to the two ears no fusion happens, but if an in-phase AM of 8 Hz is added to both they will fuse.

Moore (1982) described in his book how common changes in a set of frequency components can cause those components to group together and to segregate from others that are changing in a different pattern. This subgroup of components stands out perceptually from the rest only by making them vary in a coherent way either in frequency, amplitude, or both. This group can then emerge as a figure against a stochastically stationary background.

McAdams (1984) studied FM in his doctoral thesis. In his experiment, one partial had one pattern of fluctuation and the rest of the partials had a different pattern. He found that when the depth of the fluctuation reached 0.5% subjects always said that the incoherently modulated sound had more than one source, but not for the coherently modulated sound. Moreover, the results also showed that for the first five harmonics the listeners could hear out a separate pitch, but with higher harmonics they did not. Subjects reported a choral effect as if you listen to a choir and you cannot distinguish individual voices, but based on the incoherence you know that more than one voice is involved.

Darwin, Ciocca and Sandell (1994) found that when they used a common FM (6 Hz) on all the harmonics and one harmonic was mistuned the effect of the mistuned harmonic was attenuated and the tolerance of the harmonic sieve was increased. They also found that the tolerance was not increased when they employed amplitude modulation with the rate of 17 Hz. In another experiment they found that there is an increase in sieve tolerance for common FM, but not AM at both 6 and 17 Hz modulation. Their results support the view that the auditory system is more likely to bind together sounds that are coherently modulated than those that are unmodulated (both AM and FM). However, in another study Darwin and Sandell (1995) found that when mistuning the fourth harmonic of a vowel and all the harmonics are given a coherent 6 Hz frequency modulation the coherent modulation does not prevent the mistuned harmonic from popping out of the complex.

For testing the effects of AM, Kubovy (1981) created a non-harmonically related complex and played it with equal intensity which made it sound like a complex tone. However, when the intensity of a partial was attenuated and then brought back to the original intensity the pitch of that partial became prominent. The auditory system seemed to be sensitive to differences in the amplitude pattern for different partials and it was possible to use them to segregate them.

Taken together, the evidence shows that inharmonicity, onset asynchrony, different source location, different AM and FM rates help segregate concurrent sounds, whereas sounds belonging to the same harmonic template, sounds emanating from the same, or very close source locations, and sounds with same modulation patterns are integrated into one sound object. Bendixen and colleagues (2015) found that the brain of newborn infants is sensitive to some cues of concurrent segregation (at least to mistuning and delaying a partial). Thus humans are capable of segregating concurrent sounds by detecting harmonic relation between co-occurring sounds from birth, apparently without attending to the sounds. However, Alain (2007) found that listeners' ability to segregate concurrent vowels improves with training. Thus, although these forms of concurrent sound segregation may be regarded as "primitive" process (see Bregman, 1990), they can be extended to different types of stimuli by experience.

1.1.4. An alternative to sequential/concurrent sound grouping

Shamma and colleagues (Shamma, Elhiali and Micheyl, 2011; Krishnan, Elhiali and Shamma, 2014; Lu, Xu, Yin, Oxenham, Fritz, & Shamma, 2017) suggested that the principle of grouping by temporal coherence provides an explanation of both sequential and concurrent grouping phenomena. Indeed, the notion of coherence appears to solve the binding problem (i.e., the problem of associating different sound features with the correct sound source) in

auditory scene analysis and it can also link temporally distinct sounds. This hypothesis is an extension of the common fate principle: acoustic features characterizing a given source are present when the source is active and absent when inactive and different sound sources are unlikely to fluctuate with the same temporal schedule. However, temporal coherence does not explain the effects of predictable patterns (such as familiar melodies) on sequential stream segregation, or some of the concurrent segregation phenomena, such as those listed for harmonic sieves (see *Section 1.1.3.2*).

Section 1.2. Auditory object perception

When we talk about object perception we have to ask first why we need to parse the incoming information into objects? There are two sides of object perception:

1) We parse the input into objects because the world can be interpreted in terms of objects (i.e., there are parts of the input that can be segregated from the rest of the input which behaves coherently) and because perception aids us in reaching our goals and avoid possible dangers. Therefore we must be able to identify the parts of our environment the presence (e.g., food) or behaviour (e.g., predator) of which is essential for us. This notion represents the functionalist view of perception (e.g., Brunswik, 1955, referred by Gordon, 2004). That is, for interacting with the world (achieving behavioural goals) one needs to have a representation of the potential targets of these interactions.
2) The object-representation also serves as the unit of mental operations (such as retrieving knowledge about the object in order to predict its behaviour). Perception without mental representations cannot support flexible interactions with the world, such as humans and many animals are capable of.

This duality of perceptual object representations is especially clear in the auditory modality (Kubovy and Van Valkenburg, 2001). Some of the auditory perceptual objects describe sound sources (such as a person's voice), while others describe sound patterns, which can be manipulated independently of the object producing them (such as a melody).

Objects are the building blocks of experience, although the concept of an 'object' has generally been used as a visuocentric terminology. When we refer to what we perceive we refer to the environmental object that produced the physical pattern of stimulation at the receptor level, thus we say that we see a book and not the array of points or edges, or we hear

a bell not the complex of partials. Therefore in vision, the study of object perception concentrates on the details of further processing of the peripheral representations, whereas in hearing no direct peripheral representation is given, all sensory information is compressed into a pair of acoustical pressure waveforms. To be able to detect an acoustic event from a single sound source one must be capable of integrating information at multiple time scales and to extract specific patterns from the background that contribute to the input. Another important difference between vision and audition is that the objects in vision are usually not transparent therefore they cover the objects behind them, whereas sounds are always transparent. A sound never obscures another sound, but the two waves are added together and this mixture provides the input to the ear.

The definition of auditory object is an issue of controversy to this day. In this section, a list of concepts will be provided that attempt to define auditory objects.

Handel (1988) states that both the visual and auditory worlds are temporal and spatial, as for both senses the events and objects are perceived in a framework created by spatial and temporal changes. The framework of space and time acts to segregate events and achieve figure-ground segregation. In both the auditory and visual world, signals overlap spatially and temporally and to perceive objects it is essential that the intermixed signals are parsed into segregated events.

Auditory objects are acoustical waveforms that evoke a mental reference to the source of the waveform; we hear the objects themselves and we are generally unaware of its temporal and spectral structure (Wightman and Jenison, 1995), even though a reference to a concrete identifiable object is not a necessary condition. Wightman and Jenison (1995) distinguish between two forms of auditory objects, the concrete object representations, which are formed

from sounds that were emitted by actual objects in the environment (e.g., an orchestra), and the abstract objects that often do not correspond to actual objects (e.g., a melody).

Darwin and Hukin (1999) claim - with regards to speech sounds - that auditory objects are the results of non-spatial grouping processes. Once the object is formed one can direct their attention to it.

Alain and Arnott (2000) define an auditory object as the percept of a group of sounds, as a coherent whole coming from a single source. They clarify the terms auditory event and auditory object. An auditory object refers to the perception of a sound source and its behaviour over time, while auditory event refers to the dimension of hearing a sound occurring at a time point in a particular space and has particular attributes and it can be a part of a larger entity, such as an auditory object. The role of selective attention is to connect the elements of the auditory input at the focus of attention in order to create a perceptual object (Treisman and Gelade, 1980).

A perceptual object is what is susceptible to figure-ground segregation according to Kubovy and Van Valkenburg (2001). These authors proposed the following view on the relation between early processing, grouping, figure-ground segregation, and attention: during early processing elements are formed that require grouping which occurs based on the principles of Gestalt theory. These principles form perceptual organizations that are putative objects from which attention selects one to become figure while the rest is assumed to be background. This view is compatible with those of Bregman (1990), who also assumes that the processes of grouping and feature integration are pre-attentive (as opposed to the view of Treisman - Treisman and Gelade, 1980; Treisman, 1998, but in agreement to those of Duncan and colleagues – Duncan and Humphreys, 1989, 1992). Thus, whereas for example Handel (1988) defines auditory objects based on analogies between visual and auditory object

features, Kubovy and Van Valkenburg (2001) build their definition on analogies between the mental operations.

Kubovy and Van Valkenburg (2001) also introduced the concept of cross-modal objecthood focusing on the similarities between modalities instead of the differences as was the traditional perspective. Our auditory system is dealing with vibrating audible sources and not the surfaces that reflect the sound as does our visual system with objects and the light reflected on them. They also propose that the auditory system has two parallel streams of processing, like vision (the 'what' and 'where' subsystems). Here, the 'what' subsystem deals with auditory objects, whereas the 'where' system provides information to the visual localization as space is not a main aspect of forming auditory objects. The spatial localization of sounds in humans is very malleable. When conflicting visual information is provided the latter influences the auditory information, however, where vision is weakest auditory localization can be of great help. An excellent example for this phenomenon is the ventriloquism effect (Jack and Thurlow, 1973): this phenomenon occurs when the speech sounds are perceived as though coming from a direction other than their true direction due to the influence of a visual stimulus from an apparent speaker. The effect can be explained as a phenomenon in which the sensory modality with the higher acuity dominates over the modalities with the lower acuity (Warren, Welch and McCarthy, 1981).

Griffiths and Warren (2004) describe an object as something that can be considered a perceptual entity that depends on the brain mechanisms available to represent and analyse sensory information. They proposed four principles as the basis for studying object analysis in any sensory domain: "1) object analysis involves information analysis corresponding to things in the sensory world; 2) object analysis involves the separation of information related to the object and related to the rest of the world; 3) object analysis involves the abstraction of sensory information so that information can be used to generalize between particular sensory

experiences in any sensory domain; 4) object analysis involves generalization between the senses" (Griffiths and Warren, 2004, *pp. 887*). The first three principles are regarded as essentials for auditory object analysis, but the fourth principle is debatable: the correspondence between the other senses is less clear.

Based on the first principle, auditory object analysis should be based on the analysis of their sources. On the output level this is often the case, however, on the level of the representation of the input this distinction is less clear: an object representation corresponds to a particular combination of sound source and event information. Griffiths and Warren (2004) claim that auditory objects can be categorized in more than one way (if a person talks a certain sound can be categorized as the voice of the individual or as a vowel), e.g., through generic auditory pattern analysis and then perceptual categorization. A definition of auditory object that is based on the sources implies that the auditory object analysis depends completely on *a priori* knowledge whereas in everyday situations we can perceive novel sounds as distinct auditory objects.

Based on the second principle, auditory objects should have perceptual boundaries. These boundaries are not so easily describable in hearing: a sound can be described in temporal and spectral domain and the sound envelope in either domain alone can form an object boundary. Kubovy and Van Valkenburg (2001) suggest that edges can be used to define auditory objects in frequency-time space, but according to this definition, a formant of a vowel can be regarded as a separate object. If a formant is regarded as an object then what is not regarded as one? In response to this issue, Griffiths and Warren (2004) claim that this is not a problem for the auditory system, as it analyses all sounds as if they were objects or combinations of objects.

The third principle describes that auditory object analysis involves the abstraction of information that is independent of the specific sensory representation. Object constancy can be used to extract invariant characteristics that are used to define an auditory object. An example of object constancy is depicted by our ability to recognize a familiar voice regardless of loudness or pitch.

As for the fourth principle, there is still a debate whether object analysis depends on any correspondence between different senses in the terms of the object information they process (e.g., the ventriloquism effect mentioned earlier).

Taken together, Griffith and Warren (2004) define an auditory object as an acoustic experience that produces a two-dimensional image with the dimensions of time and frequency.

O'Callaghan (2008) talks about intentional objects that may not include objects we perceive. "If an intentional object of a perceptual state is something that state concerns or represents, or at which it is directed, then the intentional objects of a perceptual state might include things apart from ordinary or material objects." (O'Callaghan, 2008, *pp. 817-818*) Therefore, auditory intentional objects can include: sounds, instances of audible properties, and relations (e.g., pitches, octaves, etc.). This notion of auditory object is more similar to that of a 'direct object'. O'Callaghan claims that auditory objects are perceptual objects which capture the critical aspects of how perceptual experience is organized. Perceptual objects are built on the notion of perceptible individuals which bear features, persist from one moment in time to the next, survive certain changes but not others, have boundaries, its parts are cohesive, and occlusion and masking do not disrupt their identification.

Winkler, Denham and Nelken (2009) give an overview about what characteristics a perceptual object should fulfil. First, an object representation must span multiple sensory

events as in a natural environment no sounds appear in isolation. Second, an object should be described by the combination of its features. Third, an object representation should describe which part of the acoustic input belongs to the object, because an object is a unit which is separable from other objects. Fourth, an object representation should generalize across the different ways the same object appears (i.e., perceptual invariance), as an object can appear quite variable to our senses. Fifth, an object representation should predict parts of the object for which the input is missing. Their main claim is that predictive (generative) regularity representations fit all these criteria. Therefore, they propose that the auditory regularities serve as perceptual objects. They hypothesize that the auditory objects are characterized in the brain by predictive rules that link them together into a unit which might be suitable for other modalities as well.

According to Bizley and Cohen (2013), auditory objects are the computational result of the auditory system's capacity to detect, extract, segregate, and group spectrotemporal regularities in the acoustic input. They also add that auditory objects have several general characteristics which are the following: 1) they are the consequences of actions or events emitted by or from things; 2) an auditory object can span several acoustic events that unfold over time, thus forming a stream; 3) the auditory system is capable of parsing the soundscape into the constituent objects based on the object's spectrotemporal properties that make it separable from other objects; 4) the listener can readily describe an auditory object by the combination of its features; 5) auditory object recognition is invariant to changes to its spectrotemporal features which result from the context the object is perceived in; 6) the auditory system expects object representations to predict parts of the object for which there is no currently available input.

Auditory information has been proposed to be processed similarly to the visual system in a 'what' (non-spatial aspects) and 'where' (spatial aspects of auditory processing)

subsystem (e.g., Rauschecker, 1997; Kubovy and Van Valkenburg, 2001). However, more recent evidence found that some object processing occurs in the dorsal pathway, whereas spatial information about the object has been found in the ventral stream (e.g., Bizley et al., 2009; Miller and Recanzone, 2009). These results suggest that a model of parallel hierarchical processing is too simplistic, and a mixture of the auditory information (spatial or non-spatial) might be more useful to achieve a consistent perceptual representation.

To sum up, an object is the focus of perception and the basis of cognition as we make sense of the world in terms of objects. Once objects are formed, they enable us to detect dangers and to perform goal-directed behaviour. In hearing, elements of the sound input are grouped together (possibly through processes employing the Gestalt principles) creating putative objects (or proto-objects). Once these proto-objects are formed attention can be directed to them, figure-ground segregation can be performed, and one can examine and react to them.

Section 1.3. Electrophysiological correlates of concurrent sound segregation

1.3.1. Electrophysiology and event-related potentials in general

Electroencephalography (EEG) is the method of recording electrical activity of the brain. The derivatives of the EEG technique include several measures, for example event-related potentials (ERPs) and neural oscillations, which will be introduced below. An event-related potential is the brain's large-scale electrical response arising as a result of a specific sensory, cognitive, or motor event. The study of event-related oscillations is based on the notion that information processing operates continuously, resulting in oscillatory electrical activity which is modulated by sensory, cognitive, and motor events.

The EEG comprises the electrical correlates of a very large number of simultaneously ongoing processes. Thus the ERP is embedded in this activity, making it invisible to the naked eye. In order to extract an ERP response, the event evoking it has to be repeated several times and the event-related activity needs to be extracted from the concurrent activity. Assuming that the non-event-related activity is independent of the event-related one, averaging across the EEG segments following the repeated events with the event onset as a reference reduces the amplitude of the brain activity unrelated to the event, while the common part of the activity time-locked to the event remains unchanged. There are also other methods of extracting ERPs from the EEG (such as the independent component analysis), which rely on somewhat different assumptions regarding the relation between the event-related and the non-event-related brain activity.

EEG techniques have the advantage that they are noninvasive and allow excellent temporal resolution as compared to some other imaging techniques. ERPs are valuable for investigating the time course of the processing of a stimulus providing a continuous measure

of the response. ERSPs can complement this by adding information about the event-related brain dynamics of the EEG spectrum induced by the stimuli.

ERP waveforms consist of a series of positive and negative voltage deflections related to underlying components. Some ERP components are referred to by an acronym (e.g., object-related negativity – ORN or mismatch negativity - MMN), whereas most components are referred to by a letter (N/P) reflecting the polarity followed by a number that either indicates the latency in milliseconds or the ordinal position of the component within the waveform (e.g., P1 - the first positive going peak in the waveform). ERPs can also be classified based on the response latency: early (0-10 ms), middle (10-50 ms) and long (50-500+ ms) (Davis, 1976; Winkler, Denham and Escera, 2013). Early components include receptor potentials from the cochlea and neurogenic responses arriving from the auditory nerve and low midbrain structures. Together, these are termed the auditory brainstem response (ABR). Generators of the middle latency responses (MLR) cannot be mapped as precisely as ABRs, but the thalamus and the cortex are the most likely contributors to the observable responses. Long latency responses have cortical generators, and they are much slower (composed of lower frequencies) than the earlier responses. Long latency responses (LLR) are highly dependent on stimulus type and may differ greatly in morphology and timing (Kraus and Nicol, 2009).

In addition to the peak latency based categorization, there are two other distinctions used for characterizing ERPs: a) ERPs can be exogenous if they are elicited by each event without respect to its relation with its acoustic context, the person's tasks or knowledge (ABRs, MLRs, and some of the LLRs are exogenous); an ERP component is endogenous if there is a relationship between the event and other events or some aspects of the person's mental state (there are many different endogenous LLRs); b) an ERP is considered "active" if it is elicited only when the person performs some explicit task involving the eliciting event,

and we talk about "passive" ERPs when the ERP response is elicited regardless of the person's task (Winkler et al., 2013).

1.3.2. Auditory event-related potentials

Long-latency auditory event-related potentials (ERPs) are typically characterized by a P1-N1-P2 potential complex peaking at about 50, 80-100, and 200 ms, respectively (Näätänen and Picton, 1987; Woods, 1995), which is related to signal detection and is present when a transient auditory stimulus is audible. The P1 was initially thought to reflect neural activity involved in extracting features (Näätänen and Winkler, 1999), but some more recent evidence connects it to the separation of auditory streams (Gutschalk et al., 2005; Snyder et al., 2006). P1 is generated in primary auditory cortex (Huotilainen et al., 1998; Reite et al., 1988) and it is often used as a landmark for localizing other auditory ERP components. The N1 component is elicited by abrupt changes in sound energy such as sound onsets or offsets (Näätänen and Picton, 1987). N1 peaks around 100 ms after the event with negative polarity. N1 is the most studied component as it basically can be related to any auditory processing steps (Winkler et al., 2013). The N1's generator and subcomponent structure is quite complex; the subcomponent that is most tightly related to the auditory processing is located in secondary auditory areas (Godey et al., 2001). Not as much is known about the P2 component; it peaks between 175 and 200 ms from the event with positive polarity and its generators are anterior to those of the N1 generators in the secondary auditory areas (Mäkelä et al., 1988). Based on its neural origin, Tremblay and Kraus (2002) have suggested that P2 is generated by a pre-attentive alerting mechanism. The P1-N1-P2 complex is often followed by the N2 component, which is a negative wave that peaks around 200-350 ms. The N2 component can reflect mismatch between the current sound and the regularities extracted from previous sounds, but it can also be elicited in paradigms of cognitive control, novelty or sequential matching as well (Folstein and Van Patten, 2008). The N2 component can be further divided

into three sub-components: the N2a or mismatch negativity (MMN), N2b and N2c. MMN is elicited when a sound violates the regular pattern in a sequence of sounds (Näätänen, Gaillard and Mäntysalo, 1978). It has a fronto-central peak and appears with inverted polarity over inferior temporal locations (i.e., the other side of the Sylvian fissure), as its main generator lies within auditory cortex. The repetition of the pattern leads to the formation of a prediction model which enables the brain to form predictions about what stimulus comes next. When this prediction is violated, the MMN is elicited (Winkler, 2007). If the deviants are task-relevant, further N2 effects are observed which are largest over central sites for auditory and over posterior sites for visual stimuli, and are labelled N2b and N2c, respectively (Simson, Vaughan and Ritter, 1977). There are several paradigms (e.g., go-no-go or stop signal paradigm) in which an N2b can be observed and it follows and overlaps MMN if the stimuli are attended. It is likely that there are several generators that contribute to the N2b. N2c is also elicited by task-relevant targets, but the functional significance of it is not yet clear (Luck, 2005). Target identification is also associated with another positive wave, the P300 or P3b (Hillyard, Squires, Bauer, and Lindsay, 1971), which peaks between 250 and 600 ms post-stimulus. It has two subcomponents, the P3a is elicited by rare task-irrelevant sounds or by novel stimuli, and it is also termed novelty-P3, whereas the P3b is elicited by target sounds (Polich, 2007).

While ERP analysis provides ample information about the temporal characteristics of stimulus processing, it does not offer information about the large-scale functional brain networks supporting information processing functions. The study of EEG oscillation can help in this respect. Oscillations are rhythmic or repetitive neural activity in the central nervous system. The study of oscillations traces back to Berger (1929) who first observed the dominant oscillations of approximately 10 Hz recorded from the human scalp. Oscillatory activity can be induced or evoked (Galambos, 1992). Induced oscillatory activity is correlated

with the experimental conditions, but it is not strictly phase-locked to the stimulus onset, while evoked oscillatory activity is phase-locked to the onset of an experimental condition across trials. Brain oscillations follow rhythmic shifting of neural activity over spatial and temporal scales (Buzsáki and Draguhn, 2004). Oscillations are characterized by their frequency, amplitude, and phase: the amplitude of EEG oscillations is typically between 0 and 10 μV, and the phase ranges between 0 and 2π (in the case of induced activity, there is no absolute 0 phase as it is not time-locked to the stimulus). The brain oscillatory theory (Basar, Basar-Eroglu, Karakas, and Schürmann, 1999; Buzsáki and Draguhn, 2004) suggests that different frequency bands are connected to the neural activity of distinct cell assemblies and thus, distinct neural and cognitive functions can be studied. It is assumed that when changes occur in the oscillation amplitudes the extent of neural activity involved in the functional process changes.

1.3.2.1. The object-related negativity (ORN)

The first group that studied the neural patterns of concurrent auditory object perception was Alain and colleagues (2001). They employed the paradigm of Moore, Glasberg and Peters (1986) where a complex tone is presented and the complex contains a mistuned harmonic: when the complex with the partial that is mistuned by at least 3% is presented the participants report hearing two separate sounds. They found that the stimuli elicited a large N1-P2 complex over the midline frontal and central electrodes and this was followed by a late positive peak (P3b) over the parietal regions. The negative difference waveform between mistuned and in-tune tone complexes was termed the object-related negativity (ORN) as it followed the perception of more than one concurrent objects being present in the auditory scene. The ORN component commences at about 100 ms from stimulus onset, and peaks about 180 ms. The maximum of its amplitude is over frontocentral electrode sites and it inverts polarity at inferior temporal sites which is consistent with

generators in the medial *planum temporale*. Alain and colleagues (2001) also found that the ORN generation was somewhat varied as a function of the attention manipulation. They claimed that ORN was a subcomponent of the N2 wave, reflecting an automatic mismatch process between the mistuned harmonic and the one expected on the basis of the template derived from the fully harmonic stimulus (Alain et al., 2001).

The presence and amplitude of ORN is correlated with manipulations that typically lead to listeners reporting two concurrent sound sources (Alain et al., 2001; Alain, Theunissen, Chevalier, Batty, and Taylor, 2003; Alain and McDonald, 2007; McDonald and Alain, 2005). The results of Alain and colleagues (2001) also suggested that a widely distributed network is involved in distinguishing simultaneous auditory objects which includes the primary and secondary auditory cortex, the medial temporal lobe, and posterior association auditory cortices.

The magnetoencephalographic counterpart of the ORN has also been investigated. Hiraumi and colleagues (2005) found right hemisphere dominance in processing the mistuned harmonics: the N100m response was significantly larger for the mistuned sound than for the harmonic sound and also the peak latency was significantly longer. Their findings suggest that the right hemisphere plays a more important role in detecting a mistuned partial than the left.

Johnson and Hautus (2010) used the dichotic pitch paradigm with both ITD and ILD and found that these cues are first processed separately by distinct neuronal populations indexed by the N100m (also showing distinct hemispheric patterns), while in a later processing stage of auditory scene analysis information common to the two cues is extracted and merged into a common code and this step is indexed by the ORNm.

Arnott, Bardouille, Ross, and Alain (2011) used a paradigm that promoted the perception of multiple auditory objects and found that the ORNm was associated with

bilateral auditory cortex sources that were distinct from those of the P1m, N1m, and P2m (associated with the sound onset). This supports the view that the detection of multiple auditory objects is associated with generators located in the auditory cortex and that these neural populations are distinct from the long latency evoked responses sensitive to sound onset. Previous studies have shown that ORN can be elicited by different cues, such as inharmonicity (Alain et al., 2001, 2002; Bendixen, Jones, Klump, and Winkler, 2010), onset asynchrony (Lipp, Kitterick, Summerfield, Bailey, and Paul-Jordanov, 2010; Weise, Schröger and Bendixen, 2012), dichotic pitch (Johnson, Hautus and Clapp, 2003; Hautus and Johnson, 2005; Hautus, Johnson and Colling, 2009), separation in the fundamental frequency of speech sounds (Alain, Reinke, He, Wang, and Lobaugh, 2005; Snyder and Alain, 2005), and simulated echo (Sanders, Joh, Keen, and Freyman, 2008; Sanders, Zobel, Freyman, and Keen, 2008). Further reports have shown the ORN being elicited by the combination of some of the above cues, e.g., inharmonicity and location difference (McDonald and Alain, 2005), or inharmonicity and onset asynchrony (Weise et al., 2012).

Bendixen and colleagues (2010) studied the probability dependence of the ORN component. They used a paradigm where not only concurrent, but sequential cues were used. They found that ORN was consistently elicited by the presence of multiple objects, although its amplitude was modulated by the probability: ORN amplitude decreased with increasing probability and its scalp distribution shifted from bilateral towards unilateral (left) activation pattern.

Studies showing that ORN can be elicited by cues other than those violating some harmonic template suggest that the template underlying ORN also includes information about the timing and source location of the partials of complex sounds. This template might be similar to what is described in *Section 1.1.3.2*. The harmonic sieve model proposes that the pitch of a complex tone corresponds to the fundamental frequency's harmonics that are

multiple integers of it (Goldstein, 1973; Gerson and Goldstein, 1978). Similarly to the harmonic sieve, it is possible to imagine that each harmonic of the complex tone fits into a hole of a sieve that represents a different feature, for example, onset and offset times, AM, or FM. ORN might accumulate evidence from these sieves one-by-one, and thus would represent a cue-based evaluation of the incoming auditory stimulus. If this was the case, ORN would sum together the outputs of each detector or sieve, the amplitude of the ORN elicited by multiple cues together would be as large as the sum of the output of each detector (each ORN elicited by a single cue). Alternatively, ORN might reflect the integration of the sound segregation cues and once a threshold is reached (which might be driven by the strongest, most salient cue), ORN signals the presence of multiple sound objects. Thus ORN would represent the overall assessment of the likelihood of more than one concurrent sound objects being encountered. If this was true, the amplitude of the ORN would be smaller for multiple cues applied together than the sum of the ORNs elicited by each cue alone. Testing this question was one of the primary goals of the studies of the current thesis.

1.3.2.2. P400

In the mistuned harmonic paradigm, Alain and colleagues (2001) first described the P400 component as a subtype of the P3b component, which was elicited when more than one concurrent sound object was present in the auditory scene and subjects were asked to listen to the stimuli. Even though ORN can be elicited in both passive and active listening situations (Alain et al., 2001, 2002; Alain, 2007) and its amplitude is independent of task demands (Alain and Izenberg, 2003), P400 can only be elicited in active listening situations, when the participants are asked to pay attention to the sounds and to decide whether they heard one or more concurrent sounds. Whereas ORN is assumed to reflect an automatic process of detecting the difference between the physical features (e.g., frequency) extracted from the incoming stimulus and a template of the complex sound (e.g., based on its fundamental

frequency), P400 appears to reflect a controlled process that uses prior knowledge to extract meaning from the incoming auditory information (Alain et al., 2002; Hautus and Johnson, 2005; Johnson et al., 2007).

The P400 is a widely distributed component that peaks around 400 ms post-stimulus, it reaches its maximum over central and posterior regions, and inverts polarity at the mastoids. The amplitude of the P400, similarly to ORN, correlates with the likelihood of perceiving two concurrent sound objects compared to one (Alain et al., 2001, 2002; Hautus and Johnson, 2005), but P400 does not follow the ORN in an obligatory manner (Johnson, Hautus, Duff, and Clapp, 2007). P400 has also been shown to be sensitive to sound duration: its amplitude was smaller for long (1000 ms) than middle (400 ms) or short (100 ms) duration sounds (Alain, Schuler and McDonald, 2001). This may be due to that the P400 amplitude was superimposed by the offset response elicited by the end of the stimulus in Alain and colleagues' (2001) study. The scalp topography shows active generators in the medial temporal lobe and posterior auditory association cortices. P400 most likely reflects top-down attentional processes as described by Bregman (1990).

Johnson and colleagues (2007) proposed that P400 is influenced by the task context. In their study, they used two different tasks: In the detection task, participants were to indicate whether they heard dichotic pitch or a control stimulus, whereas in the localization task, only dichotic pitch stimuli were presented and participants were to decide where the sound was located. In the latter case, no P400 was elicited (Johnson et al., 2007). Johnson and Hautus (2010) suggested that P400 indexes a relatively higher level and more controlled processes than ORN of the auditory scene analysis.

1.3.3. EEG oscillations

So far only a few studies have examined the neural oscillations in connection with concurrent sound segregation; here we review the most significant findings regarding auditory perception so far.

Weisz and colleagues (2011) showed that there is a pronounced alpha-like resting oscillatory activity in the auditory system. However this activity is of relatively low amplitude and it is often masked by non-auditory alpha generators. Following stimulations with sounds, a marked desynchronization can be observed between 6 and 12 Hz, which can be localised to auditory cortex. The authors suggest that this attentional alpha modulation may be involved in modulating the exhibitory-inhibitory balance across sensory modalities. Furthermore, Klimesch and colleagues (2007) found alpha suppression as a consequence of a word or other linguistic stimulation.

Most studies have investigated the auditory oddball paradigm. For example, Yordanova and colleagues (2001) showed that the decrease of induced alpha power is related to attention. They linked the above event-related desynchronization to the amplitude and latency of the auditory target P3b, which indicates that it is an attentional process. Fell and colleagues (1997) found that evoked gamma activity increased after stimulation in an auditory oddball paradigm, whereas induced activity was reduced. This indicates that the stimulation modulated the phase rather than the amplitude of the gamma activity. It has also been found that unattended deviants elicit larger induced gamma activity during the MMN interval than standard stimuli, whereas in the P3 interval, attended deviants elicited less induced gamma activity than the standard stimuli (Marshall et al., 1996; Kaiser et al., 2000).

In connection with auditory object formation, Bertrand and Tallon-Baudry (2000) suggested that induced gamma oscillatory activity may be involved in the construction of

objects and it possibly plays an important role in auditory grouping. Another question relevant to the current topic was studied by Bidet-Caulet and Bertrand (2009), who investigated the segregation of two concurrent auditory objects in patients with pharmacologically resistant partial epilepsy implanted with depth electrodes in the temporal cortex by manipulating sound onset asynchrony. These authors found that induced gamma oscillatory activity was enhanced when onset asynchrony was applied to the sound. However, so far not much evidence has been gathered as to how the oscillatory activity reflects the various cues promoting the perception of concurrent sounds.

Understanding how the brain solves the auditory scene analysis problem is a major goal in the field of auditory neuroscience. Since Bregman's (1990) book, many studies have dealt with auditory scene analysis and investigated behavioral measures related to it. However, only a handful of neurophysiological studies have addressed the concurrent grouping of auditory stimuli, although they can help to address issues, which are less accessible to behavioral measures, such as the time course of parsing the acoustic signal when multiple auditory objects can be perceived or the processes operating in the absence of attention focused on the sounds. Recent advances in neuroanatomical and physiological knowledge also enable us to investigate the underlying mechanisms of auditory scene analysis.

Section 1.4. Auditory figure-ground segregation

Auditory figure-ground discrimination is the ability to selectively abstract certain salient stimuli from the multitude of less relevant ones (Lerea, 1961). What constitutes the figure or the background in hearing is not easily definable as sounds are transparent thus they do not block one another; not when one sound is emitted by a source closer to the ear than the other. The salience of some acoustic feature determines the most likely figure for the listener. Figure-ground segregation involves at least three processes: 1) grouping simultaneous figure components from across the spectral array; 2) grouping the figure components over time; 3) after the completion of grouping, separating the figure from the rest of the acoustic input, i.e., deciding the group that attention can be directed to (i.e., the figure; Teki, Chait, Kumar, von Kriegstein, and Griffiths, 2011). In a complex scene, the auditory system utilizes sequential and simultaneous cues at the same time, which can be studied by the use of so-called tone clouds. A tone cloud is a series of sounds that consist of a large number of pure tones with random frequencies. Such stimuli have been used in informational masking paradigms to measure the effect of auditory stream segregation on detecting the repetition of a tone within a protected frequency zone of the cloud (Kidd et al., 1994; Elhiali et al., 2009; Akram et al., 2014). The tone cloud makes it possible to create a target pattern within the series of sounds for detection (Kumar et al., 2014; Barascud et al., 2016). Most recently, Teki and colleagues developed a new method based on the work of Kidd, et al. (1994), and Micheyl, Shamma and Oxenham (2007), which they called "stochastic figure-ground" stimulus. This stimulus incorporates stochastic variation of the signal components both in time and frequency. It comprises a sequence of chords of random tones that are harmonically non-related. When the stimulus contains a figure, a subset of the frequency components is repeated, which results in the percept of the figure popping out of the background (see Figure 3). In the stimulus, at any point in time the figure and the background are indistinguishable and it is only possible to

hear out the figure if the auditory system integrates information over time and frequency. This stimulus is similar to the random dot stereograms used in vision research (Julesz, 1971): when viewed through a stereoscope, a figure pops out of the pair of images due to local spatial correlation between the two images. When Julesz introduced dynamic random dot stereograms (Lehmann and Julesz, 1978), the paradigm allowed changes in horizontal disparity without associated monocularly visible motion. In response to these images, they found a negative/positive/negative sequence of peaks between 96 and 240 ms in the visual-evoked potentials.

Teki and colleagues (2011) used behavioural measures and fMRI to study the mechanisms involved in processing the stochastic figure-ground stimuli. Based on the behavioural tests, they concluded that people are capable of hearing out the figure effectively based on what they reasoned that the auditory system must possess mechanisms sensitive to cross-frequency and cross-time correlations. During the fMRI study, they found that the neurons within the banks of the superior temporal sulcus is responsible for extracting the stimulus-driven figure, whereas the intraparietal sulcus was found to be differentially activated for the effects of duration (anterior part) and coherence (posterior part) of the figure; duration denotes the number of repetitions of the same chord, whereas by coherence we mean the number of pure tone components within the repeating chord (see Figure 3).

In a second paper, Teki and colleagues (2013) report several other experiments in which they used different parameters for the same stimuli, for example shorter chord durations and interleaved noise between the chords, and altogether they found that the temporal coherence was the biggest influence for the subjects to segregate the figures from the background.

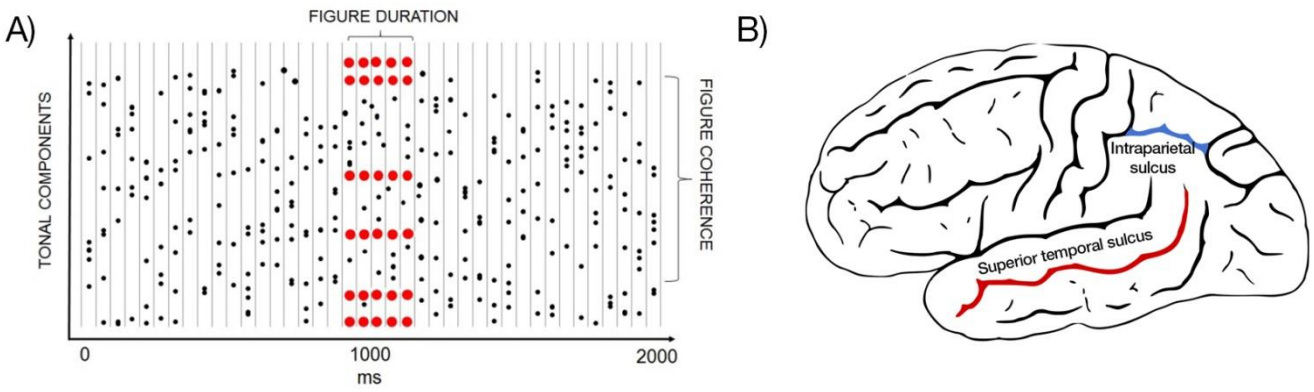

Figure 3. A) A schematic picture of the novel "stochastic figure-ground" stimuli including a figure (based on Tóth, Kocsis, Háden, Szerafin, Shinn-Cunningham, and Winkler, 2016). Black dots depict random tonal components while red ones represent repeating components. The onsets of the chords are represented as vertical lines. The x axis shows both time and the serial position of the chord within the stimulus, the y axis shows the frequency of components. B) A schematic depiction of the human brain, showing the intraparietal sulcus (marked with blue) and the superior temporal sulcus (marked with red). Teki and colleagues (2011) found that the neurons within the banks of the superior temporal sulcus are responsible for extracting the figure, whereas neural activation within the intraparietal sulcus was found to be sensitive to the duration and coherence of the figure.

O'Sullivan, Shamma and Lalor (2015) extended Teki and colleagues' findings with electroencephalography as the method possesses superior temporal resolution to fMRI. They altered the stimuli by trying to make them more naturalistic: they allowed for the repeated tones to be deviated from the original frequency throughout the figure, but not with more than two semitones and the figure tones always moved together. In their experiment, they used an active and a passive listening situation. O'Sullivan and colleagues (2015) used the temporal response function (TRF), a novel method to analyse the relationship between the coherence level of the stimuli and the EEG data. TRF can be interpreted as a filter that describes the brain's linear transformation of an input stimulus feature to continuous EEG data and it can be calculated by performing linear regression between those two variables. Based on their

results, they found that it is possible to obtain a temporally resolved neural signature of the temporal coherence of an acoustic scene. In both the active and passive situation, the response starts at 115 ms and is probably pre-attentive. During active listening, the response is larger, peaks later and lasts longer than in passive situation which indicates that the neural computations are enhanced during active listening.

Recently, Teki, Barascud, Picard, Payne, Griffiths, and Chait (2016) looked at stochastic figure-ground stimuli with MEG during passive listening situation and they found a robust neural evoked response to the emergence of the figure from the ground at about 150 ms after figure onset that was modulated by the number of temporally correlated components of the figure and was followed by a sustained-like phase until the figure offset. The source localization revealed the *planum temporale* and the intraparietal sulcus were activated during figure-ground segregation. Teki et al.'s (2016) participants did not focus on the sounds. However, training helps in better figure detection for these stimuli (Teki et al., 2011). Thus, although the processes underlying the segregation of figures in Teki and colleagues (2011 and 2016) stimuli may be categorized as "primitive" process by Bregman (1990), figure detection can be improved by experience.

In terms of the auditory scene analysis framework of Bregman (1990), figure-ground segregation involves three main steps: sequential and simultaneous grouping (first stage), segregation from the background, which involves competition between the resulting group plus the remaining background and the integrated percept (i.e., when the whole input is grouped together – resulting in the perception of no "figure" stimulus), and possibly the attentional selection of the figure over the ground. In contrast, Shamma and colleagues suggest that the main factor in organizing acoustic scenes is the temporal coherence between the brain responses that encode various sound features without the need to create alternative groupings and competition between them (Shamma and Micheyl, 2010; Shamma et al., 2011;

Teki et al., 2013). For the type of stimuli shown in Figure 3, Teki and colleagues (2013) described a two–stage version of the temporal coherence model. In the first stage, feature analysis is done by distinct populations of neurons tuned to different temporal modulation rates and spectral resolution scales, with at least pitch, timbre, and loudness being computed. Then, in the second stage, the output of the first stage is grouped according to common modulation (i.e., temporal coherence). The modelling results described by Teki and colleagues (2013) showed that temporal coherence defined this way is modulated by the "coherence" and the "duration" of the figure. Thus in this case, temporal coherence is a correlate of stimulus coherence allowing the neurons within the brain to pick out coherent groups of sound from the background. Indeed, the modulation of temporal coherence by the two figure properties was similar to how they modulated the listeners' detection performance.

2. Synopsis and rationale of the theses

The main aim of this thesis was to investigate how the human brain separates concurrent sounds or a figure from the background and what brain correlates can be found during this process. Concurrent sound segregation is a basic grouping process of auditory scene analysis, yet many features of it are not well understood. In this thesis a more ecologically valid paradigm was examined in the light of auditory objects and the combinations of simple cues of concurrent sound segregation. The goal of the thesis was to shed light on several questions: First, how does our brain cope with more realistic figure-ground segregation stimuli in an even-related potential paradigm? Is ORN elicited in response to an auditory figure when it is separated from the concurrent background? Thus *Thesis I* focuses on the generality of the ORN response. This thesis summarizes the findings in the electrophysiological study of stochastic figure-ground segregation, in which we investigated the effects of different parameters. Second, how does the auditory system use the different concurrent grouping cues? How do these different grouping cues interact with each other? What does the ERP correlate reflect in terms of cue evaluation: separate assessment of the cues or an integrated process? The rest of the studies utilize the ORN for investigating specific questions regarding concurrent sound segregation using the ORN response and brain oscillatory responses accompanying the ORN. *Thesis II* presents the conclusions from our study regarding these questions based on manipulating cues of concurrent sound segregation and their combinations in both active and passive listening situations. Third, what types of EEG oscillations are elicited by the concurrent sound segregation? *Thesis III* summarizes our results regarding the neural oscillations in concurrent sound segregation. And finally, what are the EEG correlates of concurrent sound segregation based on multiple cues that are either convergent or divergent? *Thesis IV* introduces a paradigm in which the perception of two or three concurrent sounds was promoted and cues were manipulated either convergently or

divergently. The results of the ERP study designed to test the paradigm are summarized. Taken together, what we seek to answer is whether the ORN reflects an overall readout of the available cues in the auditory system, or it shows the actual number and contribution of each cue?

2.1. Thesis I: Auditory figure-ground segregation

In everyday acoustic scenes figure and ground signals often overlap in time as well as in frequency content. Therefore to segregate the acoustic figure from the background noise one is typically required to group together sound elements over both time and frequency. Event-related brain potential (ERP) correlates of simultaneous and sequential grouping have both been studied, but they have generally been treated separately. Our aim with this study was to investigate the responses emerging in more natural situations where both grouping processes are required for perception. The salience of the figure was varied systematically by independently manipulating sequential and simultaneous cues supporting figure detection. This design allowed us to investigate the electrophysiological correlates of the emergence of an auditory object from a stochastic background. All studies tested healthy young participants without testing for musical expertise.

2.2. Thesis II: Effects of multiple congruent cues on concurrent sound segregation

Concurrent sounds segregation is based on instantaneously available cues, such as differences in pitch, sound onset, source location, and even the combination of the above. Most cues have been studied alone, or combined with another cue, but no systematic investigation has been done with several cues in both active and passive listening conditions.

Here we investigated the effects of combining different cues on the ORN and P400 ERP components. Participants were presented with complex tone sequences, half of which was manipulated: one or two harmonic partials were mistuned, delayed or presented from a

different source location that the rest. In separate blocks, one, two or three of these manipulations were combined. Participants either watched a silent, subtitled movie, or were asked to respond to each tone by pressing down one of two response buttons indicating whether they perceived one or two concurrent sounds. We tested whether the salience of the harmonicity-based cue can be further increased by mistuning two partials in a congruent manner and by investigating whether the effects of combined cues are additive, sub- or superadditive compared to the single-cue effects. If each cue elicits a separate ORN response, the ORN elicited by multiple congruent cues will be as large as the summed amplitudes of the ORN components elicited by the contributing cues. This would suggest that ORN reflects processes that are closely related to cue evaluation and farther upstream from what appears in perception. Alternatively, ORN may reflect the system's overall assessment of the likelihood that the auditory input consists of two concurrent sounds. That is, ORN could reflect the readout of a process combining the evidence from the available cues therefore we should find sub- or superadditivity between the contributing cues' ORN components.

2.3. Thesis III: Theta oscillations accompanying concurrent sound segregation

This study was aimed at assessing the large scale brain oscillations associated with concurrent sound segregation as there have not been many reports of oscillatory networks during concurrent sound segregation. The complex tones could be perceived as a single sound or as two concurrent sounds based on differences in the harmonic template, sound onset or sound source location. In separate task conditions, participants performed a visual change detection task (visual control), watched a silent movie (passive listening) or reported for each tone whether they perceived one or two concurrent sounds (active listening). Our goal was to determine the neural oscillatory correlates of concurrent sound segregation by comparing the oscillatory activity between concurrent sound segregation supported by different cues and between different attentional conditions.

2.4. Thesis IV: Two and three concurrent sound objects

The segregation of two concurrent sounds has gained the attention of researchers, yet no studies have dealt with the segregation of three concurrent sounds. Complex sounds containing sound segregation cues and cue combinations were set up to promote different separations of the tonal elements into one, two or three concurrent sounds.

Thus here we examined the effects of divergent cues on perceptual and neural (ORN) indicators of concurrent sound segregation. If ORN elicited by the divergent manipulations (three-objects conditions) is as large as the summed amplitudes of the ORNs elicited separately by the individual cues, this would be consistent with the interpretation that ORN reflects the independent evaluation of the divergent cues of concurrent sound segregation (i.e., more closely related to the evaluation of cues). Alternatively, if ORN generally shows sub-additivity relative to the ORN amplitudes elicited separately by the individual cues, this would support the interpretation that ORN reflects the auditory system's overall readout of the presence of multiple auditory objects, regardless of the congruency of the manipulations or the number of objects and different cues.

3. Studies

3.1. Study I: EEG signatures accompanying auditory figure-ground segregation

Tóth, B., Kocsis, Z., Háden, G.P., Szerafin, Á., Shinn-Cunningham, B., & Winkler, I. (2016). EEG signatures accompanying auditory figure-ground segregation. *Neuroimage*, 141, 108-119. DOI: 10.1016/j.neuroimage.2016.07.028.

NeuroImage 141 (2016) 108–119

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg

EEG signatures accompanying auditory figure-ground segregation

Brigitta Tóth [a,b,*], Zsuzsanna Kocsis [a,c], Gábor P. Háden [a], Ágnes Szerafin [a,c], Barbara G. Shinn-Cunningham [b], István Winkler [a,d]

[a] Institute of Cognitive Neuroscience and Psychology, Research Centre for Natural Sciences, Hungarian Academy of Sciences, Budapest, Hungary
[b] Center for Computational Neuroscience and Neural Technology, Boston University, Boston, USA
[c] Department of Cognitive Science, Faculty of Natural Sciences, Budapest University of Technology and Economics, Budapest, Hungary
[d] Department of Cognitive and Neuropsychology, Institute of Psychology, University of Szeged, Szeged, Hungary

ARTICLE INFO

Article history:
Received 2 March 2016
Accepted 11 July 2016
Available online 12 July 2016

Keywords:
Perceptual object
Auditory scene analysis
Figure-ground segregation
Event-related brain potentials (ERP)
Object-related negativity (ORN)
ERP source localization

ABSTRACT

In everyday acoustic scenes, figure-ground segregation typically requires one to group together sound elements over both time and frequency. Electroencephalogram was recorded while listeners detected repeating tonal complexes composed of a random set of pure tones within stimuli consisting of randomly varying tonal elements. The repeating pattern was perceived as a figure over the randomly changing background. It was found that detection performance improved both as the number of pure tones making up each repeated complex (figure coherence) increased, and as the number of repeated complexes (duration) increased – i.e., detection was easier when either the spectral or temporal structure of the figure was enhanced. Figure detection was accompanied by the elicitation of the object related negativity (ORN) and the P400 event-related potentials (ERPs), which have been previously shown to be evoked by the presence of two concurrent sounds. Both ERP components had generators within and outside of auditory cortex. The amplitudes of the ORN and the P400 increased with both figure coherence and figure duration. However, only the P400 amplitude correlated with detection performance. These results suggest that 1) the ORN and P400 reflect processes involved in detecting the emergence of a new auditory object in the presence of other concurrent auditory objects; 2) the ORN corresponds to the likelihood of the presence of two or more concurrent sound objects, whereas the P400 reflects the perceptual recognition of the presence of multiple auditory objects and/or preparation for reporting the detection of a target object.

Introduction

Selectively hearing out a sound from the background of competing sounds (referred to as auditory figure–ground segregation) is one of the main challenges that the auditory system faces in everyday situations. In ordinary acoustic scenes, figure and ground signals often overlap in time as well as in frequency content. In such cases, auditory objects are extracted by integrating sound components both over time and frequency. Auditory figure–ground segregation thus involves most of the processes of auditory scene analysis (Bregman, 1990): 1) grouping simultaneous components from disparate spectral regions and 2) across time into perceptual objects or sound streams, while 3) separating them from the rest of the acoustic scene. Event-related brain potential (ERP) correlates of simultaneous and temporal/sequential grouping have been studied, but they have generally been treated separately. As a result, little is known about the responses emerging in more natural situations where both grouping processes are required for veridical perception. The aim of the present study was to investigate electrophysiological correlates of figure–ground segregation by using auditory stimuli with high spectro-temporal complexity. The salience of the figure was varied systematically by independently manipulating sequential and simultaneous cues supporting figure detection. This design allowed us to investigate the electrophysiological correlates of the emergence of an auditory object from a stochastic background.

Auditory objects are formed by grouping incoming sound components over frequency and time (Kubovy and van Valkenburg, 2001; Griffiths and Warren, 2004; Shinn-Cunningham, 2008; Winkler et al., 2009; Bizley and Cohen, 2013) on the basis of various grouping heuristics (Bregman, 1990; Denham and Winkler, 2014). Simultaneous grouping is driven by various sound features such as common onset/offset (Lipp et al., 2010; Weise et al., 2012), location, loudness (Bregman, 1990; Darwin, 1997), as well as harmonic structure, or, more generally, spectral templates (Alain et al., 2002; for a review, see Ciocca, 2008). Feature similarity promotes sequential grouping (Van Noorden, 1975; Moore and Gockel, 2002; for reviews see Bregman, 1990; Carlyon et al., 2001). It interacts with the temporal separation of successive sounds, such that longer gaps between sounds reduce the likelihood of grouping even similar sounds into the same perceptual stream (Winkler et al., 2012; Mill et al., 2013). Temporal structure has been suggested to guide attentive grouping processes through rhythmic processing (Jones et al., 1981) and/or temporal coherence between elements of the auditory input (Shamma et al., 2011, 2013). For example,

http://dx.doi.org/10.1016/j.neuroimage.2016.07.028
1053-8119/

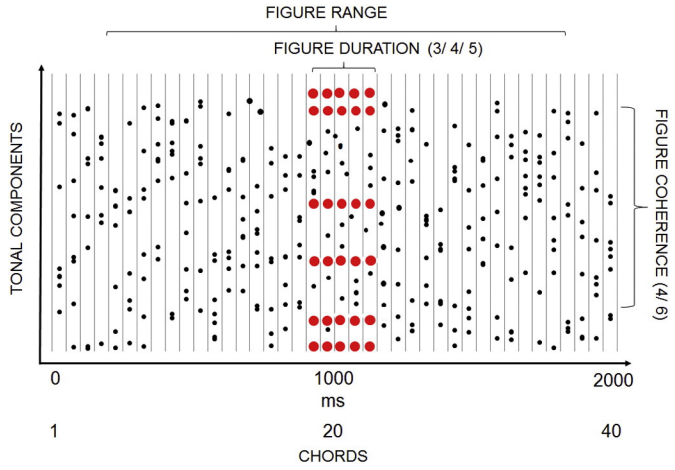

Fig. 1. Schematic illustration of a stimulus including a "figure" component. Black dots depict random tonal components while red represent repeating components. The onsets of the chords are represented as vertical lines. The *x* axis shows both time and the serial position of the chord within the stimulus. Stimuli consisted of 40 chords, each of 50-ms duration, and each containing a random set of 9 to 21 pure tone components. In half of the stimuli, an additional set of 4 or 6 tonal components was repeated 2, 3, or 4 times (resulting in 3, 4, or 5 consecutive chords) to create a "figure" that could be perceptually segregated from the rest of the random chords ("ground"). In the other half of the stimuli, random chords with the same numbers of tonal components were added to the ground ("control"). The figure/control started between 200 and 1800 ms from the stimulus onset.

within a stochastic background, the spectrotemporal regularity of a repeating cluster of synchronous tones causes them to stream together into a perceptual object distinct from the acoustic background (Elhilali et al., 2009; Elhilali et al., 2010). Indeed, temporal regularity also aids temporal/sequential segregation by allowing listeners to predict upcoming sounds (Dowling et al., 1987; Bendixen et al., 2010a, 2010b; Devergie et al., 2010; Szalárdy et al., 2014).

Few past studies addressed interactions between simultaneous and temporal grouping cues. Differences in amplitude modulation, a cue that helps simultaneous grouping through the gestalt "common fate" principle, has been also found effective for temporal grouping (Grimault et al., 2002; Szalárdy et al., 2013; Dolležal et al., 2012). Testing temporal coherence and harmonicity separately and together, Micheyl et al. (2013) found that the two cues separately facilitated auditory stream segregation. Teki et al. (2011, 2013) designed a new stimulus for testing both simultaneous and sequential grouping in auditory figure-ground segregation. The stimuli consist of a sequence of chords that are made up of pure tones with random frequency values and no harmonic relation to each other. When a subset of these tonal components is repeated several times, they form an auditory object (figure) which pops out from the rest of the stimulus (ground). The *coherence* of the figure is controlled by the number of frequencies in the subset making up the repeating chords, while the number of repetitions sets the *duration* of the figure. The separation of the figure from the ground requires integrating across both frequency and time. Specifically, there are no low-level feature differences between the figure and the ground; the subset of repeated components making up the figure chord is randomly chosen for each trial and each frequency can serve as part of the figure or of the ground, depending on the trial. Listeners are sensitive to the appearance of the spectro-temporally coherent figure in such stimuli, and figure salience systematically increases with increasing figure *coherence* and increasing figure *duration* (Teki et al., 2011; Teki et al., 2013, O'Sullivan et al., 2015).

Neural correlates of auditory stream segregation originate from a distributed network including the primary and non-primary auditory cortices and the superior temporal and intraparietal sulci (Teki et al., 2011; Alain, 2007; Alain and McDonald, 2007; Alain et al., 2002; O'Sullivan et al., 2015). Electrophysiological correlates of figure ground segregation have been investigated by using linear regression for extracting a signature of the neural processing of different temporal coherence defining a foreground object over a stochastic background (O'Sullivan et al., 2015). The results showed fronto-central activity suggesting early pre-attentive neural computation of temporal coherence between 100 and 200 ms poststimulus, which was extended beyond 250 ms when listeners were instructed to detect the figure. Further, a frontocentrally negative event-related potential (ERP) component of sound segregation, which typically peaks between 150 and 300 ms from cue onset, is elicited by auditory objects segregated by simultaneous cues (Alain and Izenberg, 2003, 2001; Alain and McDonald, 2007, McDonald and Alain, 2005). The object-related negativity (ORN) appears to reflect the outcome of the simultaneous segregation process (i.e., the perceptual decision that the acoustic input carries two or more concurrent sounds) rather than the processes leading to the perceptual decision (Kocsis et al., 2014). Sound segregation by simultaneous cues interacts with the temporal/sequential probability of the presence of these cues within the sound sequence, thus providing some evidence for joint processing of simultaneous and sequential cues of auditory stream segregation (Bendixen et al., 2010a; Bendixen et al., 2010b). When listeners are instructed to report whether they heard one or two sounds, ORN is followed by the centro-parietal P400 component peaking at about 450 ms from cue onset (Alain et al., 2001, 2002). P400 amplitude correlates with the likelihood of consciously perceiving two concurrent sound objects (Alain et al., 2001, 2002; Johnson et al., 2003). As for the ERP correlates of sequential sound segregation, the auditory *P*1 and N1 have been shown to be modulated by whether the same sound sequence is perceived in terms of a single (integrated) or two separate (segregated) streams (Gutschalk, 2005; Micheyl et al., 2007; Snyder and Alain, 2007; Szalárdy et al., 2013). The mismatch negativity (MMN) ERP can also be used as an index of sequential auditory stream segregation when the auditory regularities that can be detected from the stimulus sequences differ between the alternative sound organizations (Sussman et al., 1999; for reviews, see Winkler et al., 2009; Spielmann et al., 2014). However, MMN does not reflect auditory stream segregation per se; it can only be used as an indirect index of segregation in certain paradigms where the way in which the auditory scene is organized determines whether or not a particular sound will be perceived as a predicted or an unexpected event.

In two experiments, we employed the figure-ground stimuli adapted from Teki and colleagues' study (Teki et al., 2011) to analyze figure-ground segregation-related ERPs as a function of figure coherence and duration. Experiment 1 used behavioral methods a) to assess the optimal parameter ranges for figure coherence and duration to be used in the electrophysiological experiment (Experiment 2) and b) to test whether location difference between the frequency components assigned to the figure and the ground enhanced their separation. For Experiment 2, we hypothesized that concurrent sound segregation will lead to the elicitation of ORN and P400 (as listeners were instructed to detect the emergence of the figure) and further that the P400 and possibly the ORN amplitude will increase together with figure coherence, whereas figure duration may gate the emergence of these components. We further hypothesized that interactions between the effects of these parameters on the ERP components would arise, supporting the view that simultaneous (figure coherence) and temporal/sequential (figure duration) grouping cues interact when listeners parse complex acoustic scenes.

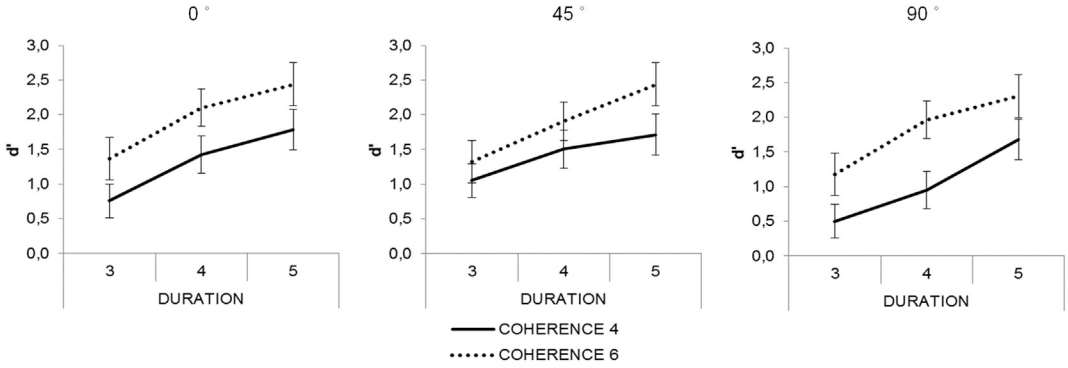

Fig. 2. In Experiment 1, detection improved with increasing figure coherence and increasing figure duration, but was worse when the figure and background were separated by a large spatial separation (see text). Group-averaged ($N = 20$) d′ values (standard error of mean represented by bars) are shown as a function of figure duration separately for the two coherence levels (marked by the different line types). The three levels of location difference between the figure and the ground are shown in the three separate panels.

Experiment 1

Methods

Participants

20 young adults (10 female; mean age: 22.4 years) participated in the experiment. They received modest financial compensation for participation. All participants had normal hearing and reported no history of neurological disorders. The United Ethical Review Committee for Research in Psychology (EPKEB; the institutional ethics board) approved the study. At the beginning of the experimental session, written informed consent was obtained from participants after the aims and methods of the study were explained to them.

Stimuli

The auditory stimuli (see a schematic example in Fig. 1) were adapted from Teki and colleagues' study (Teki et al., 2011). Each sound consisted of a sequence of 40 random chords of 50 ms duration with no inter-chord interval (total sound duration: 2000 ms). Chords consisted of 9–21 pure tone components. Component frequencies were drawn with equal probability from a set of 129 frequency values equally spaced on a logarithmic scale between 179 and 7246 Hz. The onset and offset of the chords were shaped by 10 ms raised-cosine ramps. In half of the stimuli, the same chord (containing 4 or 6 tonal components) was repeated 2, 3, or 4 times in a row (resulting in 3, 4, or 5 identical chords, respectively), thus forming a "figure" over the background of random chords. In the other half of the stimuli, random chords of 4 or 6 tonal components ("control") were added to 3, 4, or 5 consecutive chords (control chords). Past work showed that listeners could segregate repeating chords (but not additional random chords) from the other concurrent chords ("ground"), resulting in the perception of a foreground auditory object and a variable background (Teki et al., 2011). Each figure/control chord had a unique spectral composition with their frequencies randomly chosen from the set. The figure/control chords appeared at a random time between 200 and 1800 ms from stimulus onset (between the 5th and the 35th position within the sequence of 40 chords).

The figure chord sequences differed across trials on three dimensions: duration (the number of chords: 3, 4, or 5), coherence (the number of tonal components comprising the chord: 4 or 6), and perceived difference in lateral direction relative to the background (no difference, roughly 45° difference, or roughly 90° difference). The tones forming the background were always presented dichotically (perceived as originating from a midline location). In contrast, the interaural time and level differences (ITDs and ILDs, respectively) of the figure/control chords were manipulated to change their perceived laterality, either set to zero (heard at the same midline location as the background), heard at a lateral angle of roughly ±45° (ITD = ±395 μs and

Table 1
Group-average ($N = 25$) central (Cz) ORN (top) and parietal (Pz) P400 amplitudes and peak latencies (bottom) of the figure-minus-control difference waveforms, separately for the six stimulus conditions.

	Coherence 4			Coherence 6		
ORN	Duration 3	Duration 4	Duration 5	Duration 3	Duration 4	Duration 5
Mean amplitude at Cz (μV)	−0.37	−0.86	−1.17	−1.30	−1.70	−2.85
SD	1.22	1.72	1.42	1.58	1.90	2.03
t(24)	−1.48	−2.44*	−4.04***	−4.03***	−4.38***	−6.87***
Amplitude measurement window (ms)	200–300	200–300	232–332	172–272	200–300	232–332
ORN peak latency	258.08	263.04	272.16	242.40	268.80	273.28
SD	5.61	6.37	6.02	4.76	5.05	6.34
P400						
Mean amplitude at Pz (μV)	0.33	1.60	4.08	1.58	4.35	6.79
SD	1.39	2.04	2.76	1.72	2.93	4.05
t(24)	1.16	3.84***	7.23***	4.48***	7.27***	8.23***
Amplitude measurement window (ms)	452–552	520–620	580–680	500–600	480–580	480–580
P400 peak latency	554.08	561.92	556.32	542.24	545.12	536.80
SD	8.14	6.09	7.46	6.61	7.48	6.99

Significant differences from zero are marked by asterisks.
* $p < 0.05$.
*** $p < 0.001$.

Table 2
Group-average ($N = 25$) central (Cz) ORN (top) and parietal (Pz) P400 amplitudes and peak latencies (bottom) of the hit-minus-miss difference waveforms, separately for the four tested stimulus conditions.

	Coherence 4		Coherence 6	
ORN	Duration 4	Duration 5	Duration 3	Duration 4
Mean amplitude at Cz (μV)	−0.84	−2.57	−0.02	−2.03
SD	1.89	2.24	1.71	1.71
t(24)	−2.17*	−5.62***	−0.04	−5.82***
Amplitude measurement window (ms)	200–300	240–340	200–300	200–300
P400				
Mean amplitude at Pz (μV)	4.10	4.67	3.51	5.40
SD	2.87	3.89	3.47	3.81
t(24)	6.99***	5.88***	4.96***	6.95***
Amplitude measurement window (ms)	552–652	500–600	472–572	500–600

Significant differences from zero are marked with asterisks; due to the low number of Coherence-4/Duration-3 hit trials and Coherence-6/Duration-5 miss trials (<30% of all trials), the ERP measures are not reliable for these conditions.
* $p < 0.05$.
*** $p < 0.001$.

ILD = ±5.7 dB), or heard at a lateral angle of roughly ±90° (ITD = ±680 μs and ILD = ±9.08 dB). Thus, the figure and the ground overlapped spectrally; they could only be separated based on the figure's coherence and, when different from the background, the differences in perceived location.

Consecutive trials were separated by an inter-trial interval of 2000 ms. Listeners were presented with 20 trials of each stimulus type (figure vs. control × 2 coherence levels × 3 duration levels × 3 perceived location difference levels = 72 stimulus types, each appearing with equal probability) in a randomized order.

Stimuli were created using MATLAB 11b software (The MathWorks) at a sampling rate of 44.1 kHz and 16-bit resolution. Sounds were delivered to the listeners via Sennheiser HD600 headphones (Sennheiser electronic GmbH & Co. KG) at a comfortable listening level of 60–70 dB SPL (self-adjusted by each listener). Presentation of the stimuli was controlled by Cogent software (developed by the Cogent 2000 team at the FIL and the ICN and Cogent Graphics developed by John Romaya at the LON) under MATLAB.

Procedure

Listeners were tested in an acoustically attenuated room of the Research Centre for Natural Sciences, MTA, Budapest, Hungary. Each trial consisted of the presentation of the 2000-ms long sound, during which they were asked to focus their eyes on a fixation cross that appeared simultaneously at the center of a 19″ computer screen (directly in front of the listener at a distance of 125 cm). After the stimulus ended, a black screen was presented for 2000 ms. Listeners were instructed to press one of two response keys either during the stimulus or the subsequent inter-trial interval to indicate whether or not they detected the presence of a "figure" (repeating chord). The instruction emphasized the importance of responding correctly over response speed. The response key assignment (left or right hand) remained the same throughout the experiment and was counterbalanced across participants.

Prior to conducting the main experiment, listeners performed a 15 min practice session with feedback. The practice session consisted of two parts. In the first part, six stimulus sequences were presented. Each sequence consisted of 5 examples of the figure and 5 of the control condition, delivered in a randomized order (60 trials, altogether). In the practice session, the duration and coherence values used covered a larger range than in the main experiment, but all components were presented dichotically (no spatial location difference was employed). The figure stimuli were categorized into easy-to-detect (duration = 5, coherence = 6 and duration = 3, coherence = 8), moderately-difficult-to-detect (duration = 4, coherence = 4 and duration = 3, coherence = 6), and difficult-to-detect (duration = 3, coherence = 4 and duration = 2, coherence = 3) groups. In order to help listeners to learn the task, practice trials were organized into sequences consisting of sounds with the same difficulty level; these sequences were presented in descending order of detectability, from easy-to-detect to difficult-to-detect. All other parameters were identical to those described for the main experiment. To accustom listeners to the perceived location manipulation, 6 additional practice blocks were presented, one for each of the six levels of perceived location difference presented (0, 15, 30, 45, 60, and 90°). In these practice sequences, the figure duration was always 5 and the coherence level 6. Each level of the perceived location difference was presented for 12 trials (6 with a figure and another 6 with the control; 72 overall). These were presented in a fixed order (90 60, 0, 45 30, and 15°). All other stimulus parameters were identical to those described for the main experiment.

No feedback was provided to listeners in the main experiment, which lasted for about 1.5 h. The main experiment was divided into 20 blocks, each consisting of 72 trials. The order of the different types of trials was randomized separately for each listener. Listeners were allowed a short rest between stimulus blocks.

Data analysis

Reaction times were not analyzed, because listeners were instructed to respond accurately rather than as fast as they could. For the d′ values (the standard measure for discrimination sensitivity; see, for example, Green and Swets, 1988) a repeated-measures ANOVA was performed with the factors of Coherence (2 levels: 4 vs. 6 tonal components) × Duration (3 levels: 3 vs. 4 vs. 5 chords) × Location difference (3 levels: 0 vs. 45 vs. 90°). Statistical analyses were performed with the Statistica software (version 11.0). When the assumption of sphericity was violated, degrees of freedom values were adjusted using the Greenhouse-Geisser correction. Bonferroni's posthoc test was used to qualify significant effects. All significant results are described. The ε correction values for the degree of freedom (where applicable) and the partial η^2 values representing the proportion of explained variance are shown.

Results and discussion

The results of Experiment 1 are presented in Fig. 2. The fact that the d′ values exceeded 2 for several parameter combinations demonstrates that listeners were sensitive to the appearance of figure in the stimuli, confirming that the auditory system possesses mechanisms that process cross-frequency/time correlations (Teki et al., 2011). The main effect of Coherence ($F(1,19) = 97{,}05$, $p < 0.001$; $\eta^2 = 0.83$) demonstrates that listeners were better at detecting figures containing six tonal components than those comprising four components. The main effect of Duration was also significant ($F(2,38) = 114.98$, $p < 0.001$; $\eta^2 = 0.85$).

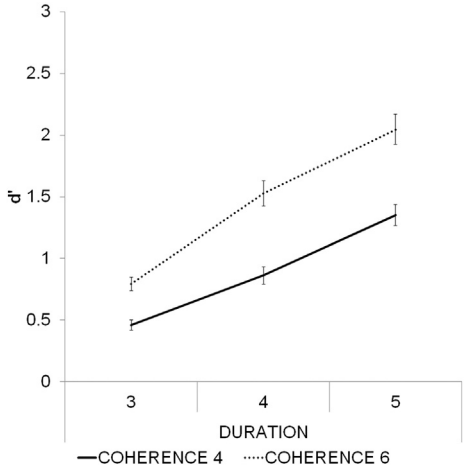

Fig. 3. In Experiment 2, detection improved with increasing figure coherence and increasing figure duration, consistent with Experiment 1. Group-averaged ($N = 25$) d′ values (standard error of mean represented by bars) are shown as a function of figure duration separately for the two coherence levels (marked by the different line types).

Pairwise post-hoc comparisons showed that the d′ values were significantly higher for figure duration of 5 than for durations of 3 or 4 chords ($p < 0.001$, both), and that the d′ for figure duration of 4 chords was significantly higher than for duration of 3 chords ($p < 0.001$). Location difference also yielded a significant main effect ($F(2,38) = 9{,}96$, $p < 0.01$; $\eta^2 = 0.34$). Post hoc pairwise comparisons showed that the d′ for figures with 90° difference from the ground was significantly lower than that for figures with 0 or 45° location difference ($p < 0.01$, both). There were no significant interactions between the three factors.

Similarly to previous results (Teki et al., 2011), we found that increasing figure coherence and duration helped listeners to separate the figure from the ground in the expected way and without interactions between these factors. We expected that increasing location difference between the figure and the ground would help figure-ground segregation, helping the detection of the figure. Instead we found that a large separation between the figure and ground interfered with detection of the figure. We ascribe this difference to an effect of top-down attention: the figure could appear at any lateral angle, from roughly −90° to +90°; listeners may have adopted a strategy of listening for the figure near midline (at the center of the range). If the actual figure was too far from this attended direction (e.g., at the extreme locations of ±90°), it may have fallen outside the focus of attention. Given that our focus was on bottom-up, automatic processes involved in segregating figure and group, we excluded the location manipulation from Experiment 2.

Experiment 2

Methods

Participants

27 young adults (17 female; mean age 21.9 years) with normal hearing and no reported history of neurological disorders participated in the experiment. None of the participants were taking medications affecting the nervous system and none of them participated in Experiment 1. The study was approved by the institutional ethics board (EPKEB). At the beginning of the experimental session, written informed consent was obtained from participants after the aims and methods of the study were explained to them. Participants were university students who received course credit for their participation. Data of one participant was excluded from the analysis due to a technical problem in the data recording.

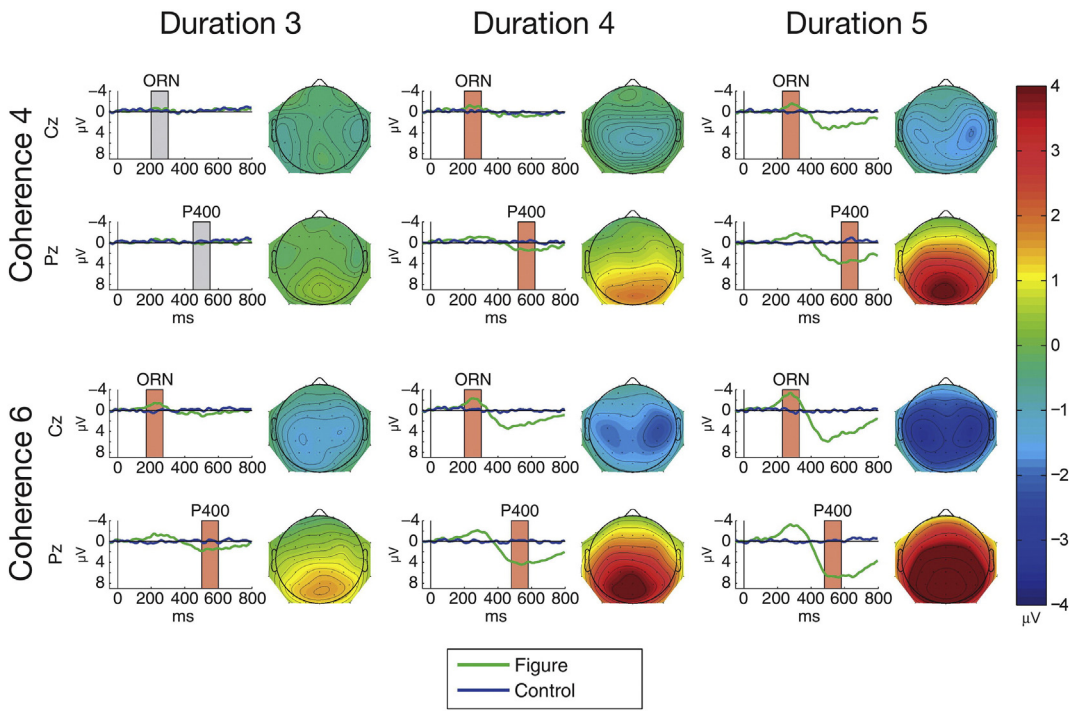

Fig. 4. Group-average ($N = 25$) ERPs elicited by figure (green lines) and control stimuli (blue lines) triggered from the figure/control segment onset (0 ms at the x axis) at Cz (top of each panel) and at Pz (bottom of each panel) for the 6 stimulus conditions (Coherence: 4 or 6; Duration: 3, 4, or 5). Boxes mark the measurement windows for ORN at Cz and P400 at Pz; a red box indicates that the figure-minus control difference significantly differed from zero ($p < 0.05$) within the measurement window, while a grey box indicates no significant amplitude difference. The scalp distribution of the mean difference amplitude within the measurement window is shown to the right of each panel. Color calibration is at the right side of the figure.

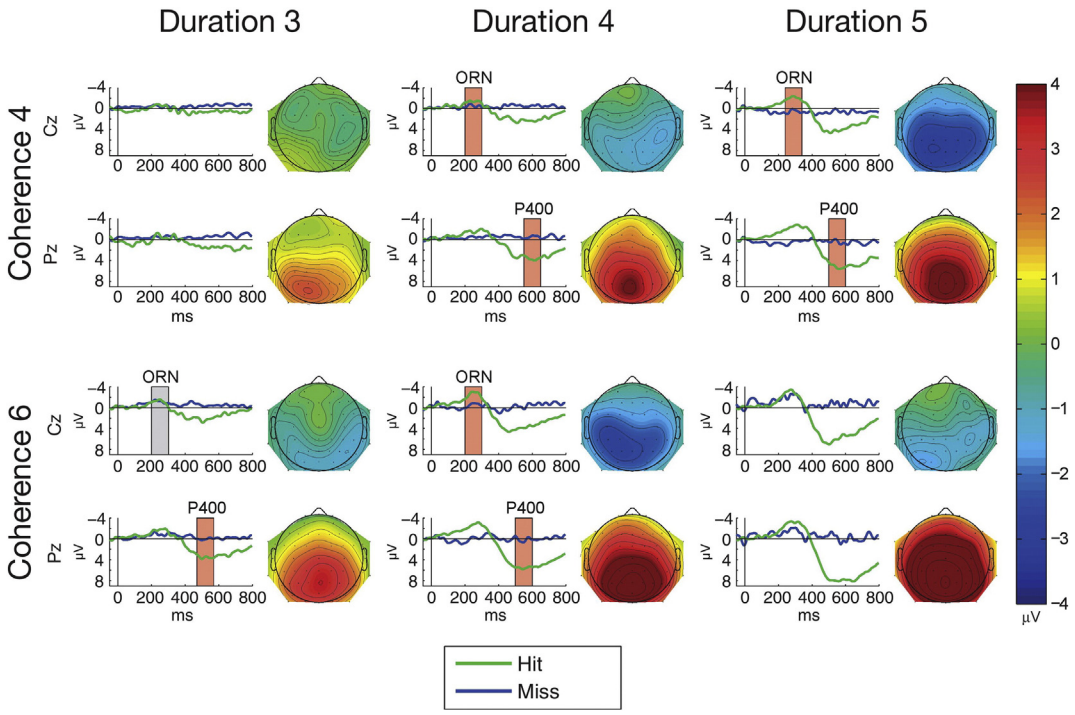

Fig. 5. Group-average ($N = 25$) ERPs elicited for hit (green lines) and miss trials (blue lines) triggered from the figure segment onset (0 ms at the x axis) at Cz (top of each panel) and at Pz (bottom of each panel) for the 6 stimulus types (Coherence: 4 or 6; Duration: 3, 4, or 5. Boxes mark the measurement windows for ORN at Cz and P400 at Pz; a red box indicates significant amplitude difference ($p < 0.05$) between hit and corresponding miss trials within the measurement window, a grey box indicates no significant amplitude difference. Note that due to the low number of hit or miss trials in the Coherence-4/Duration-3 and Coherence-6/Duration-5 conditions, no response amplitudes were measured. The scalp distribution of the mean hit-minus-miss difference amplitudes within the measurement window is shown to the right of each panel. Color calibration is at the right side of the figure.

Stimuli

The stimuli were identical to those delivered in the "no location difference" condition of Experiment 1 except that the test sounds were composed of 41 tonal segments. The stimulus set in the EEG experiment therefore comprised six stimulus conditions: 2 coherence levels (4, 6 tonal components) × 3 duration levels (3, 4, 5 chords). Fifty percent of the sounds carried a figure, which appeared between 200 and 1800 ms (5th–35th chord) from onset.

Procedure

Participants were tested in an acoustically attenuated and electrically shielded room of the Research Centre for Natural Sciences, MTA, Budapest, Hungary. Each trial started with the delivery of the sound with a concurrent presentation of the letter "S" at the center of a 19″ computer screen placed directly in front of the participant (distance: 125 cm). Following the stimulus presentation, the letter "S" was replaced by a question mark on the screen denoting the response period which lasted until a response was made. After the response was recorded, the screen was blanked for a random inter-trial interval of 500–800 ms (uniform distribution) before the next trial began. Listeners were instructed to press one of two response keys during the response period to mark whether or not they detected the presence of a "figure" (repeating chord). The instruction emphasized the importance of confidence in the response over speed. The response key assignment (left or right hand) remained the same during the experiment and was counterbalanced across participants.

Before the main experiment, participants completed a short practice session (10 min) during which they received feedback. The practice session was identical to the first part of the practice session of Experiment 1. (The second part, training for the perceived location manipulation, was skippEd.)

The main experiment lasted about 90 min. Overall, listeners received 130 repetitions of each stimulus type (2 coherence levels × 3 duration levels × figure present vs. absent), divided into 10 stimulus blocks of 156 trials each. The order of the different types of trials was separately randomized for each listener. Participants were allowed a short rest between stimulus blocks.

Data analysis

Behavioral responses. Figure detection was assessed by means of the sensitivity index (d′ value), separately for each figure type, with the control trials serving as distractors. For the d′ data, a repeated-measures ANOVA was performed with the factors of Coherence (2 levels: 4 vs. 6 tonal components) × Duration (3 levels: 3 vs. 4 vs. 5 chords).

EEG recording and preprocessing. EEG was recorded from 64 locations of the scalp with Ag/AgCl electrodes placed according to the international 10–20 system with Synamps amplifiers (Neuroscan Inc.) at 1 kHz sampling rate. Vertical and horizontal eye movements were recorded by electrodes attached above and below the left eye (VEOG) and lateral to the left and right outer canthi (HEOG). The tip of the nose was used as reference and an electrode placed between Cz and FCz was used as ground (AFz). The impedance of each electrode was kept below 15 kΩ. Signals were filtered on-line (70 Hz low pass, 24 dB/octave roll off).

The analysis of EEG data was performed using Matlab 7.9.1 (Mathworks Inc.) The continuous EEG signal was filtered between 0.5 and 45 Hz by band-pass finite impulse response (FIR) filter (Kaiser windowed, Kaiser $\beta = 5.65$, filter length 4530 points). EEG signals were converted to average reference. In order to exclude EEG segments containing infrequent electrical artifacts (rare muscle and movement artifacts etc.), the data were visually screened and the affected segments were rejected. Next the Infomax algorithm of Independent Component

Analysis (ICA) (as implemented in EEGlab; for detailed mathematical description and validation, see Delorme and Makeig, 2004, Delorme et al., 2007) was performed on the continuous filtered dataset of each subject, separately. ICA components constituting blink artifacts were removed via visual inspection of their topographical distribution and frequency content.

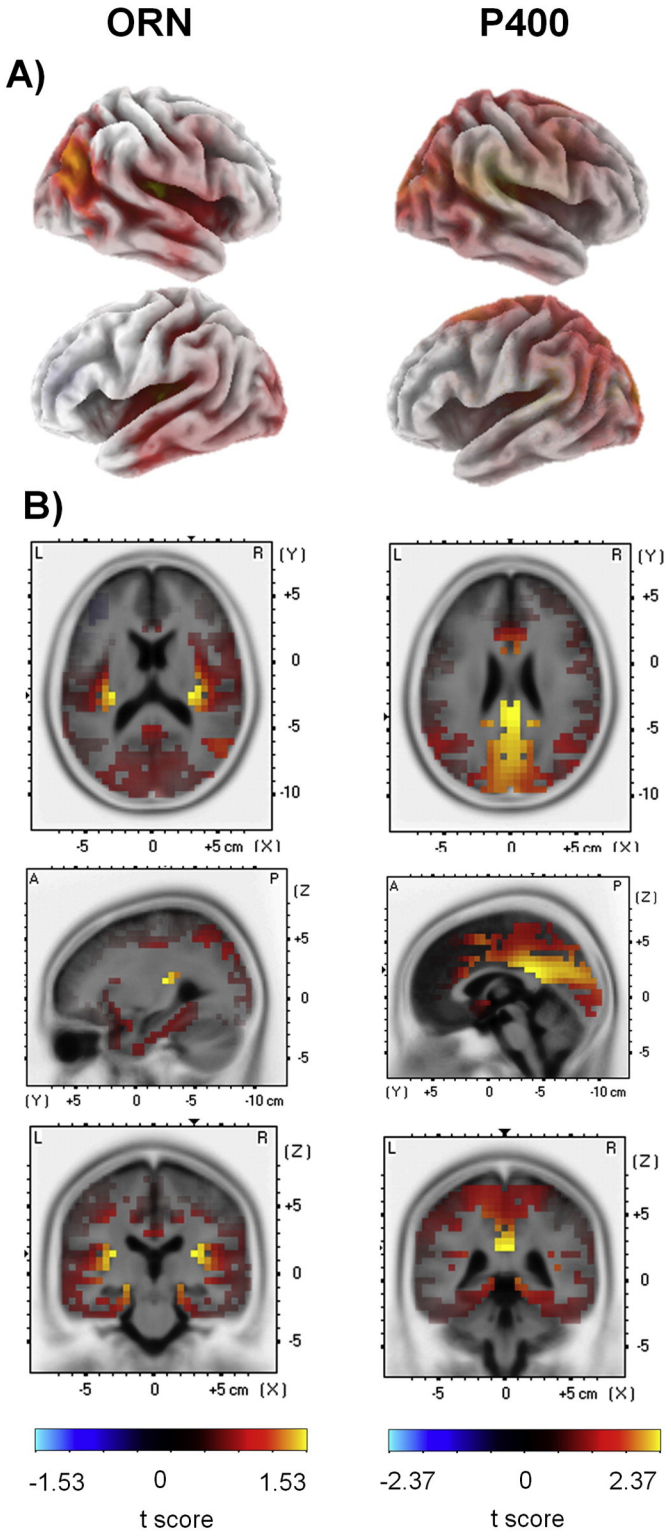

Fig. 6. LORETA t-value maps from voxel-by-voxel paired t-tests contrasting current density values between figure and control stimuli for the ORN (left) and P400 (right) latency range. Red color corresponds to higher current source density magnitudes (indexed by positive t values) for the figure compared to control trials (color scales are at the bottom of the left and right panels). A) Maps are displayed on the 3D inflated cortex. The 3D inflated cortex plots present the right hemisphere on the top and left hemisphere below. B) Maps shown on the MNI152 standard brain template. Coordinates are scaled in cm; origin is at the anterior commissure; (X) = left (−) to right (+); (Y) = posterior (−) to anterior (+); (Z) = inferior (−) to superior (+). The maps corresponding to the ORN time window (200–350 ms) are shown at the x = −40 mm, y = −25 mm, z = 0 mm MNI coordinates; the maps corresponding to the P400 time window (460–600 ms) are shown at the x = 30 mm, y = −25 mm, z = 15 mm MNI coordinates.

ERP data analysis. For the ERP analysis, the EEG signals were downsampled to 250 Hz and filtered between 0.5 and 30 Hz by a band-pass finite impulse response (FIR) filter (Kaiser windowed, Kaiser β = 5.65, filter length 4530 points). EEG epochs of 850 ms duration were extracted separately for each stimulus from 50 ms before the onset of the figure/control within each trial and baseline corrected by the average voltage in the pre-stimulus period. Epochs with an amplitude change exceeding 100 μV at any electrode were rejected from further analysis. The data of one subject were excluded from further analysis due to low signal to noise ratio: we obtained fewer than 20 artifact free epochs for one of the stimulus types. Overall, 84.2% of the data was retained.

Difference waveforms were calculated between ERPs elicited by the figure- and the control-trial responses. Inspecting the group-averaged difference waveforms elicited by the figure trials in each condition, we observed an earlier negative and a later positive centroparietal response in most conditions. We tentatively identified them as ORN and P400, respectively. Using the typical latency windows for ORN (150–300 ms) and P400 (450–600 ms) we performed peak detection for ORN and P400 at their typical maximal scalp location (maximal negative value at Cz and maximal positive value Pz within the ORN and P400 time window, respectively) on the group-averaged waveforms, separately for each condition. Based on these peak latencies, ORN and P400 amplitudes were then averaged from 100 ms wide windows centered on the detected peaks (see Table 1 for descriptive statistics of the ERP amplitudes). Individual peak latencies were determined from the same latency windows and electrode location as was described above. For assessing whether ORN and/or P400 were elicited, ERP amplitude differences were tested against zero by one-sample t-tests, separately for each stimulus condition and time window. For testing the effects of coherence and duration on figure vs. control trials, central (Cz) ORN and parietal (Pz) P400 amplitudes and peak latencies were compared by repeated-measures ANOVA with the factors of Coherence (2 levels: 4 vs. 6 tonal components) x Duration (3 levels: 3 vs. 4 vs. 5 chords).

For testing the effects of coherence and duration on hit and miss trials, difference waveforms were calculated between ERPs elicited by hit (correct response to figure trials) and miss trials (no response to figure trials). Peak latency and subsequent amplitude measurements were performed by the same procedure as those described for figure vs. control trial analyses. Measurement windows and descriptive statistics are shown in Table 2. Because both this and the following analyses were based on the figure trials alone, only half of the trials were used. In the Coherence-4/Duration-3 and in the Coherence-6/Duration-5 conditions, very few hit or miss trials were obtained because of the very low and very high detection rates (respectively). Therefore, these stimulus conditions were excluded from further analysis. Paired-samples t-tests were performed separately for the remaining four stimulus types to compare the trial types (hits vs. misses). In order to determine whether the processes indexed by ORN and P400 are related to the inter-individual variability in figure detection sensitivity, the amplitude differences between hit and miss trials in the ORN (Cz) and P400 (Pz) time windows were correlated with d′ (Pearson correlation), separately for each stimulus condition.

Statistical analyses were performed with the Statistica software (version 11.0). When the sphericity assumption was violated, the

Table 3
Summary of significant differences of LORETA-based estimates of neural activity for figure versus control in the Coherence 6 conditions in the time ORN window (200–350 ms). The anatomical regions, MNI coordinates, and BAs of maximal t-values are listed.

Region	BA	MNI coordinates (mm)			Voxels (N)	t-Value	*p* value
		x	y	z			
Transverse temporal gyrus	41	40	−25	10	3	1,33	<0.001
Superior temporal gyrus	39	45	−60	30	1	1,27	<0.001
Angular gyrus	39	50	−60	30	1	1,26	<0.001
Anterior cingulate	25	0	0	−5	3	1,55	<0.001
Parahippocampal gyrus	25, 27, 28, 30, 34. 35	0	−35	0	27	1,41	<0.001

Note: Positive t-values indicate stronger current density for figure than for control trials. The numbers of voxels exceeding the statistical threshold ($p < 0.01$) are also reported. The origin of the MNI space coordinates is at the anterior commissure; (X) = left (−) to right (+); (Y) = posterior (−) to anterior (+); (Z) = inferior (−) to superior (+).

degrees of freedom were adjusted using the Greenhouse-Geisser correction. Bonferroni's post hoc test was used to qualify significant effects. All significant results are described. The ε correction values for the degree of freedom (where applicable) and the partial η^2 values representing the proportion of variance explained are shown.

Source localization by sLORETA. The sLORETA software (standardized Low Resolution Brain Electromagnetic Tomography; Pascual-Marqui et al., 2002) allows the location of the neural generators of the scalp-recorded EEG to be estimated. The algorithm limited the solution to the cortical and hippocampal grey matter according to the probability template brain atlases based on template structural MRI data provided by the Montreal Neurological Institute (MNI). Electrode locations were calculated according to the 10–20 system without individual digitization. The solution space is divided into 6239 voxels ($5 \times 5 \times 5$ mm resolution). Source localization computations are based on a three-shell spherical head model registered to the Talairach human brain atlas. Because the highest-amplitude sound segregation related ERP responses were obtained for the Coherence-6 stimuli, current density maps were generated from the ORN (200–350 m) and P400 (460–600) measurement windows of the figure and control trials collapsing across durations 3–5, separately for each participant. For comparisons of the electrical source activity between the figure and the control trials, Student's t value maps were generated using the LORETA-Key software package's statistical nonparametric mapping voxel-wise comparison calculation tool.

Results

Behavioral responses

Group-averaged d′ values are presented in Fig. 3. There was a significant main effect of Coherence ($F(1,24) = 153.84$, $p < 0.001$, $\eta^2 = 0.865$), confirming that d′ was greater for figures consisting of 6 compared to 4 tonal components. The main effect of Duration was also significant ($F(2,48) = 193.51$, $p < 0.001$, $\eta^2 = 0.89$, $\varepsilon = 0.89$). Pairwise post hoc comparisons showed that the d′ values for figure duration of 5 chords were significantly higher than those for durations of 3 or 4 chords ($p < 0.001$, both), and the d′ values for figure duration of 4 chords were significantly higher than those for duration of 3 chords ($p < 0.001$). There was also a significant interaction between Duration and Coherence ($F(2,48) = 18.52$, $p < 0.001$, $\eta^2 = 0.44$). All post hoc pairwise comparisons between different figure types yielded significant ($p < 0.001$) results, except that between Coherence-6/Duration-3 and Coherence-4/Duration-4. These results are compatible with those of Teki et al. (2011) and of Experiment 1.

ERP responses

Comparison between the figure and control responses

Mean ERP responses elicited by all figure and control sounds are shown in Fig. 4. Figure-minus-control difference amplitudes measured from the ORN and P400 time windows (at Cz and Pz, respectively) significantly differed from zero for all stimulus types except for Coherence-4/Duration-3 (see Table 1). The ORN shows a lateral central maximum extending to central and parietal scalp locations with increasing Coherence and Duration. The P400 shows a midline parietal maximum extending towards lateral and central scalp locations with increasing Coherence and Duration. Table 2 shows all significant results for the ANOVAs of the ORN and P400 amplitudes.

The ANOVA comparing the central (Cz) ORN amplitudes showed a significant main effect of Coherence ($F(1,24) = 24.61$, $p < 0.001$, $\eta^2 = 0.506$), which was due to significantly larger amplitudes for Coherence-6 than for Coherence-4 stimuli ($p < 0.001$). The main effect of Duration was also significant ($F(2,48) = 8.288$, $p < 0.001$, $\eta^2 = 0.257$); post-hoc pairwise comparisons showed significantly larger amplitudes for Duration 5 than for the 3 or 4 conditions ($p < 0.001$ and $p = 0.047$, respectively). The ANOVA comparing the ORN peak latencies showed a significant main effect of Duration ($F(2, 48) = 9.12$, $p < 0.001$, $\eta^2 = 0.275$) with post-hoc pairwise comparisons indicating

Table 4
Summary of significant differences of LORETA-based estimates of neural activity for figure versus control in the Coherence 6 conditions in the P400 time window (460–600 ms). The anatomical regions, MNI coordinates, and BAs of maximal t-values are listed.

Region	BA	MNI coordinates (mm)			Voxels (N)	t-Value	*p* value
		x	y	z			
Superior temporal gyrus	41	40	−40	10	1	1.86	<0.001
Medial frontal gyrus	6,32	0	5	50	20	2.01	<0.001
Paracentral gyrus	5,31	−15	−40	50	7	1.91	<0.001
Superior frontal gyrus	6	0	5	55	21	1.99	<0.001
Cingulate gyrus	23,24,31,32	0	−40	25	178	2.38	<0.001
Anterior cingulate gyrus	33	5	10	25	5	1.98	<0.001
Posterior cingulate gyrus	23, 29, 30,31	5	−40	25	50	2.38	<0.001
Parahippocampal gyrus	27, 30	10	−35	0	13	2.06	<0.001
Cuneus	7,18,19	0	−75	20	138	2.21	<0.001
Precuneus	7,19,31	0	−50	30	152	2.23	<0.001
Middle occipital gyrus	18	−15	−90	15	10	1.94	<0.001

Note: Positive t-values indicate stronger current density for figure than for control trials. The numbers of voxels exceeding the statistical threshold ($p < 0.01$) are also reported. The origin of the MNI space coordinates is at the anterior commissure; (X) = left (−) to right (+); (Y) = posterior (−) to anterior (+); (Z) = inferior (−) to superior (+).

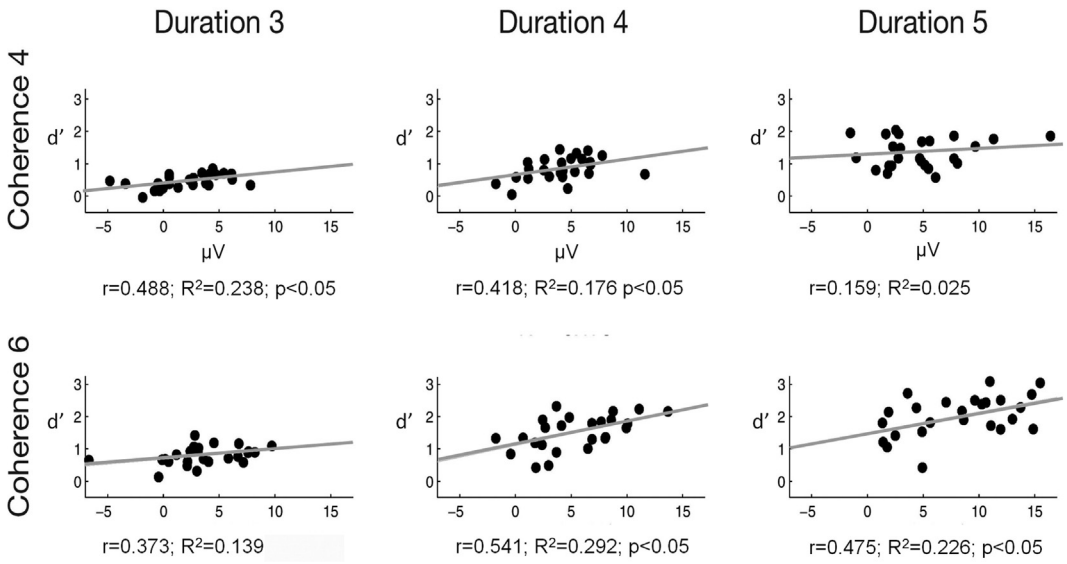

Fig. 7. Across individual subjects, the change in the size of the P400 amplitude difference for hit-miss trials (measured at Pz) correlates with figure-detection performance (d′) for four of the six stimulus conditions. The dots represent the different listeners' data. Pearson correlation r values and R2 determination coefficients and *p*-values are shown on each panel. A regression line is shown on each panel representing the relationship between P400 amplitudes and d′.

significantly shorter ORN latencies in the 3 than the 4 or 5 chords conditions ($p < 0.02$ and $p < 0.001$, respectively). Note that the peak-latency effect was caused by the increased ORN duration and amplitude elicited at longer figure durations (see Fig. 4).

The ANOVA comparing the parietal (Pz) P400 amplitudes showed significant main effects of Coherence ($F(1,24) = 37.856$, $p < 0.001$, $\eta^2 = 0.611$) due to significantly higher amplitudes for the 6 tonal components than for 4 tonal components ($p < 0.001$) and Duration ($F(2,48) = 51.944$, $p < 0.001$, $\eta^2 = 0.684$), post-hoc pairwise comparisons showed significantly higher amplitudes for 5 than for 3 or 4 chords and for 4 than for 3 chords; $p < 0.001$ in all comparisons. There was also a significant interaction between Coherence and Duration ($F(2,48) = 4.005$, $p = 0.025$, $\eta^2 = 0.143$). Posthoc ANOVAs were performed with the factors of Coherence (2 levels: 4 vs. 6 tonal components) separately for each level of Duration. These revealed significant Coherence main effects at each level of Duration ($F(1,24) = 9.32$, $p = 0.005$, $\eta^2 = 0.279$; $F(1,24) = 29.11$, $p < 0.001$, $\eta^2 = 0.548$; $F(1,24) = 21.91$, $p < 0.001$, $\eta^2 = 0.477$; for Durations levels 3, 4, and 5, respectively). The Coherence main effect size was lower for stimuli with Duration 3 than for stimulus with Duration 4 or 5. These results indicate that the source of interaction between Coherence and Duration is that the effect of Coherence is larger at the two longer than at the shortest duration. The ANOVA comparing the P400 peak latencies showed a significant main effect of Coherence ($F(1, 24) = 11.49$, $p = 0.002$; $\eta^2 = 0.323$) due to significantly shorter ERP latency for Coherence-6 than for Coherence-4 stimuli.

Comparison between the hit and miss figure trial responses

ERP responses from the hit and miss figure trials are shown in Figure 5. The central (Cz) hit and miss amplitudes measured in the ORN latency range significantly differed from each other for all but one of the tested stimulus condition: Coherence-4/Duration-3 (see Table 2).[1] The parietal (Pz) amplitudes measured from the P400 latency range significantly differed between hit and miss trials for each of the tested conditions (see Table 2).

[1] Note that the number of trials averaged for the compared hit and miss responses differed from each other. However, the difference never exceeded the ~1:2 ratio, because the *t* tests were only conducted for those conditions in which the number of hit and miss trials separately exceeded 30% of the total number of trials. The Coherence-4/Duration-3 and Coherence-6/Duration-5 conditions were dropped from these analyses due to this reason.

ORN and P400 source localization

LORETA paired-sample t-tests revealed significantly higher current source density in response to figure than control trials corresponding to the sources of ERPs at the ORN and P400 time windows. LORETA t value maps superimposed on the MNI152 standard brain are shown in Figure 6, while the statistical results are shown in Tables 3 and 4 for the ORN and P400 ERPs, respectively. In both time windows, Brodmann area 41 (BA 41) on the right hemispheres, the anterior transverse temporal part of the primary auditory cortices, and the anterior cingulate cortex (ACC, BA 25, 33) were found to be more active during figure compared to control trials. At the ORN time window, activity was greater for figure than control trials also in the cortical regions of BA 39, including areas of the superior temporal gyrus and the inferior parietal sulcus (angular gyrus). In the time window of P400, several other brain regions were observed to be more active for figure than for control stimuli. These include frontal cortical areas such as the medial and superior frontal gyri (BA 6, 32, 31), the cingulate cortices (BA 23,24, 29, 30,31,32), and also areas in the visual cortices (BA 7,18, 19).

Correlation between behavioral and ERP measures

Discrimination sensitivity (d′) was correlated with the amplitude difference between hit and miss trials in the ORN and P400 time window. No significant correlation was found for the central (Cz) amplitude difference in the ORN time window. However, significant positive correlations were obtained between the parietal (Pz) hit-minus-miss amplitude difference measured from the P400 time window and d′ for four of the six stimulus conditions (see Fig. 7).

General discussion

In accordance with the findings of Teki et al. (2011 and 2013), the results of both Experiment 1 and 2 showed that both the coherence of the figure and its duration promoted figure–ground segregation: Figure detection performance improved as the number of repeated tonal components increased and as the number of repetitions of the figure elements increased. In other words, the perceptual salience of the figure increased parametrically with increasing figure coherence and duration. This result confirms that the segregation of the figure from the concurrently presented stochastic background required the integration of acoustic elements over time and frequency. Teki et al. (2013)

showed that the effects of figure coherence and duration on figure–ground segregation can be explained by the temporal coherence principle (Shamma et al. 2011, 2013). In the temporal coherence model, auditory features (such as location, pitch, timbre, loudness) are first extracted in auditory cortex by distinct neuron populations. Correlations between the dynamic activity of these distinct cortical populations cause perceptual streams to emerge, as described by the resulting correlational matrix of activity patterns.

We found no evidence that spatial separation between the figure and the background led to an automatic enhancement of figure–ground segregation; instead, when the figure came from the most extreme lateral locations, detection of the figure was poorer than when it came from closer to midline. Taken together with the results of previous studies of simultaneous sound segregation (McDonald and Alain, 2005; Kocsis et al., 2014, Lee and Shinn-Cunningham, 2008), this finding supports the idea that spectrotemporal cues contribute automatically to figure–ground segregation, while spatial cues are more influential in directing top-down, volitional attention. This conclusion is also compatible with that of Bregman (1990), who argued that source location is a weak cue of auditory stream segregation.

Correct identification of the figure resulted in the elicitation of a centrally maximal negative response between 200 and 300 ms from the figure onset and a parietally maximal positive response between 450 and 600 ms (Experiment 2). Based on the observed scalp distributions, their cortical source origin, and the latency range, these ERP responses could be identified as the ORN and P400 (Alain and McDonald, 2007; Lipp et al., 2010; Johnson et al., 2007, Bendixen et al., 2010a, 2010b), respectively, which are known to be elicited when two concurrent sounds are attentively segregated (Alain et al. 2001, 2002). However, ORN (and P400) have been previously observed only in the context of one vs. two discrete concurrent complex tones, whereas the present figure stimuli formed a coherent stream that was separated from the randomly changing background. Thus, the current results demonstrate that ORN and P400 are elicited also in cases when concurrent sound segregation requires integrating spectral cues over time to form a new stream. In turn, the elicitation of these ERP components suggests that the brain mechanisms underlying figure–ground segregation by spectral coherence over time may reflect some common processes with those involved in simpler forms of simultaneous sound segregation, such as some common segregation mechanism or common consequence of detecting two concurrent sounds. If ORN is based on deviation from some template (Alain et al., 2002), then the current results suggest that the template does not have to be fixed, such as a template of harmonicity (Lin and Hartman, 1998). Rather, it can be built dynamically by extracting higher-order spectro-temporal statistics of the input stimulus. This conclusion is also supported by the results of O'Sullivan et al. (2015), who manipulated the coherence level of the figure under both active and passive listening conditions. These authors found that a neural response appearing in the same latency range as the present ORN was correlated with the coherence level of the figure stimuli. It is possible that this neural activity (extracted from the EEG by a linear regression method) corresponds to or at least overlaps with the ORN response obtained with the ERP method in the current study. It is then likely that the early negative response reported in the present and in O'Sullivan et al.'s (2015) study reflect at least partly the same underlying spectrotemporal computations. O'Sullivan et al., however found an effect of the coherence level on the onset latency (the first time point that significantly differed from zero) of their response: lower levels of coherence elicited responses with longer onset latencies. This effect held for stimuli with 6,8,or 10 coherence levels, but not for coherence levels of 2 or 4. In the current study stimuli with 4 vs. 6 coherence levels were tested and no coherence effect on the peak latency of the ORN response was found. One explanation is that the correlation between coherence level and the onset latency of the response only holds for more salient auditory objects. Another alternative is that the onset latency is more sensitive to coherence levels than the peak latency.

There are, however, other event-related brain responses that may also be related to the current early response. Most notable of them is the auditory evoked awareness related negativity (ARN, Gutschalk et al., 2008). ARN was described in an auditory detection task in which listeners were instructed to detect a repeating tone embedded in a stochastic multi-tone background (masker). This paradigm is similar to the current one. The main differences are that in Gutschalk et al. (2008) study, only a single tone was repeated and that it was separated in frequency from the tones of the background by a protected band surrounding the frequency of the target tone. Gutshalk and colleagues observed an auditory cortical magnetoencephalographic response in the latency range of 50–250 ms, which was elicited by detected targets and also in a passive condition (with higher amplitudes for cued than uncued repeating tones). The authors did not discuss the relation of the response they termed ARN to the ORN. One possibility is that the two components are similar and the current early response matches both. However, the ORN and the ARN may also be separate components. One possible difference between them is that whereas ORN was found rather insensitive to task load (Alain and Izenberg, 2003), no ARN was obtained when the ARN-eliciting stimulus was presented to one ear while attention was strongly focused on sound a presented to the opposite ear (Gutschalk et al., 2008). However, the two tests of attention are not compatible. Thus they do not definitively prove whether ORN and ARN are different responses or not. In the current study, the auditory stimuli were always task-relevant. Therefore, if the ORN and ARN components differ from each other, further experiments are needed to determine which if any matches the observed early negative response.

The N2 ERP responses are also elicited in the same latency range. However, the current early negative ERP response cannot be analogous to either the N2b or the MMN component. Unlike to the N2b, the current early response was found to be generated in the temporo-parietal regions (see source localization results), and unlike to the MMN, the current early response was elicited even though the figure and control trials were delivered with equal probabilities.

The ORN and the P400 amplitude increased together with figure coherence and duration, both of which increase the salience of the figure, as shown by the behavioral results. Further the P400 peak latency decreased with increasing figure coherence. These findings suggest that both the ORN and P400 reflect processes affected by the integrated impact of the different cues of concurrent sound segregation rather than processes affected by individual cues (cf. Kocsis et al., 2014). This conclusion is also compatible with results of studies in the visual domain, which demonstrated that in a visual figure identification task neural responses emerging at about 200 ms reflect perceptual salience rather than physical cue contrast (Straube et al., 2010). The fact that the ORN peak latency increased together with figure duration increasing from 3 to 4 but not from 4 to 5 segments suggests that ORN reflects the outcome of temporal integration of the cues, at least until some threshold is reached (sufficient evidence is gathered for the presence of multiple concurrent sounds).

The P400 amplitude was significantly correlated with figure detection performance, at least when figure salience was sufficiently high so that detection performance was above chance level. Hence, the inverse relationship between P400 amplitude and task difficulty is clear for stimuli above the perceptual threshold. A similar relationship to behavioral sensitivity has been reported for the P300 component (see Polich and Kok, 1995; Polich, 2007). Convergent results were obtained in a visual figure identification task: Straube et al. (2010) found that increasing the salience of the visual object resulted in increasing P300 amplitudes. An alternative explanation would suggest that P400 reflects attention capture by the presence of the figure. Although one cannot rule out this alternative based on the current results, P400 was found to be elicited by mistuning a partial of a complex tone even when tones with mistuned partials appeared with higher probability than

fully harmonic ones within the sequences (Alain et al., 2001), making it unlikely that they would have captured attention. There is one more result dissociating ORN and P400 within the current data: Whereas no significant interaction was observed between the effects of the two cues of figure–ground segregation on the ORN amplitude, the effects of the two cues interacted significantly for the P400 amplitude as well as for discrimination performance (in Experiment 2). Thus, the P400 amplitude is linked directly to behavioral performance in two different ways, whereas the ORN amplitude does not show a similar correspondence to behavior. Furthermore, while ORN is elicited in passive situations (similarly to the brain electric activity observed by O'Sullivan et al., 2015) and has been observed in newborns and 6-month-old infants (Bendixen et al., 2015; Folland et al., 2012), P400 is only elicited when listeners are instructed to report whether they heard one or two concurrent objects (e.g., Alain et al., 2001; McDonald and Alain 2005; Kocsis et al., 2014). These results suggest that ORN reflects the likelihood of the presence of two or more concurrent sounds (the outcome of cue evaluation), whereas P400 relates to the outcome of perceptual decisions (Alain, 2007; Synder and Alain, 2007). The lack of interaction between the effects of the spectral and the temporal figure–ground segregation cue on ORN suggests that these cues independently affect the auditory system's assessment of the likelihood that multiple concurrent sounds are present in an acoustic mixture. Moreover, the significant interaction found between the P400 amplitude and discrimination performance hints that perceptual decisions are nonlinearly related to this likelihood, at least for high likelihoods.

Our source localization results suggest that in both the early (ORN) and the late (P400) time intervals, the temporal cortices are involved in the segregation of the figure from the rest of the acoustic scene. This result is in line with previous reports about the sources of concurrent sound segregation-related ERP components (Alain and McDonald, 2007; Snyder et al., 2006; Wilson et al., 2007) and also with the location of the effects of concurrent sound segregation on transient and steady-state evoked responses, as well as induced gamma oscillations (Bidet-Caulet et al., 2007; Bidet-Caulet et al., 2008; Bidet-Caulet and Bertrand, 2009). ERP studies showed that the source waveforms of ORN and P400 were located in bilateral regional dipoles of the primary auditory cortex, whereas direct electrophysiological recording from auditory cortex revealed the involvement of secondary auditory areas, such as the lateral superior temporal gyrus. Furthermore, in auditory cortex, attention to a foreground object leads to sustained steady state power and phase coherence (regular auditory targets) compared to attention to an irregular background (Elhilali et al., 2009). In Elhilali and colleagues' study, the enhancement varied with the salience of the target. For the same type of stimuli as the current study, a previous fMRI study showed that activity in the intraparietal and superior temporal sulci increased when the stimulus parameters promoted the perception of two streams as opposed to one (Teki et al., 2011). However, in contrast to our experimental design, the BOLD responses were recorded during a passive listening condition and analyzed over the whole duration of the stimuli. Thus it is possible that whereas the auditory cortical electrophysiological responses evoked or induced by the emergence of the figure reflect processes directly involved in detecting the emergence of auditory objects and making perceptual decisions, the full network of perceptual object representations extends also to higher auditory cortical and parietal areas. Consistent with this, we find that in the ORN time window, stimuli including a figure elicited higher activity than control trials in areas of the superior temporal gyrus and the inferior parietal sulcus (angular gyrus), which are also linked with attention towards salient features (for review see Seghier, 2012). The scalp distributions of the figure–ground segregation related neural activity found by O'Sullivan et al. (2015) are compatible with the current observations. The angular gyrus is known to receive connections from the parahippocampal gyrus (Rushworth et al., 2006), which have been shown to have greater activity in response to figure than control stimuli at both the ORN and the P400 time windows. Further, the anterior cingulate cortex (ACC, BA 25, 33), which also showed higher activity for figure than for control stimuli in both time windows, has previously been associated with attentional control processes (Wang et al., 2009). Finally, further brain regions associated with attention control, such as the medial and superior frontal gyri (BA 6, 32, 31) showed higher activation during figure than control trials in the P400 time window. Although the current localization results are either compatible with those of previous studies localizing the neural generators responsible of figure–ground segregation or they can be interpreted in a consistent manner, nevertheless, the precision of our source localization is restricted by the relatively low number of electrodes ($N = 64$), the lack of individual digitization of structural MRI scans and the general limitations of the solutions for EEG source localization (the accuracy with which a source can be located is affected by the factors such as head-modelling errors, source-modelling errors, and instrumental or biological EEG noise, for review see Grech et al., 2008; Whittingstall et al., 2003).

Summary

Figures with multiple temporally coherent tonal components can be perceptually separated from a randomly varying acoustic ground. Two ERP responses, the ORN and the P400, were elicited when listeners detected the emergence of figures in this situation. Both of these components were at least partly generated in auditory cortex. The ORN and P400 amplitudes were correlated with the salience of the figure, but only the P400 amplitude was correlated with behavioral detection performance. The figures used in our study were defined by their spectro-temporal structure: their emergence depended jointly on integrating information over both time (duration) and frequency (coherence). Our results suggest that auditory cortex is involved in both the integration across time and frequency and the grouping of sound that leads to the emergence of such a figure. ORN probably reflects the likelihood of the presence of multiple concurrent sounds based on the evaluation of the available perceptual cues, whereas P400 appears to be related to the perceptual decision. These ERP components are reliably elicited even in stimulus configurations the complexity of which approaches that of real-life auditory scenes.

Conflict of interest statement

The manuscript has not been previously published or submitted for publication elsewhere. The authors declare no conflict of interest. Our study complies with the ethical standards laid down in the 1964 Declaration of Helsinki. I will serve as the corresponding author for this manuscript. All of the coauthors have agreed to the order of authorship and to submission of the manuscript in the present form.

Acknowledgments

This work was funded by the Hungarian Academy of Sciences (Magyar Tudományos Akadémia [MTA], post-doctoral fellowship and internship of Erasmus Mundus Student Exchange Network in Auditory Cognitive Neuroscience to B.T. and the MTA Lendület project (LP2012-36/2012) to I.W and National Institutes of Deafness and Communication Disorders R01 DC013825 to Barbara G. Shinn-Cunningham. The authors are grateful to Tamás Kurics for programming assistance and Emese Várkonyi, Zsófia Zavecz, Csenge Török for collecting the EEG data.

References

Alain, C., 2007. Breaking the wave: effects of attention and learning on concurrent sound perception. Hear. Res. 229 (1–2), 225–236.
Alain, C., Izenberg, A., 2003. Effects of attentional load on auditory scene analysis. J. Cogn. Neurosci. 15 (7), 1063–1073.
Alain, C., McDonald, K.L., 2007. Age-related differences in neuromagnetic brain activity underlying concurrent sound perception. J. Neurosci. 27 (6), 1308–1314.

Alain, C., Arnott, S.R., Picton, T.W., 2001. Bottom-up and top-down influences on auditory scene analysis: evidence from event-related brain potentials. J. Exp. Psychol. Hum. Percept. Perform. 27, 1072–1089.
Alain, C., Schuler, B.M., McDonald, K.L., 2002. Neural activity associated with distinguishing concurrent auditory objects. The Journal of the Acoustical Society of America 111 (2), 990–995.
Bendixen, A., Denham, S.L., Gyimesi, K., Winkler, I., 2010a. Regular patterns stabilize auditory streams. The Journal of the Acoustical Society of America 128 (6), 3658–3666.
Bendixen, A., Jones, S.J., Klump, G., Winkler, I., 2010b. Probability dependence and functional separation of the object-related and mismatch negativity event-related potential components. NeuroImage 50, 285–290.
Bendixen, A., Háden, G.P., Németh, R., Farkas, D., Török, M., Winkler, I., 2015. Newborn infants detect cues of concurrent sound segregation. Dev. Neurosci. 37 (2), 172–181.
Bidet-Caulet, A., Bertrand, O., 2009. Neurophysiological mechanisms involved in auditory perceptual organization. Front. Neurosci. 3 (09), 182–191.
Bidet-Caulet, A., Fischer, C., Besle, J., Aguera, P.-E., Giard, M.-H., Bertrand, O., 2007. Effects of selective attention on the electrophysiological representation of concurrent sounds in the human auditory cortex. J. Neurosci. 27, 9252–9261.
Bidet-Caulet, A., Fischer, C., Bauchet, F., Aguera, P.-E., Bertrand, O., 2008. Neural substrate of concurrent sound perception: direct electrophysiological recordings from human auditory cortex. Front. Hum. Neurosci. 1, 5.
Bizley, J.K., Cohen, Y.E., 2013. The what, where and how of auditory-object perception. Nat. Rev. Neurosci. 14 (10), 693–707.
Bregman, A.S., 1990. Auditory Scene Analysis: The Perceptual Organization of Sound. MIT Press, Cambridge, MA.
Carlyon, R.P., Cusack, R., Foxton, J.M., Robertson, I.H., 2001. Effects of attention and unilateral neglect on auditory stream segregation. J. Exp. Psychol. Hum. Percept. Perform. 27, 115–127.
Ciocca, V., 2008. The auditory organization of complex sounds. Front. Biosci. 13, 148–169.
Darwin, C.J., 1997. Auditory grouping. Trends Cogn. Sci. 1 (9), 327–333.
Delorme, A., Sejnowski, T., Makeig, S., 2007. Improved rejection of artifacts from EEG data using high-order statistics and independent component analysis. NeuroImage 34, 1443–1449.
Delorme, A., Makeig, S., 2004. EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. J. Neurosci. Methods 134 (1), 9–21.
Denham, S.L., Winkler, I., 2014. Auditory perceptual organization. In: Wagemans, J. (Ed.), Oxford Handbook of Perceptual Organization. Oxford University Press, Oxford, U.K., pp. 601–620.
Devergie, A., Grimault, N., Tillmann, B., Berthommier, F., 2010. Effect of rhythmic attention on the segregation of interleaved melodies. J. Acoust. Soc. Am. 128, EL1–EL7.
Dolležal, L.V., Beutelmann, R., Klump, G.M., 2012. Stream segregation in the perception of sinusoidally amplitude-modulated tones. PLoS One 7 (9), e43615.
Dowling, W.J., Lung, K.M., Herrbold, S., 1987. Aiming attention in pitch and time in the perception of interleaved melodies. Percept. Psychophys. 41, 642–656.
Elhilali, M., Xiang, 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 Biol. 7 (6), 1000129.
Elhilali, M., Ma, L., Micheyl, C., Oxenham, A.J., Shamma, S.A., 2010. Representation of auditory scenes. Computer 61 (2), 317–329.
Folland, N., Butler, B.E., Smith, N., Trainor, L.J., 2012. Processing simultaneous auditory objects: infants' ability to detect mistuning in harmonic complexes. The Journal of the Acoustical Society of America 131 (1), 993.
Grech, R., Cassar, T., Muscat, J., Camilleri, K.P., Fabri, S.G., Zervakis, M., Xanthopoulos, P., Sakkalis, V., Vanrumste, B., 2008. Review on solving the inverse problem in EEG source analysis. Journal of NeuroEngineering and Rehabilitation 5, 25.
Green, D.M., Swets, J.A., 1988. Signal Detection Theory and Psychophysics. Peninsula Publishing, Los Altos, CA.
Griffiths, T.D., Warren, J.D., 2004. What is an auditory object? Nature reviews. Neuroscience 5 (11), 887–892.
Grimault, N., Bacon, S.P., Micheyl, C., 2002. Auditory stream segregation on the basis of amplitude-modulation rate. J. Acoust. Soc. Am. 111, 1340–1348.
Gutschalk, A., 2005. Neuromagnetic correlates of streaming in human auditory cortex. J. Neurosci. 25 (22), 5382–5388.
Gutschalk, A., Micheyl, C., Oxenham, A.J., 2008. Neural correlates of auditory perceptual awareness under informational masking. PLoS Biol. 6 (6), e138.
Johnson, B.W., Hautus, M., Clapp, W.C., 2003. Neural activity associated with binaural processes for the perceptual segregation of pitch. Clin. Neurophysiol. 114 (12), 2245–2250.
Johnson, B.W., Hautus, M.J., Duff, D.J., Clapp, W.C., 2007. Sequential processing of interaural timing differences for sound source segregation and spatial localization: evidence from event-related cortical potentials. Psychophysiology 44 (4), 541–551.
Jones, M., Kidd, G., Wetzel, R., 1981. Evidence for rhythmic attention. J. Exp. Psychol. Hum. Percept. Perform. 7, 1059–1073.
Kocsis, Z., Winkler, I., Szalárdy, O., Bendixen, A., 2014. Effects of multiple congruent cues on concurrent sound segregation during passive and active listening: an event-related potential (ERP) study. Biol. Psychol. 100 (1), 20–33.
Kubovy, M., van Valkenburg, D., 2001. Auditory and visual objects. Cognition 80 (1–2), 97–126.
Lee, A.K., Shinn-Cunningham, B.G., 2008. Effects of frequency disparities on trading of an ambiguous tone between two competing auditory objects. J. Acoust. Soc. Am. 123, 4340–4351.
Lipp, R., Kitterick, P., Summerfield, Q., Bailey, P.J., Paul-Jordanov, I., 2010. Concurrent sound segregation based on inharmonicity and onset asynchrony. Neuropsychologia 48 (5), 1417–1425.
McDonald, K.L., Alain, C., 2005. Contribution of harmonicity and location to auditory object formation in free field: Evidence from event-related brain potentials. J. Acoust. Soc. Am. 118 (3), 1593–1604.
Micheyl, C., Carlyon, R.P., Gutschalk, A., Melcher, J.R., Oxenham, A.J., Rauschecker, J.P., Courtenay Wilson, E., 2007. The role of auditory cortex in the formation of auditory streams. Hear. Res. 229 (1–2), 116–131.
Micheyl, C., Kreft, H., Shamma, S., Oxenham, A.J., 2013. Temporal coherence versus harmonicity in auditory stream formation. J. Acoust. Soc. Am. 133 (3), 188–194.
Mill, R.W., Böhm, T.M., Bendixen, A., Winkler, I., Denham, S.L., 2013. Modelling the emergence and dynamics of perceptual organisation in auditory streaming. PLoS Comput. Biol. 9 (3), 1–21.
Moore, B.C.J., Gockel, H., 2002. Factors influencing sequential stream segregation. Acta Acustica United with Acustica 88 (3), 320–333.
O'Sullivan, J.A., Shamma, A.S., Lalor, E.C., 2015. Evidence for neural computations of temporal coherence in an auditory scene and their enhancement during active listening. J. Neurosci. 35 (18), 7256–7263.
Pascual-Marqui, R.D., Esslen, M., Kochi, K., Lehmann, D., 2002. Functional imaging with low resolution brain electromagnetic tomography (LORETA): review, new comparisons, and new validation. Japanese Journal of Clinical Neurophysiology 30, 81–94.
Polich, J., 2007. Updating P300: an integrative theory of P3a and P3b. Clin. Neurophysiol. 118 (10), 2128–2148.
Polich, J., Kok, A., 1995. Cognitive and biological determinants of P300: an integrative review. Biol. Psychol. 41, 103–146.
Rushworth, M.F., Behrens, T.E., Johansen-Berg, H., 2006. Connection patterns distinguish 3 regions of human parietal cortex. Cereb. Cortex 16, 1418–1430.
Seghier, M.L., 2012. The angular gyrus: multiple function ad multiple subdivisions. Neuroscientist 19 (1), 43–61.
Shamma, S.A., Elhilali, M., Micheyl, C., 2011. Temporal coherence and attention in auditory scene analysis. Trends Neurosci. 34 (3), 114–123.
Shamma, S., Elhilali, M., Ma, L., Micheyl, C., Oxenham, A.J., Pressnitzer, D., Yin, P., Xu, Y., 2013. Temporal coherence and the streaming of complex sounds. Adv. Exp. Med. Biol. 787, 535–543.
Shinn-Cunningham, B.G., 2008. Object-based auditory and visual attention. Trends Cogn. Sci. 12 (5), 182–186.
Snyder, J.S., Alain, C., 2007. Toward a neurophysiological theory of auditory stream segregation. Psychol. Bull. 133 (5), 780–799.
Snyder, J.S., Alain, C., Picton, T.W., 2006. Effects of attention on neuroelectric correlates of auditory stream segregation. J. Cogn. Neurosci. 18 (1), 1–13.
Spielmann, M.I., Schröger, E., Kotz, S.A., Bendixen, A., 2014. Attention effects on auditory scene analysis: insights from event-related brain potentials. Psychol. Res. 78 (3), 361–378.
Straube, S., Grimsen, C., Fahle, M., 2010. Electrophysiological correlates of figure-ground segregation directly reflect perceptual saliency. Vis. Res. 50 (5), 509–521.
Sussman, E.S., Ritter, W., Vaughan, H.G., 1999. An investigation of the auditory streaming effect using event-related brain potentials. Psychophysiology 36 (1), 22–34.
Szalárdy, O., Bendixen, A., Tóth, D., Denham, S.L., Winkler, I., 2013. Modulation frequency difference acts as a primitive cue for auditory stream segregation. Learning & Perception 5 (2), 149–161.
Szalárdy, O., Bendixen, A., Böhm, T.M., Davies, L.A., Denham, S.L., Winkler, I., 2014. The effects of rhythm and melody on auditory stream segregation. J. Acoust. Soc. Am. 135 (3), 1392–1405.
Teki, S., Chait, M., Kumar, S., von Kriegstein, K., Griffiths, T.D., 2011. Brain bases for auditory stimulus-driven figure-ground segregation. J. Neurosci. 31, 164–171.
Teki, S., Chait, M., Kumar, S., Shamma, S., Griffiths, T.D., 2013. Segregation of complex acoustic scenes based on temporal coherence. E Life 2, 00699.
Van Noorden, L.P.A.S., 1975. Temporal Coherence in the Perception of Tone Sequences Unpublished doctoral dissertation Eindhoven University of Technology.
Wang, L., Liu, X., Guise, K.G., Knight, R.T., Ghajar, J., Fan, J., 2009. Effective connectivity of the fronto-parietal network during attentional control. J. Cogn. Neurosci. 22 (3), 543–553.
Weise, A., Bendixen, A., Müller, D., Schröger, E., 2012. Which kind of transition is important for sound representation? An event-related potential study. Brain Res. 1464, 30–42.
Whittingstall, K., Stroink, G., Gates, L., Connolly, J.F., Finley, A., 2003. Effects of dipole position, orientation and noise on the accuracy of EEG source localization. Biomedical Engineering Online 2, 14.
Wilson, E.C., Melcher, J.R., Micheyl, C., Gutschalk, A., Oxenham, A.J., 2007. Cortical FMRI activation to sequences of tones alternating in frequency: relationship to perceived rate and streaming. J. Neurophysiol. 97, 2230–2238.
Winkler, I., Denham, S.L., Nelken, I., 2009. Modeling the auditory scene: predictive regularity representations and perceptual objects. Trends Cogn. Sci. 13 (12), 532–540.
Winkler, I., Denham, S.L., Mill, R., Böhm, T.M., Bendixen, A., 2012. Multistability in auditory stream segregation: a predictive coding view. Philos. Trans. R. Soc. B 367, 1001–1012.

3.2. *Study II: Effects of multiple congruent cues on concurrent sound segregation during passive and active listening: An event-related potential (ERP) study*

Kocsis, Z., Winkler, I., Szalárdy, O., & Bendixen, A. (2014). Effects of multiple congruent cues on concurrent sound segregation during passive and active listening: An event-related potential (ERP) study. Biological Psychology, 100, 20-33.DOI: 10.1016/j.biopsycho.2014.04.005.

Biological Psychology 100 (2014) 20–33

Contents lists available at ScienceDirect

Biological Psychology

journal homepage: www.elsevier.com/locate/biopsycho

Effects of multiple congruent cues on concurrent sound segregation during passive and active listening: An event-related potential (ERP) study

Zsuzsanna Kocsis[a,b,*], István Winkler[a,c], Orsolya Szalárdy[a], Alexandra Bendixen[d,e]

[a] Institute of Psychology and Cognitive Neuroscience, Research Centre for Natural Sciences, Hungarian Academy of Sciences, Budapest, Hungary
[b] Budapest University of Technology and Economics, Budapest, Hungary
[c] Institute of Psychology, University of Szeged, Szeged, Hungary
[d] Department of Psychology, Cluster of Excellence "Hearing4all", European Medical School, Carl von Ossietzky University of Oldenburg, Oldenburg, Germany
[e] Department of Psychology, University of Leipzig, Leipzig, Germany

ARTICLE INFO

Article history:
Received 25 October 2013
Accepted 30 April 2014
Available online 9 May 2014

Keywords:
Object-related negativity (ORN)
P400
Multiple congruent cues
Concurrent sound segregation
Active and passive listening

ABSTRACT

In two experiments, we assessed the effects of combining different cues of concurrent sound segregation on the object-related negativity (ORN) and the P400 event-related potential components. Participants were presented with sequences of complex tones, half of which contained some manipulation: one or two harmonic partials were mistuned, delayed, or presented from a different location than the rest. In separate conditions, one, two, or three of these manipulations were combined. Participants watched a silent movie (passive listening) or reported after each tone whether they perceived one or two concurrent sounds (active listening). ORN was found in almost all conditions except for location difference alone during passive listening. Combining several cues or manipulating more than one partial consistently led to sub-additive effects on the ORN amplitude. These results support the view that ORN reflects a combined, feature-unspecific assessment of the auditory system regarding the contribution of two sources to the incoming sound.

1. Introduction

In everyday situations, we are constantly confronted with mixtures of sounds emitted by concurrently active sources. The human auditory system needs to parse this mixture to allow us to perceive the world in terms of meaningful objects and events. Cues that support the parsing process are traditionally divided into two main categories (Bregman, 1990; Carlyon, 2004; Haykin & Chen, 2005; Snyder & Alain, 2007): those that group together sound elements along time (horizontal or sequential sound organization) and those that group them at one particular moment of time (vertical or concurrent sound organization). Concurrent segregation is based on instantaneously available cues, such as differences in pitch, sound onset, and source location. Whereas no direct event-related potential (ERP) correlate of sequential segregation has been discovered yet, concurrent segregation appears to have such an ERP correlate: The object-related negativity (ORN) component has been shown to follow the listener's perception of two concurrent sounds (Alain, Arnott, & Picton, 2001). The present study was designed to systematically investigate how combinations of the three most well-known cues of concurrent sound segregation (different source location, onset asynchrony, and inharmonic relation between the partials of complex tones) affect the ORN component. Specifically, we wished to assess whether ORN sums together the outputs of three independent detectors of concurrent sound segregation, or whether it is a read-out of the system's overall assessment of the likelihood that the sound input carries contributions from two sound sources.[1]

The ORN peaks between 150 and 180 ms from cue onset, reaches its maximum at frontocentral electrode sites, and inverts polarity at the mastoids (Alain, Schuler, & McDonald, 2002; Alain & McDonald, 2007). Alain and colleagues (2001) found that ORN was larger at the mastoid electrodes during active listening (listeners were required

* Corresponding author at: Institute of Cognitive Neuroscience and Psychology, Research Centre for Natural Sciences, Hungarian Academy of Sciences, Magyar tudósok körútja 2, Budapest H-1117, Hungary. Tel.: +36 13826809.
E-mail address: kocsis.zsuzsanna@ttk.mta.hu (Z. Kocsis).

http://dx.doi.org/10.1016/j.biopsycho.2014.04.005
0301-0511/

[1] Another possibility is that the amplitude of the ORN reflects the number of perceived auditory objects, although modulation of the ORN amplitude by the amount of mistuning (Alain et al., 2001) makes this alternative unlikely.

to judge whether they heard one or two concurrent sounds) than passive listening situations (listeners had no task related to the sounds), indicating attentional modulation of the ORN amplitude.

The presence and amplitude of ORN is correlated with manipulations that typically lead to listeners reporting two sound sources compared to one (Alain, Theunissen, Chevalier, Batty, & Taylor, 2003; Alain & McDonald, 2007; McDonald & Alain, 2005). Previous studies have shown that ORN can be elicited by different cues, such as inharmonicity (Alain et al., 2001, 2002; Bendixen, Jones, Klump, & Winkler, 2010), onset asynchrony (Lipp, Kitterick, Summerfield, Bailey, & Paul-Jordanov, 2010; Weise, Schröger, & Bendixen, 2012); dichotic pitch (Johnson, Hautus, & Clapp, 2003; Hautus, Johnson, & Colling, 2009), separation in the fundamental frequency of speech sounds (Alain, Reinke, He, Wang, & Lobaugh, 2005; Snyder & Alain, 2005), and simulated echo (Sanders, Joh, Keen, & Freyman, 2008; Sanders, Zobel, Freyman, & Keen, 2008). There are also some reports of ORN emerging with a combination of some of the above cues, such as inharmonicity and location difference (McDonald & Alain, 2005) or inharmonicity and onset asynchrony (Weise et al., 2012).

ORN is elicited in both passive and active listening situations (Alain et al., 2001, 2002; Alain, 2007) and its amplitude is independent of the task demands (Alain & Izenberg, 2003). In active listening situations, ORN elicitation is accompanied by a late positive wave that peaks about 400 ms after stimulus onset, the P400 component. P400 amplitude also correlates with the likelihood of perceiving two concurrent sound objects compared to one (Alain et al., 2001, 2002; Hautus & Johnson, 2005), but P400 does not follow the ORN in an obligatory manner (Johnson, Hautus, Duff, & Clapp, 2007). Johnson et al. (2007) proposed that P400 is influenced by the task context. In their study, they used two different tasks: In the detection task, participants were to indicate whether they heard dichotic pitch or a control stimulus, whereas in the localization task, only dichotic pitch stimuli were presented, and participants were to decide where the sound was located. In the latter case, no P400 was elicited (Johnson et al., 2007).

Whereas ORN is assumed to reflect an automatic process of detecting the difference between the physical features (e.g., frequency) extracted from the incoming stimulus and a template of the complex sound (e.g., based on its fundamental frequency), P400 appears to reflect a controlled process that uses prior knowledge to extract meaning from the incoming auditory information (Alain et al., 2002; Hautus & Johnson, 2005; Johnson et al., 2007). Studies showing that ORN is not only elicited by harmonic cues suggest that the template underlying ORN also includes information about the timing and source location of the partials of complex sounds.

Previous studies suggested that the ORN amplitude is modulated by the strength or salience of the cues supporting the segregation of concurrent sounds. For example, Alain and colleagues (2001) found larger ORN amplitudes with increasing amounts of inharmonicity (larger ORN amplitude for the 16% than 8%, or 4% mistuning). In this study, participants reported hearing two sounds more often with higher amounts of mistuning. In another paradigm, using dichotic pitch, Clapp, Johnson, and Hautus (2007) found that the largest ORN was elicited in response to the most salient dichotic pitch cue, and the ORN amplitude decreased with decreasing cue saliency. These authors also found a similar pattern for P400 amplitude.

Perception of concurrent sounds can be made more likely not only by strengthening one particular cue (e.g., increasing the amount of mistuning for inharmonicity-based segregation, cf. Alain et al., 2001), but also by combining two different cues (e.g., frequency and location). In this case, the multiple congruent cues may strengthen the impression of the presence of separate sound sources. Using MEG, such a combined effect was found in a speech segregation task (Du et al., 2011). Du and colleagues (2011) hypothesized that separation in both base frequency and source location contribute to speech segregation, and combining these cues would result in additivity or superadditivity between the ORN components elicited by the two cues, separately. They found that the ORN elicited by the combination of the base-frequency separation and the location cue equaled the sum of the responses elicited by the two cues alone. A similar effect of summing two different types of cues was obtained by Hautus and colleagues (2009), although these authors did not directly test whether the effect of the cue combination was strictly additive when compared to the sum of the effects of the single cues alone (see also McDonald & Alain, 2005; Weise et al., 2012).

Here we report the results of a study in which we systematically investigated combinations of inharmonicity, onset asynchrony, and location difference under passive (Experiment 1) and active (Experiment 2) listening conditions. Based on previous studies (Alain et al., 2001, 2002, 2003; McDonald & Alain, 2005), we expected that ORN will be present in both listening situations, whereas P400 will only be present in the active listening situation. First, we tested whether the salience of the harmonicity-based cue can be further increased by mistuning two partials in a congruent manner (as opposed to mistuning only one partial). We hypothesized that mistuning two partials would enhance the ORN amplitude by providing redundant information for harmonicity-based segregation. Second, we aimed to assess the effects of combining different cues of concurrent segregation on ORN and P400. As the cues are congruent in supporting the same decomposition of the input into two sounds in perception, we regard them as redundant with respect to concurrent sound segregation. By investigating whether the effects of combined cues are additive, sub- or superadditive compared to the single-cue effects, our goal was to separate two possible interpretations of the ORN component. It is possible that each cue elicits a separate ORN response and the observed response sums together the individual ORN components. In this case, the ORN elicited by multiple congruent cues will be as large as the summed amplitudes of the ORN components elicited by the contributing cues. This would suggest that ORN reflects processes that are closely related to cue evaluation and farther upstream from what appears in perception. Alternatively, ORN may reflect the system's overall assessment of the likelihood that the auditory input consists of two concurrent sounds. That is, ORN could reflect the readout of a process combining the evidence from the available cues. In this case, depending on the way the cues are combined, we should find sub- or superadditivity between the contributing cues' ORN components. Subadditivity of the contributing cues' ORN amplitudes would occur for example if the cue-combination algorithm evaluated the likelihood of the presence of two concurrent sounds by passing on the signal resulting from the most salient cue. Superadditivity of the ORN amplitudes would occur if cue combination took into account partial cues, which alone would not be sufficient to support the presence of two concurrent sounds. The two methods are not mutually exclusive; thus one may find both sub- and superadditivity depending on the strength of the available contributing cues. Either one of these possibilities would mean that the process reflected by ORN is less directly related to cue evaluation; rather it is closer to what appears in perception.

2. Experiment 1

2.1. Methods

2.1.1. Participants

Twenty healthy volunteers (eight female, mean age 23.5 years, SD = 2.42) participated in the experiment. Participants received modest financial compensation. None of the participants were taking any medication affecting the central nervous system. Prior to the beginning of the experiment, written informed consent was obtained from each participant according to the Declaration of Helsinki after the experimental

procedures and aims of the study were explained to them. The study was approved by the Ethical Committee of the Institute of Cognitive Neuroscience and Psychology, Research Centre for Natural Sciences, Hungarian Academy of Sciences.

2.1.2. Apparatus, stimuli, and procedure

The study was conducted in a sound-attenuated experimental chamber at the Institute of Cognitive Neuroscience and Psychology, Research Centre for Natural Sciences, Hungarian Academy of Sciences.

Complex tones with an intensity of 40 dB sensation level (above hearing threshold, adjusted individually for each participant) were presented binaurally via headphones with a 1100 ms onset-to-onset interval. In each stimulus block, 2 types of tones were presented in random succession with equal probabilities: the "base" tone, a fully harmonic tone of 250 ms duration (including 10 ms rise and 10 ms fall times) comprising the 5 lowest partials (all having the same amplitude and starting in sine phase), and a manipulated version of this tone. The manipulated tones had the same base frequency (see below) and duration as their base versions. The manipulations were administered either to one (the 2nd) or two (2nd and 4th) partials. Three simple manipulations and their combinations (altogether 11 different manipulations) were tested. The simple manipulations were: (a) mistuning the 2nd partial (or 2nd and 4th partials) by +8%, or (b) delaying the same partial(s) by 100 ms (but ending at the same time as the other partials), or (c) delivering the same partial(s) with a different interaural time (ITD) and level difference (ILD) compared to those of the other harmonics (location difference). For the purpose of adding the location manipulation in some of the conditions without making these conditions stand out from the other conditions, the location of each individual tone could take one of two positions in each condition, regardless of whether a location manipulation was applied or not. Hence in each condition, half of the tones were presented with parameters promoting the listener to hear the tones as originating from ca. 45° right and the other half from 45° left from the midline (ITD of ±200 μs and ILD of ±5 dB, applied congruently). Tones with the two perceived locations were delivered in a fully randomized order, which was independent from the manipulation (i.e., the probability of a manipulation was equal for the left and right tones). Thus in all conditions, base-left, base-right, manipulated-left, and manipulated-right tones each made up 25% of the stimuli. The location difference cue for the manipulated partial was set up by employing the opposite parameter combination than for the rest of the partials, thus creating a ca. 90° location difference between the manipulated and the other partials. A summary of the experimental manipulations is given in Table 1.

Each stimulus condition was presented in a separate block consisting of 200 base and 200 manipulated tones. Stimulus blocks commenced with 10 base-version tones, which were excluded from the analyses. In each stimulus block, all tones had the same fundamental frequency, whilst the fundamental frequency changed from block to block. Eleven fundamental frequencies were used with the lowest frequency being 200 Hz, and the rest following in one-semitone steps (i.e., the highest fundamental frequency being 378 Hz). The order of the fundamental frequencies and the order of the different stimulus blocks (conditions) were randomized separately for each participant.

Participants watched a silent, subtitled movie of their own choice on a computer screen that was placed in front of them at a distance of 1.15 m. They were asked to ignore the sounds. Total duration of the experimental blocks amounted to 83 min. Short breaks were inserted between stimulus blocks with at least one longer break, set between the 6th and the 7th stimulus block, when the participant was allowed to leave the chamber. Further longer breaks were inserted if the participant asked for it. The total time of the session (including electrode mounting and removal) was ca. 3 h.

2.1.3. Electrophysiological recording and data analysis

Electroencephalogram (EEG) was continuously recorded with Ag/AgCl electrodes. 63 electrodes were placed on the scalp according to the extended international 10–20 system (Chatrian, Lettich, & Nelson, 1985; Jasper, 1958). An additional electrode was placed on the tip of the nose, which served as the reference. Eye movements were monitored by bipolar recordings from two electrodes placed above and below the left eye (vertical electrooculogram, VEOG) and two placed lateral to the outer canthi of both eyes (horizontal electrooculogram, HEOG). EEG and EOG signals were amplified (0–40 Hz) by SynAmps amplifiers (Neuroscan Inc.), sampled at 500 Hz. Data were resampled to 250 Hz and filtered off-line using a 0.1–30 Hz band-pass finite impulse response (FIR) filter (Kaiser windowed, Kaiser $\beta = 5.65$, filter length 4530 points).

For each tone, an epoch of 400-ms duration including a 100 ms pre-stimulus baseline was extracted from the continuous EEG record. Epochs with an amplitude change exceeding 100 μV at any electrode were rejected from further analysis, which led to retaining 84.0% of the responses on average. Epochs for the two stimulus types (base version and manipulated) were separately averaged for each of the 11 conditions, collapsing over the two possible locations (left vs. right presentation).

Difference waveforms were calculated between ERPs elicited by the manipulated and the corresponding base tones for identifying and measuring the ORN component. Except for the scalp topography analyses, all measurements were taken from the recordings at the Cz electrode. Average ORN amplitudes were measured from 72-ms wide windows centered on the average peak latency for each condition. To account for the observed latency variation between conditions, peak latencies were determined separately for each condition by the jackknifing method (Kiesel, Miller, Jolicœur, & Brisson, 2008; Miller, Ulrich, & Schwarz, 2009). Epochs for the base versions and manipulated tones were averaged separately.

Following visual inspection of the responses, the N1 amplitude differences were also investigated. N1 difference amplitudes were measured from 40-ms wide windows centered on the average peak latency for each condition.

All ERP difference amplitudes were tested against zero using one-sample, two-tailed t tests. For testing possible differences in the ORN amplitudes and scalp distributions across the three single-cue manipulations, ORN amplitudes were averaged separately for the following six electrode clusters: left frontal (Fp1, AF7, AF3, F7, F5, F3, F1), left central (FT7, FC5, FC3, FC1, C5, C3, C1), left parietal (CP5, CP3, CP1, P7, P5, P3, P1), right frontal (Fp2, AF8, AF4, F8, F6, F4, F2), right central (FT8, FC6, FC4, FC2, C6, C4, C2), right parietal (CP6, CP4, CP2, P8, P6, P4, P2). ORN amplitudes and scalp topographies were then compared by a repeated-measures ANOVA with the factors *Manipulation* (3 levels: 2nd partial mistuned vs. 2nd partial delayed vs. 2nd partial with location difference) × *Frontality* (3 levels: frontal vs. central vs. parietal) × *Laterality* (2 levels: left vs. right).

The effects of providing multiple congruent cues were tested by a repeated-measures ANOVA of the ORN amplitudes at Cz with the factors *Number of mistuned partials* (2 levels: 2nd partial vs. 2nd and 4th partials) × *Delay* (2 levels: delay present vs. absent) × *Location difference* (2 levels: location difference present vs. absent). Additivity between cue effects was tested with paired two-tailed t tests comparing the multiple-cues ORN amplitudes with the summed amplitudes of the ORN components elicited by the corresponding cues.

All significant statistical results are reported. ANOVA effects are reported together with the partial η^2 effect size measure. The Greenhouse–Geisser correction was applied when the assumption of sphericity was violated; the ε correction factor is reported in these cases. Post hoc tests for repeated-measures ANOVAs were carried out with the Bonferroni correction of the confidence level for multiple comparisons.

2.2. Results

2.2.1. ORN

ERP responses elicited by the base and the manipulated tones as well as the corresponding difference waveforms are shown in Fig. 1 for all experimental conditions at Cz. In the conditions with delay, the delayed partials commenced 100 ms later, causing the resulting ORN to be delayed. ORN amplitudes were found to be significant in almost all conditions, except for the condition where the 2nd partial was presented with location difference alone (see Fig. 1 and Table 2 for the full list of results).

The ANOVA comparing ORN amplitudes and topographies across the three single cues showed a significant main effect of *Frontality* [$F(2,38) = 12.840$, $p = 0.001$, $\eta^2 = 0.403$, $\varepsilon = 0.63$], which was due to significantly larger amplitudes at frontal ($p = 0.015$) and central ($p < 0.001$) than parietal electrodes. This verifies the typical topography pattern of ORN (e.g. Alain et al., 2002). Importantly, there was no significant interaction between *Manipulation* and either one of the topography factors (*Frontality* or *Laterality*), suggesting that ORN topography did not significantly differ between the 3 types of manipulation. The scalp topographies are shown in Fig. 2, top row.

In the ANOVA assessing the effects of providing multiple congruent cues, no significant effects or interactions were observed for any of the experimental manipulations (*Number of mistuned partials*, *Delay* or *Location difference*), all p values > 0.07. This means that adding delay and/or location difference, and/or mistuning more than one partial, did not significantly change the ORN amplitude as compared to that elicited by mistuning only one partial, the most commonly used condition for studying ORN.

Multiple congruent cues always elicited numerically smaller ORN amplitudes than the sum of the ORN amplitudes elicited by the contributing cues, although the differences did not reach significance in each case. The sum of the contributing cues' ORN amplitude values, the corresponding multiple-cue ORN amplitude and the results of the additivity tests are given in Table 3 for each comparison.

2.2.2. N1

Significant differences between the base and the manipulated tones in the N1 latency range (i.e., preceding the ORN) were found

Table 1
Summary of experimental manipulations.

Condition	Experimental manipulation(s)	1 or 2 partials	Mistuning	Delay	Location
1	2nd partial mistuned	1	+	–	–
2	2nd and 4th partials mistuned	2	+	–	–
3	2nd partial delayed	1	–	+	–
4	2nd partial delayed and mistuned	1	+	+	–
5	2nd and 4th delayed and mistuned	2	+	+	–
6	2nd partial with location difference	1	–	–	+
7	2nd partial mistuned with location difference	1	+	–	+
8	2nd and 4th partials mistuned with location difference	2	+	–	+
9	2nd partial delayed with location difference	1	–	+	+
10	2nd partial mistuned, delayed and with location difference	1	+	+	+
11	2nd and 4th partials mistuned, delayed and with location difference	2	+	+	+

for several conditions, mostly those where the manipulated partials were delayed and/or had a different location than the rest: 2nd partial delayed, 2nd partial mistuned and delayed, 2nd and 4th partials mistuned and delayed, 2nd partial with location difference, 2nd partial delayed with location difference and 2nd and 4th partials mistuned, delayed and with location difference (see Table 2 for a full list of the mean difference amplitudes and the results of the corresponding *t* tests against zero).

2.3. Discussion

In Experiment 1, we studied the ORN components elicited by three different cues of concurrent sound segregation and their combination in a passive listening situation. All of these cues and cue combinations elicited significant ORN components, except for location difference alone, which appeared to be a weaker cue of ORN elicitation with the current parameters.

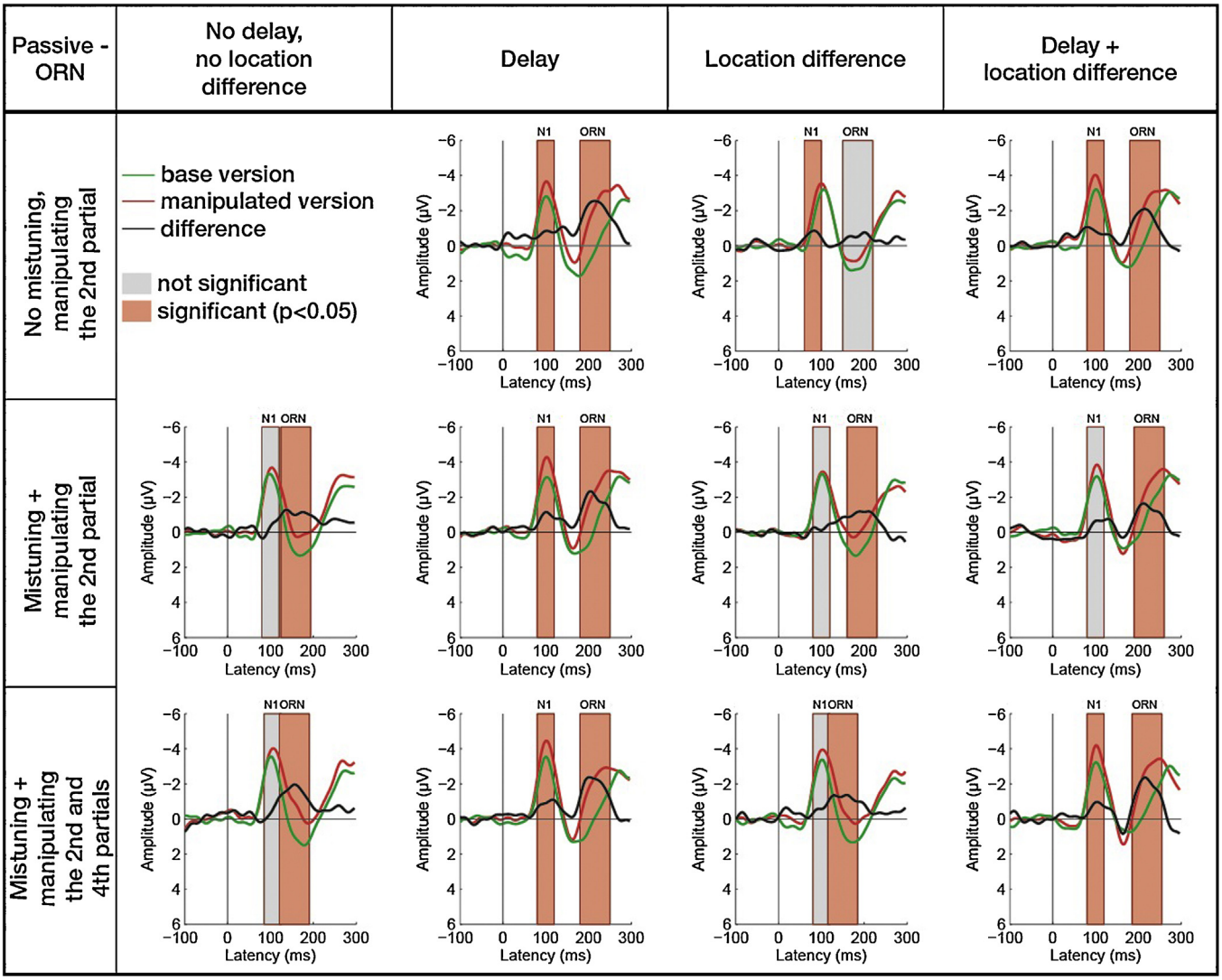

Fig. 1. Grand-average ($N = 20$) ERPs elicited at Cz in the 11 conditions of Experiment 1 (passive listening) by the manipulated (red lines) and base-version tones (green), together with their difference waveforms (black). Stimulus onset is at the crossing of the *x* and *y* axes. Note that in the conditions with delay, the delayed partials commenced 100 ms later and the resulting ORN was also delayed.

Table 2
Grand-average ($N = 20$) ERP amplitudes at Cz measured in the N1 (top) and the ORN (bottom) latency range of the manipulated-minus-base difference waveform for the 11 stimulus conditions of Experiment 1 (passive listening).

	2nd partial mistuned	2nd and 4th partials mistuned	2nd partial delayed	2nd partial mistuned and delayed	2nd and 4th mistuned and delayed	2nd partial with location difference	2nd partial mistuned with location difference	2nd and 4th partials mistuned with location difference	2nd partial delayed with location difference	2nd partial mistuned, delayed and with location difference	2nd and 4th partials mistuned, delayed and with location difference
N1											
Mean amplitude at Cz (μV)	−0.1504	−0.5522	−0.6983	−0.8475	−0.8587	−0.6698	−0.2249	−0.5698	−0.8764	−0.3931	−0.6877
$t(19)$	−0.53	−1.4025	−2.2885	−3.4962	−2.8153	−2.3047	−0.9353	−1.8026	−3.184	−1.0468	−2.3775
p	0.6023	0.1769	0.0337*	0.0024**	0.0110*	0.0326*	0.3614	0.0873	0.0049**	0.3084	0.0281*
Time window for measurement (ms)	80–120	84–125	80–120	80–120	80–120	60–100	80–120	80–120	80–120	80–120	80–120
ORN											
Mean amplitude at Cz (μV)	−1.0827	−1.5911	−2.1033	−1.7498	−1.9277	−0.5292	−1.0487	−1.2053	−1.6150	−1.2779	−1.7862
$t(19)$	−3.917	−4.6625	−5.0669	−5.0297	−5.2885	−1.6832	−4.4899	−3.3015	−4.5769	−3.3739	−6.3604
p	<0.001***	<0.001***	<0.001***	<0.001***	<0.001***	0.1087	<0.001***	0.004**	<0.001***	0.003**	<0.001***
Time window for measurement (ms)	124–196	124–196	180–252	180–252	180–252	152–224	160–232	116–188	180–252	192–264	188–260

Note: Significant differences from zero are marked with asterisks.
* $p < .05$.
** $p < .01$.
*** $p < .001$.

We found no significant increase of the ORN amplitude when congruently manipulating multiple partials, i.e., there was no significant difference between those conditions where only one partial was manipulated as compared to those conditions where two partials were manipulated. Similarly, adding delay and/or location difference on top of mistuning did not lead to a significant increase in the ORN amplitude. Furthermore, combining several cues always elicited numerically (and in most cases significantly) smaller ORN amplitudes than the sum of the contributing ORN amplitudes. In other words, multiple congruent cues were processed in a subadditive manner. We found no evidence pointing toward superadditivity for any of the combinations, nor did any of the combinations appear to follow a strictly additive model. Note that the amount of mistuning employed in the current study (+8%) did not force a ceiling effect on the ORN amplitude, as a previous study found an increase of the ORN amplitude by increasing the amount of mistuning from 8 to 16% (Alain et al., 2001). Taken together, these results suggest that ORN may reflect a combined assessment of the likelihood of the presence of two concurrent sounds, as opposed to summing the strength of sensory evidence for the presence of two concurrent sounds.

Some effects of the cues of concurrent sound segregation were observed in a latency range preceding that of the ORN. Specifically, significantly larger N1 components were elicited in conditions where one or two partials were delayed or presented with location difference and in some of the conditions where these cues appeared in combination. These results were unexpected, and the current paradigm was not designed to separate whether the N1 increase was related to the presence of two concurrent sounds or to the specific acoustic manipulations. A test of this issue was therefore included in the follow-up Experiment 2. We introduced control blocks in which only one sound object was delivered at any time. To control for the delay, we tested tones with the two (2nd and 4th) partials omitted; thus the initial 100-ms segment was identical to the delay manipulation, but no additional partials commenced after 100 ms as that would promote concurrent sound segregation. To control for the location of the tones, we recorded responses separately for the two source locations, using only the base versions of the tones. If the acoustic manipulations accounted for the N1 effect, the difference would be apparent between the base versions presented in the two different locations.

Besides these control blocks, the main purpose of the follow-up Experiment 2 was to repeat the manipulations employed in Experiment 1 in an active listening situation. Previous studies (Alain et al., 2001, 2002) have shown that ORN is also elicited during active listening, and that it is followed by a late positive peak (the P400) when listeners are asked to give perceptual judgments as to the presence of one or two concurrent sounds. Thus in Experiment 2 we investigated whether (a) a similar pattern for the processing of the cues and their combinations is observed when participants are asked to attend to the sounds, (b) whether this pattern translates into perceptual judgments of the sounds as coming from one or two sources, (c) how the different cues and combinations affect the P400 response. Finally, we also assessed (d) whether attention affects the ORN amplitude with multiple concurrent cues.

3. Experiment 2

3.1. Methods

3.1.1. Participants

Twenty-three healthy volunteers (twelve female, mean age 22.1 years, SD = 1.62) participated in the experiment. None of the participants had taken part in Experiment 1.

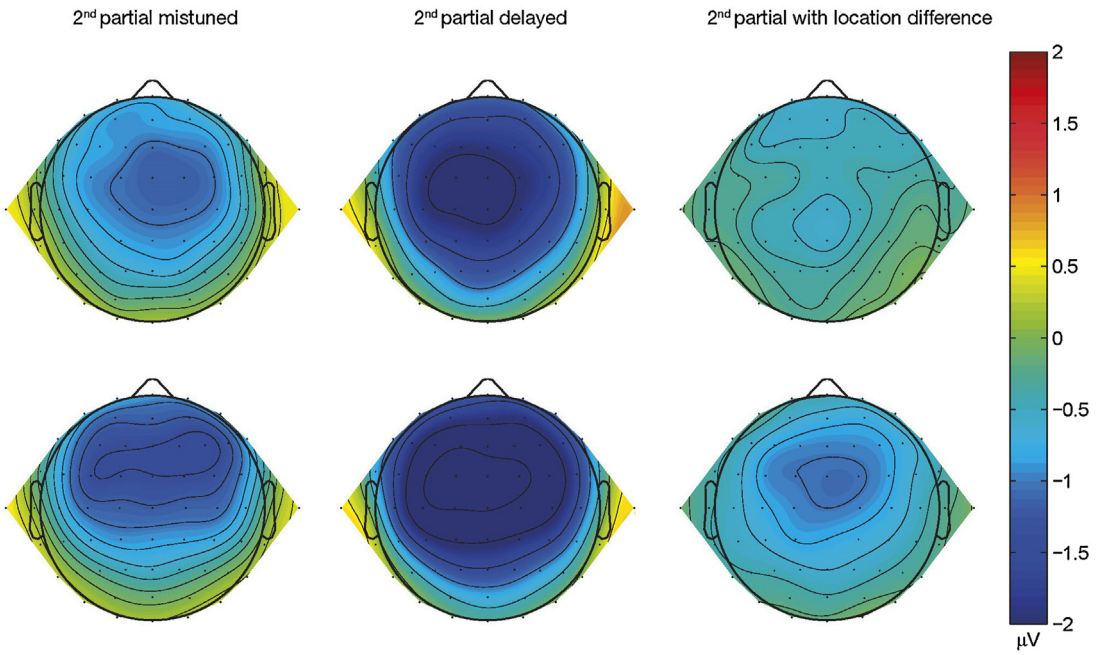

Fig. 2. Grand-average ORN scalp topographies under passive (Experiment 1: $N = 20$; top row) and active (Experiment 2: $N = 23$; bottom row) listening conditions for the three single-cue conditions (mistuning: left column; delay: middle column; location difference: right column). The common voltage scale is placed at the right side of the figure.

3.1.2. Apparatus and stimuli

The stimulus paradigm employed in Experiment 2 was identical to that of Experiment 1 with the following exceptions.

In the beginning of the experimental session, two N1 control blocks were administered. In one stimulus block, two partials (2nd and 4th) were omitted from the complex tone taking the place of the manipulated sounds of the corresponding stimulus block of Experiment 1 (50%), whereas the base version of the sound stayed the same (50%). The order of the base and manipulated sounds was fully randomized. In the other control block, only the base versions of the left and right tones were delivered. As in Experiment 1 as well as all other stimulus blocks of Experiment 2, in the two control blocks, half of the tones were presented with parameters promoting the listener to perceive the tones as originating from ca. 45° right and the other half from 45° left from the midline; tones with the two perceived locations were delivered in a fully randomized order. All other stimulus parameters were identical to those of Experiment 1. During these control blocks, participants watched a subtitled, silent movie and were asked to disregard the sounds.

For the remainder of the session, participants were given two response keys (one in each hand), and were instructed to perform tasks as detailed below by pressing one or the other key with their left or right thumb.

The next (3rd) stimulus block served as control for a different analysis, which is not reported here. In this stimulus block, half of the sounds were base-version complex tones, whereas for the other half, the 2nd and 4th partials were mistuned, delayed, and with location difference as described for Experiment 1 (condition 11). 140 stimuli of the base version and 140 of the manipulated version were delivered with an onset-to-onset interval of 1400 ms. Participants were instructed to watch a fixation cross continuously present at the center of the computer screen placed at 1.15 m directly in front of them (visual angle of 0.4°) and to press either one of the responses buttons when the fixation cross changed to an "X" for 100 ms, after which it returned to the regular "+" sign. The change appeared at a random time

Table 3
Additivity tests for combining cues of concurrent sound segregation during passive listening (Experiment 1). The sum of the amplitudes of the ORN components elicited by the contributing cues and the corresponding multiple-cue ORN amplitudes (at Cz) are given together with the t and p values for the paired two-tailed t tests between them.

Conditions	Mean amplitude at Cz (μV)	$t(19)$	p
2nd partial mistuned + 2nd partial delayed (condition 1 + 3)	−3.186		
2nd partial mistuned and delayed (condition 4)	1.74979	−1.991	0.061
2nd partial mistuned + 2nd partial with location difference (condition 1 + 6)	−1.61194		
2nd partial mistuned and with location difference (condition 7)	−1.04872	−1.306	0.207
2nd partial delayed + 2nd partial with location difference (condition 3 + 6)	−2.63249		
2nd partial delayed and with location difference (condition 9)	−1.61503	−1.59	0.128
2nd partial mistuned + 2nd partial delayed + 2nd partial with location difference (condition 1 + 3 + 6)	−3.71521		
2nd partial mistuned, delayed with location difference (condition 10)	−1.27789	−3.143	0.005**
2nd partial mistuned and delayed + 2nd partial with location difference (condition 4 + 6)	−2.27901		
2nd partial mistuned, delayed and with location difference (condition 10)	−1.27789	−1.871	0.077
2nd partial mistuned and with location difference + 2nd partial delayed (condition 7 + 3)	−3.152		
2nd partial mistuned, delayed and with location difference (condition 10)	−1.27789	−2.431	0.025*
2nd partial delayed and with location difference + 2nd partial mistuned (condition 9 + 1)	−2.69775		
2nd partial mistuned, delayed and with location difference (condition 10)	−1.27789	−2.225	0.038*

Note: Significant differences are marked with asterisks.
* $p < .05$.
** $p < .01$.

point between 550 and 750 ms after each tone onset. Participants were asked to ignore the tones.

Participants then received two blocks of training (blocks 4 and 5) in the task they were asked to do during the rest of the stimulus blocks. In the first training block, 20 base-version tones and 20 tones with both the 2nd and 4th partials mistuned, delayed, and with location difference (see Experiment 1, condition 11) were delivered in a randomized order. In the second training block, 40 tones were presented in a randomized order, 10 of which were of the base version, 10 with the 2nd partial being mistuned, 10 with the 2nd partial being delayed, and 10 with the 2nd partial with location difference. The participants' task was to mark for each tone whether he/she perceived one or two concurrent sounds by depressing one or the other pre-assigned response button. Button assignment remained the same for the rest of the experiment within one participant; it was counterbalanced across participants. Responses were scored as "corresponding" (participant responded 'one sound' for a base tone or 'two sounds' for a manipulated tone) or "non-corresponding" (the converse cases: participant responded 'two sounds' for a base tone or 'one sound' for a manipulated tone). The training blocks were repeated when the percentage of "corresponding" responses was below 65%. None of the subjects needed more than two training sessions.

From the remaining 12 stimulus blocks, 11 blocks (blocks 6–10 and 12–17) matched the stimuli and experimental conditions of Experiment 1, except that the onset-to-onset interval was increased to 1400 ms, and only 140 tones (instead of 200) of both the base and the manipulated tone versions were delivered in each of the 11 conditions. Participants were instructed to indicate whether they perceived one or two sound objects, but mark their answer only once the fixation cross changed to "X" on the screen, which occurred at a random time between 550 and 750 ms after the tone onset. Stimulus blocks commenced with 10 base version sounds, which were not included in either the behavioral or the electrophysiological data analysis.

Between the main stimulus blocks 10 and 12, participants received another control stimulus block (11), the data of which are not reported here. In this stimulus block, no sounds were presented. Participants were instructed to press either one of the response keys when the fixation cross changed to "X". The temporal schedule of delivering the cross-changes was the same as in the other control block (3).

The total net time of the experiment was 104 min. Short and long breaks were inserted as in Experiment 1. The session lasted for ca. 4 h (including instructions, electrode mounting and removal).

3.1.3. Electrophysiological recording and data analysis

Parameters for the EEG recording were identical to Experiment 1, except that signals were sampled at 2000 Hz, and resampled offline to 250 Hz for data analysis.

For each tone, an epoch of 650 ms duration including a 100 ms pre-stimulus baseline was extracted from the continuous EEG record. Epochs with an amplitude change exceeding 100 μV at any electrode were rejected from further analysis, which led to retaining 84.6% of the epochs on average.

For evaluating the perceptual judgments, the percentage of correspondence between the presence or absence of a manipulation and the listeners' judgments (two vs. one sound) was calculated separately for each condition for the base and manipulated versions of the tones. The effects of stimulus condition on the perceptual judgments were assessed by a repeated-measures ANOVA with the factors *Type of tone* (2 levels: base version vs. manipulated version of tones) × *Condition* (11 levels).

Difference waveforms were calculated as described in Experiment 1. For N1 and ORN, the measurements are identical as Experiment 1 and P400 amplitudes were measured in 100-ms wide intervals centered on the average peak latency per condition.

ORN amplitudes underwent the same statistical analyses as employed in Experiment 1. For P400 amplitudes, a one-way repeated-measures ANOVA with the factor *Manipulation* (3 levels: 2nd partial mistuned vs. 2nd partial delayed vs. 2nd partial with location difference) as well as two-tailed paired-sample *t* tests investigating additivity effects were administered.

ORN amplitudes and scalp topographies were also compared between the two experiments by two mixed-model ANOVAs. The ANOVA comparing ORN amplitudes was based on amplitude measures from Cz and had the factors *Listening condition* (2 levels, across groups: active listening vs. passive listening) × *Number of mistuned partials* (2 levels: 2nd partial vs. 2nd and 4th partials) × *Delay* (2 levels: delay present vs. absent) × *Location difference* (2 levels: location difference present vs. absent). The ANOVA comparing the scalp distributions of the single-cue based ORN components had the factors *Listening condition* (2 levels, across groups: active listening vs. passive listening) × *Manipulation* (3 levels: 2nd partial mistuned vs. 2nd partial delayed vs. 2nd partial with location difference) × *Frontality* (3 levels: frontal vs. central vs. parietal) × *Laterality* (2 levels: left, right), where each electrode cluster (as defined for Experiment 1) was represented by the mean amplitude measured from the electrodes in the cluster.

For the N1 control blocks, difference waveforms were calculated between the responses elicited by the two types of tones. N1 amplitudes were measured in 40-ms wide intervals centered on the average peak latency per condition. N1 difference amplitudes were tested against zero using one-sample, two-tailed *t* tests.

In all other respects, methods were the same as in Experiment 1.

3.2. Results

3.2.1. Behavioral data

The percentage of the corresponding one-sound answers to the base versions was 95.43%, while the percentage of corresponding two-sound answers to the manipulated versions was 85.69% (averaged across the 11 conditions; see Table 4).

The repeated-measures ANOVA showed a significant main effect of *Type of tone* [$F(1,22) = 28.699$, $p < 0.001$, $\eta^2 = 0.566$], where the number of corresponding answers to base versions was significantly higher than the corresponding answers to the manipulated versions of tones. There was also a significant main effect of *Condition* [$F(10,220) = 74.298$, $p < 0.001$, $\eta^2 = 0.772$, $\varepsilon = 0.413$], which was due to a significantly smaller number of corresponding answers in the condition with the 2nd partial with location difference than in all the other conditions (all p values < 0.001). A significant interaction of *Type of tone* and *Condition* was also found [$F(10,220) = 63.953$, $p < 0.001$, $\eta^2 = 0.744$, $\varepsilon = 0.235$]. This was due to a significant main effect of *Condition* for corresponding answers to the manipulated versions [$F(10,220) = 95.622$, $p < 0.001$, $\eta^2 = 0.813$, $\varepsilon = 0.286$]. On the other hand, no significant main effect of *Condition* for corresponding answers to the base versions was found [$F(10,220) = 1.578$, $p = 0.2$, $\eta^2 = 0.067$, $\varepsilon = 0.317$].

3.2.2. Electrophysiological data

3.2.2.1. ORN. Fig. 3 shows the ERP responses elicited by the base and the manipulated tones as well as the corresponding difference waveforms for all experimental conditions at Cz. In the conditions with delay, the delayed partials commenced 100 ms later, causing the resulting ORN to be delayed. Significant ORN responses were elicited in all conditions (see Table 5, middle).

The ANOVA comparing ORN amplitudes and scalp topographies (Fig. 2, bottom row) across the three single-cue conditions showed a significant main effect of *Manipulation* [$F(2,44) = 3.840$, $p = 0.029$, $\eta^2 = 0.149$], but the post hoc test only showed a tendency toward significance for the 2nd partial delayed having larger amplitudes than the 2nd partial with location difference ($p = 0.081$). There was also a significant main effect of *Frontality* [$F(2,44) = 19.787$, $p < 0.001$, $\eta^2 = 0.474$, $\varepsilon = 0.635$], which resulted from significantly larger amplitudes in the frontal ($p = 0.005$) and central ($p < 0.001$) than the parietal clusters. No significant interaction of *Manipulation* and either one of the topography factors (*Frontality* or *Laterality*) was observed, suggesting that ORN topography did not differ between the 3 types of manipulation.

In the ANOVA assessing the effects of providing multiple congruent cues, no significant effects or interactions were observed for any of the experimental manipulations (*Number of mistuned partials*, *Delay* or *Location difference*), all p values > 0.12. This is consistent with the pattern of results found in Experiment 1.

In the cue additivity tests, we found that multiple congruent cues always elicited numerically smaller ORN amplitudes than the sum of the ORN amplitudes elicited by the contributing cues, although not all of these differences were significant (see the corresponding amplitude values and statistical test results in Table 6).

3.2.2.2. P400. P400 difference amplitudes were not significant in the following four stimulus conditions: 2nd and 4th partials mistuned, 2nd partial with location difference, 2nd and 4th partials mistuned with location difference and 2nd and 4th partials mistuned, delayed and with location difference (see Table 5, bottom for all results). In the ANOVA comparing the three single-cue conditions, we found a significant main effect of *Manipulation* [$F(2,44) = 6.145$, $p = 0.004$, $\eta^2 = 0.218$,] due to significantly larger amplitudes in the 2nd partial delayed condition than in the 2nd partial with location difference condition ($p = 0.015$). P400 amplitude in the 2nd partial mistuned condition did not significantly

Table 4
Group mean ($N = 23$) percentages of correspondence between the presence vs. absence of cue manipulation and the listener's judgment, as well as percentage of corresponding responses separately for the base and manipulated versions in the 11 experimental conditions.

Conditions	Percentage of responses corresponding to the base version	Percentage of responses corresponding to the manipulated version
2nd partial mistuned	94.78	90.78
2nd and 4th partials mistuned	96.34	90.06
2nd partial delayed	94.88	87.48
2nd partial mistuned and delayed	96.21	93.70
2nd and 4th partials mistuned and delayed	97.30	94.13
2nd partial with location difference	90.75	18.32
2nd partial mistuned and with location difference	94.47	93.91
2nd and 4th partials mistuned and with location difference	97.05	93.79
2nd partial delayed and with location difference	95.96	91.86
2nd partial mistuned, delayed and with location difference	95.90	93.79
2nd and 4th partials mistuned, delayed and with location difference	96.09	94.75

differ from either of the other conditions (both p values > 0.069). The scalp topographies for the P400 components in the three single-cue conditions are shown in Fig. 4.

The cue additivity tests for the P400 showed less homogeneous results than those for the ORN. About half of the comparisons numerically pointed toward sub-, the other half toward super-additivity, while only one comparison in either direction was significant (subadditivity for the combination of 2nd partial delayed plus location difference with 2nd partial mistuned; superadditivity for the combination of 2nd partial delayed with 2nd partial with location difference). No other significant results were obtained (see Table 7 for all results).

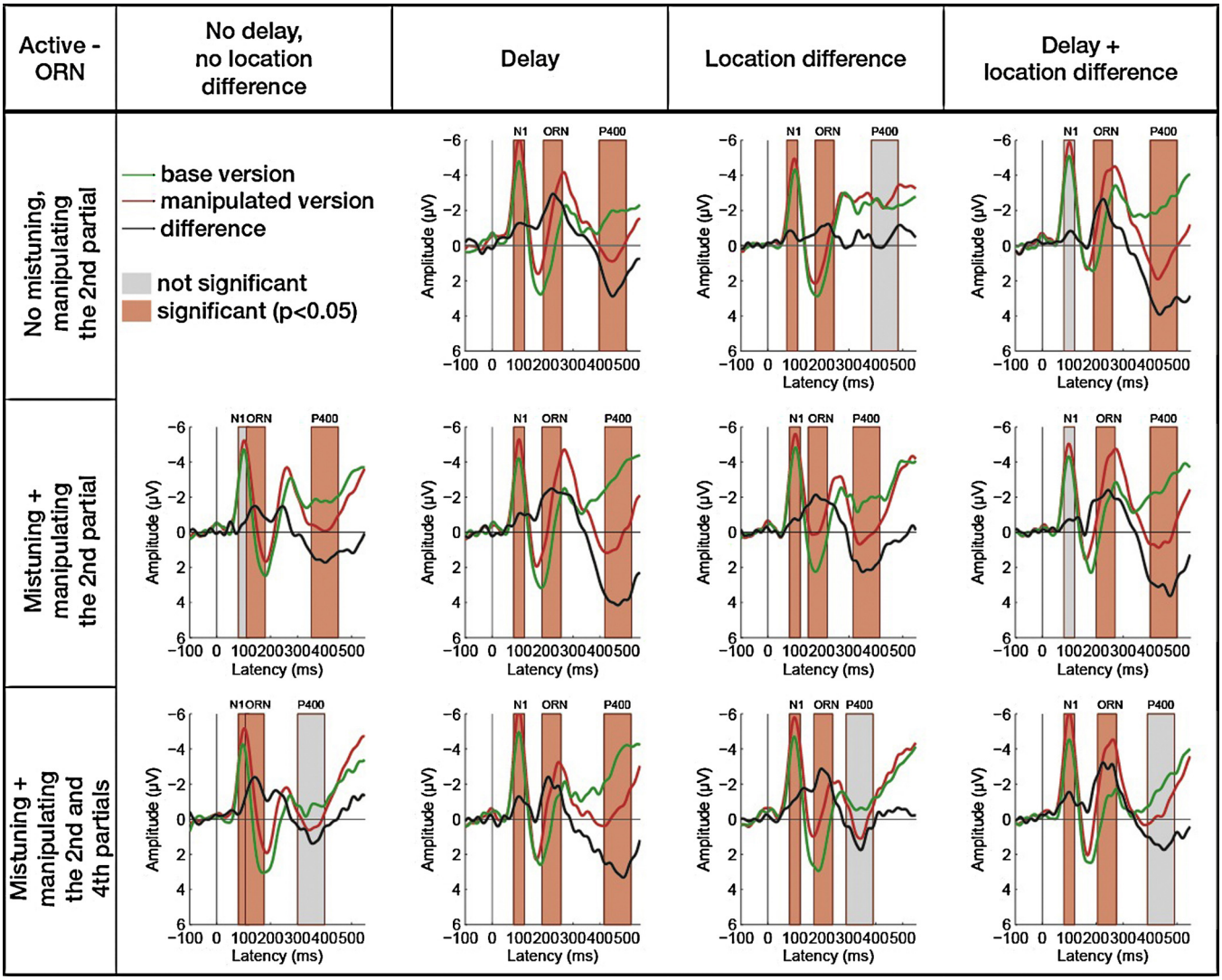

Fig. 3. Grand-average ($N = 23$) ERPs elicited at Cz in the 11 conditions of Experiment 2 (active listening) by the manipulated (red lines) and base-version tones (green), together with their difference waveforms (black). Stimulus onset is at the crossing of the x and y axes. Note that in the conditions with delay, the delayed partials commenced 100 ms later and the resulting ORN was also delayed.

Table 5
Grand-average ($N = 23$) ERP amplitudes at Cz measured in the N1 (top), ORN (middle), and P400 (bottom) latency range of the manipulated-minus-base difference waveforms for the 11 stimulus conditions of Experiment 2 (active listening).

	2nd partial mistuned	2nd and 4th partials mistuned	2nd partial delayed	2nd partial mistuned and delayed	2nd and 4th mistuned and delayed	2nd partial with location difference	2nd partial mistuned with location difference	2nd and 4th partials mistuned with location difference	2nd partial delayed with location difference	2nd partial mistuned, delayed and with location difference	2nd and 4th partials mistuned, delayed and with location difference
N1											
Mean amplitude at Cz (μV)	−0.43356	−0.84169	−1.1528	−0.84069	−1.0484	−0.70872	−0.65448	−1.0609	−0.64766	−0.63116	−1.3187
$t(22)$	−0.96744	−2.3124	−3.5627	−2.5776	−2.8308	−2.4446	−2.1502	−2.8216	−1.603	−1.9076	−5.3944
p	0.344	0.03*	0.002**	0.017*	0.01*	0.023*	0.043*	0.01*	0.123	0.07	<0.001***
Time window for measurement (ms)	80–120	80–120	80–120	80–120	80–120	72–112	80–120	80–120	80–120	80–120	80–120
ORN											
Mean amplitude at Cz (μV)	−1.1785	−1.9523	−2.5055	−2.2847	−1.9592	−0.9867	−1.8816	−2.4056	−2.0920	−2.0828	−2.9394
$t(22)$	−3.3805	−5.2447	−6.0678	−5.061	−3.8983	−2.6315	−4.9561	−5.4187	−5.8903	−3.5371	−11.027
p	0.003**	<0.001***	<0.001***	<0.001***	<0.001***	0.015*	<0.001***	<0.001***	<0.001***	0.002**	<0.001***
Time window for measurement (ms)	112–184	108–180	192–264	188–260	188–260	176–248	152–224	172–244	192–264	200–272	208–280
P400											
Mean amplitude at Cz (μV)	1.4768	0.8754	2.197	3.8848	2.832	−0.2733	1.9441	0.9587	3.429	3.1226	1.5044
$t(22)$	2.5041	1.479	3.3252	4.862	4.032	−0.7086	3.1062	1.7178	5.5549	4.4944	1.4776
p	0.02*	0.153	0.003**	<0.001***	<0.001***	0.486	0.005**	0.1	<0.001***	<0.001***	0.162
Time window for measurement (ms)	352–452	300–400	400–500	420–520	416–516	384–484	316–416	292–392	400–500	400–500	392–492

Note: Significant differences from zero are marked with asterisks.
* $p < .05$.
** $p < .01$.
*** $p < .001$.

Table 6
Additivity tests for combining cues of concurrent sound segregation during active listening (Experiment 2). The sum of the amplitudes of the ORN components elicited by the contributing cues and the corresponding multiple-cue ORN amplitudes (at Cz) are given together with the *t* and *p* values for the paired two-tailed *t* tests between them.

Conditions	Mean amplitude at Cz (μV)	*t*(19)	*p*
2nd partial mistuned + 2nd partial delayed (condition 1 + 3)	−3.683962		
2nd partial mistuned and delayed (condition 4)	−2.284653	−2.474	0.022*
2nd partial mistuned+ 2nd partial with location difference (condition 1 + 6)	−2.165166		
2nd partial mistuned and with location difference (condition 7)	−1.881584	−0.445	0.661
2nd partial delayed + 2nd partial with location difference (condition 3 + 6)	−3.49219		
2nd partial delayed and with location difference (condition 9)	−2.092038	−2.359	0.028*
2nd partial mistuned + 2nd partial delayed + 2nd partial with location difference (condition 1 + 3 + 6)	−4.670659		
2nd partial mistuned, delayed with location difference (condition 10)	−2.082845	−3.073	0.006**
2nd partial mistuned and delayed + 2nd partial with location difference (condition 4 + 6)	−3.271350		
2nd partial mistuned, delayed and with location difference (condition 10)	−2.082845	−1.856	0.077
2nd partial mistuned and with location difference + 2nd partial delayed (condition 7 + 3)	−3.270507		
2nd partial mistuned, delayed and with location difference (condition 10)	−2.082845	−1.408	0.173
2nd partial delayed and with location difference + 2nd partial mistuned (condition 9 + 1)	−4.387077		
2nd partial mistuned, delayed and with location difference (condition 10)	−2.082845	−3.414	0.002**

Note: Significant differences are marked with asterisks.
* $p < .05$.
** $p < .01$.

Table 7
Additivity tests for combining cues of concurrent sound segregation during active listening (Experiment 2). The sum of the amplitudes of the P400 components elicited by the contributing cues and the corresponding multiple-cue P400 amplitudes (at Cz) are given together with the *t* and *p* values for the paired two-tailed *t* tests between them.

Conditions	Mean amplitude at Cz (μV)	*t*(22)	*p*
2nd partial mistuned + 2nd partial delayed (condition 1 + 3)	3.673826		
2nd partial mistuned and delayed (condition 4)	3.884773	−0.223	0.825
2nd partial mistuned+ 2nd partial with location difference (condition 1 + 6)	1.203539		
2nd partial mistuned and with location difference (condition 7)	1.944099	−1.147	0.264
2nd partial delayed + 2nd partial with location difference (condition 3 + 6)	1.923776		
2nd partial delayed and with location difference (condition 9)	3.428958	−2.398	0.025*
2nd partial mistuned + 2nd partial delayed + 2nd partial with location difference (condition 1 + 3 + 6)	3.40057		
2nd partial mistuned, delayed with location difference (condition 10)	3.122613	0.338	0.739
2nd partial mistuned and delayed + 2nd partial with location difference (condition 4 + 6)	3.611517		
2nd partial mistuned, delayed and with location difference (condition 10)	3.122613	0.781	0.443
2nd partial mistuned and with location difference + 2nd partial delayed (condition 7 + 3)	4.14113		
2nd partial mistuned, delayed and with location difference (condition 10)	3.122613	1.39	0.178
2nd partial delayed and with location difference + 2nd partial mistuned (condition 9 + 1)	4.905753		
2nd partial mistuned, delayed and with location difference (condition 10)	3.122613	2.147	0.043*

Note: Significant differences are marked with asterisks.
* $p < .05$.

3.2.2.3. Comparing ORN between the active and the passive listening conditions. In the mixed-model ANOVA comparing the ORN amplitudes between Experiments 1 and 2, only *Listening condition* (i.e., the difference between the two experiments) had a significant effect [$F(1,19) = 6.9535$, $p = 0.016$, $\eta^2 = 0.268$]. ORN amplitudes were larger in the active than in the passive listening situation. No other main effects were observed, and no interactions between *Listening condition* and any of the other factors (*Number of mistuned partials*, *Delay*, and *Location difference*; all *p* values > 0.078).

When comparing the scalp topographies of the ORN components between the passive and active listening conditions, no

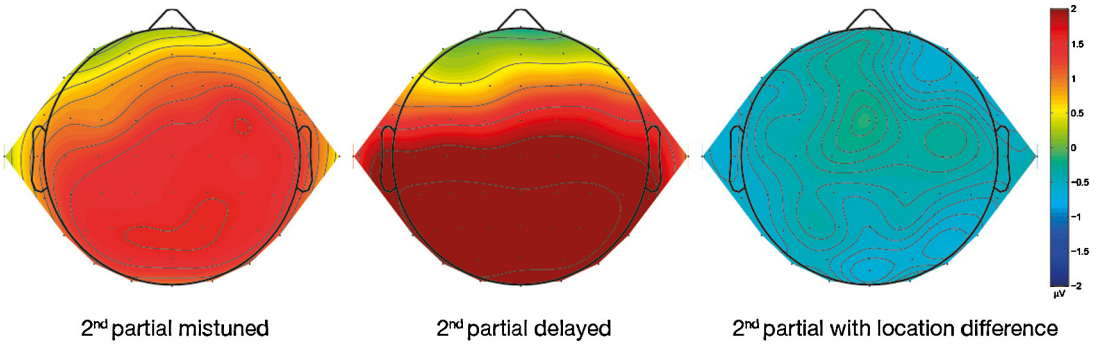

Fig. 4. Grand-average ORN scalp topographies active listening (Experiment 2: $N = 23$) for the three single-cue conditions (mistuning: left column; delay: middle column; location difference: right column) in the P400 time window. The common voltage scale is placed at the right side of the figure.

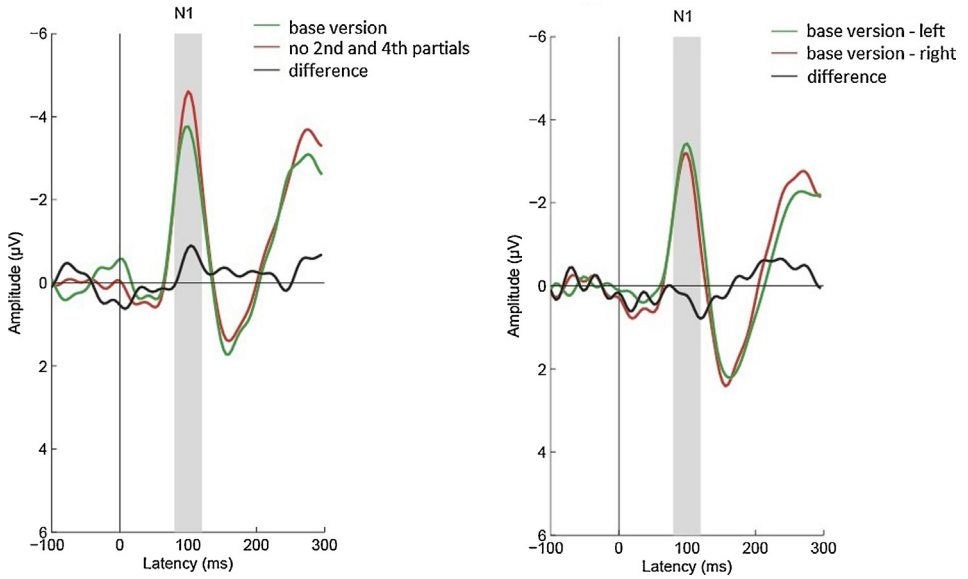

Fig. 5. Grand-average ($N = 23$) ERPs elicited at Cz by the control tones together with their difference waveforms (left: testing the effects of partials being delayed; right: testing the effects of partials delivered from different locations).

main effect of or interaction involving the *Listening condition* was found. The significant main effects of *Manipulation* [$F(2,38) = 4.349$, $p = 0.019$, $\eta^2 = 0.186$] and of *Frontality* [$F(2,38) = 35.206$, $p < 0.001$, $\eta^2 = 0.649$, $\varepsilon = 0.635$] replicated the similar results obtained in the analyses that were conducted separately for the two experiments, with the post hoc tests also showing the same origin for these effects.

3.2.2.4. N1. No significant N1 differences were found between the two different tones in either of the N1 control conditions. In the condition controlling for the effects of delay, the mean amplitude difference at Cz was −0.4809 μV ($t(22) = -1.2619$, $p = 0.22$), and in the condition controlling for the location difference the mean amplitude difference was 0.2798 μV ($t(22) = 0.9284$, $p = 0.36$). The N1 effects are shown in Fig. 5.

In contrast, manipulated tones elicited significantly larger N1 components than the base versions of the tones in most of the 11 stimulus conditions (see Table 5, top for all results).

3.3. Discussion

In Experiment 2, all investigated cues and cue combinations elicited significant ORN responses. Even the condition with location difference alone showed a significant ORN response, although participants' perceptual judgments showed that few of the manipulated tones evoked the perception of two concurrent sounds in this stimulus condition. This suggests that the cues of concurrent sound segregation employed in the current experiment were picked up by the processes underlying ORN generation. On the other hand, ORN cannot fully govern perceptual judgments. Many psychophysical studies have shown that perceptual judgments are codetermined by the criteria used by the observer in deciding between the behavioral alternatives (Green & Swets, 1966). The present instructions did not attempt to manipulate the decision level – listeners were free to choose their criteria. Given that in most stimulus blocks the cues (when present) strongly promoted perception of two concurrent sounds, it is plausible to assume that listeners accommodated to the high distinctiveness of these cues and set their criteria to be high. This should have resulted in dismissing most exemplars with the relatively weak location cue while the obligatory evaluation of the auditory system – as reflected by the elicitation of the ORN component – suggested the presence of two concurrent sound sources. The significantly higher percentage of corresponding responses for the base version tones is also compatible with this interpretation.

In full correspondence with the results obtained in Experiment 1, ORN amplitude did not significantly increase with congruently manipulating two partials (as compared to manipulating only one), or with adding delay and/or location difference on top of mistuning. Furthermore, combining several cues tended to elicit ORN components of smaller amplitude than the sum of the contributing ORN components. No superadditive effects were found for any cue combination. Further, no significant differences were found between the ORN effects or scalp topographies between the two experiments. The results of Experiment 2 thus replicated those obtained in Experiment 1 and support the interpretation given for them in the Discussion of Experiment 1.

The P400 results were not as clear-cut as those for the ORN component. In some cases, no significant P400 was observed, even though the ORN was elicited in those conditions as well. This contrasts the results of some previous studies (Alain et al., 2002). Note, however, that even in the conditions with non-significant P400, positive deflections in the P400 latency range were observed in the difference waveforms between the ERPs elicited by the manipulated and the base tones (cf. Fig. 3). It is possible that the lack of significant detection of the P400 component was caused by some of the preceding ORN components not yet having terminated at the onset of the P400.

In terms of cue redundancy, results for the P400 were equivocal: we found significant superadditivity in one case and significant subadditivity in another case. Altogether there was no clear tendency toward either pattern. Again, these results may be partly obscured by the preceding ORN components.

Significant N1 differences were obtained in most conditions. Unlike in the passive listening situation (Experiment 1), in the active listening situation, these N1 differences were not confined to conditions where delay and location difference manipulations were employed, but also extended to conditions with mistuning alone. The two control conditions suggest that these N1 effects were not solely due to the acoustic differences between the base and

the manipulated tones, but rather they may reflect some aspect of processing concurrent sounds.

4. General discussion

In the present study, we systematically combined three cues of concurrent sound segregation (mistuning, onset asynchrony, and location difference) for testing how the ORN event-related potential component (and in Experiment 2, also the P400 component as well as perceptual judgments) reflects the joint evaluation of these cues. We employed two listening conditions in two separate experiments in which participants were instructed to either disregard the tones (passive listening) or to focus their attention on the tones and judge whether they heard one or two sounds (active listening). The pattern of ORN elicitation in response to the different cues and their combination was highly similar under the two listening conditions. This pattern is consistent with the notion that the ORN response reflects the auditory system's overall assessment of the likelihood that the sound input carries contributions from two sound sources, rather than summing together the outputs of independent detectors of concurrent sound segregation.

We found that ORN was elicited by all of the tested combinations of cues and also by each of the cues individually, with the exception of location difference alone during passive listening. Location difference was also the weakest cue during active listening; although it elicited a small-amplitude ORN, it seldom (<20%) led to the perception of two separate sounds. Our location-cue results are fully consistent with those of McDonald and Alain (2005) who also showed a small ORN for location difference alone during active listening, no significant ORN during passive listening, and little effect of location difference on perceptual segregation. Notably, these authors used a similar amount of location difference (90°) but with free field presentation, which suggests that the lack of an effect in the present study should not be attributed to the artificial manipulation of location via headphones. Instead, the present and previous results (McDonald & Alain, 2005) suggest that location difference alone is not a strong cue of concurrent sound segregation, or at least its effects are easily counteracted by other cues pointing toward integration (i.e., harmonicity and common onset). Alternatively, it is possible that the location cue was not sufficiently salient, although the locations used for the different harmonics could be clearly distinguished as determined by informal perceptual reports. The saliency of the location cue may have been reduced by the fact that for 25% of the tones, all harmonics were delivered at the same location where the location-manipulated harmonics of half of the manipulated tones appeared (see Methods). That is, unlike the mistuned and delayed harmonics, the harmonics separated in location from the other harmonics of the same tone appeared with equal probability as part of tones in which all harmonics were delivered at the same (perceived) location. Bendixen and colleagues (2010) have found an effect of the probability of mistuning on the ORN amplitude. The lack of significant ORN elicitation by the current location-separation cue may indicate a similar contextual effect on ORN. Finally, the lack of consistent perceptual judgments for the location-cue manipulated tones demonstrated that although the information provided by the processes underlying ORN may reflect the full assessment of the auditory system regarding the presence of two concurrently active sound sources, perceptual judgments are codetermined by other effects (cf. the Discussion of the results of Experiment 2).

Mistuning one of the partials of the complex sound or delaying its onset elicited a clear ORN component and led to a robust two-object percept, as was shown in previous studies (e.g. Alain et al., 2001, 2002, 2003; Lipp et al., 2010; Weise et al., 2012). Importantly, ORN amplitude remained unchanged when manipulating not only one (the 2nd) but two (the 2nd and 4th) partials in a congruent manner. We had hypothesized that involving two partials would increase the saliency of the manipulation and thereby boost effects on ORN and perception. Such result patterns have been previously reported for increasing the amount of mistuning (e.g. Alain et al., 2001) or increasing the strength of a dichotic pitch manipulation (Clapp et al., 2007). The present results suggest that manipulating more than one partial of a complex sound does not lead to a similar increase in strength or saliency. Alternatively, it is possible that the 4th partial is not sufficiently influential in assessing the source of complex sounds (cf. Alain et al., 2001, who showed that mistuning the 4th partial alone causes weaker effects than mistuning the 2nd partial alone; therefore, the 2nd partial may have dominated the present results).

Besides manipulating more than one partial, we pursued a second approach for increasing the strength of the sensory evidence in favor of concurrent sound segregation. This approach was based on employing multiple cues in parallel (i.e., onset asynchrony and/or location difference in addition to mistuning). Previous studies suggested that ORN increases with such cue combinations (Hautus et al., 2009; McDonald & Alain, 2005; Weise et al., 2012), and that this increase may follow a fully additive pattern (Du et al., 2011). In contrast, here we mostly found subadditivity for the amplitude of ORN elicited by multiple cues. That is, the combinations of cues of concurrent segregation elicited lower-amplitude ORN components compared with the sum of the ORN amplitudes elicited by each contributing cue separately. In most cases, the increase of the ORN amplitude from single to multiple cues was non-significant, as was shown by the ANOVAs assessing the effects of multiple congruent cues.

One reason for these weak, subadditive effects of the cue combinations may be that the employed cues, at least as far mistuning and onset asynchrony are concerned, were clearly supra-threshold: Each of them alone was sufficient to elicit as much as 90% correspondence between the presence of the cue and the perceptual judgment. Hence the cues can be regarded as fully redundant with respect to each other. Similarly, adding the (weak) location cue on top of a strong cue of concurrent segregation probably also provided only redundant information (cf. McDonald & Alain, 2005, for high amounts of mistuning). Therefore, one may not be surprised that ORN amplitude does not increase further by adding more cues. Note, however, that this explanation implies that the ORN reflects the outcome of a process that combines the sensory evidence to provide an overall assessment of the likelihood that one or two sound sources were present in the environment. If, on the contrary, ORN were to reflect the strength of the sensory evidence that drives the decision between one and two sound sources, then each additional cue should increase ORN amplitude in an additive manner, even if it is redundant with the other cues in terms of the eventual perceptual decision. Our results are not consistent with this latter view; they support the interpretation that ORN underlies the actual perceptual decision. Subadditivity between the ORN components elicited by the contributing cues suggests that the likely method of combining cues of concurrent stream segregation is the selection of the individually most salient cue. Several authors have already argued that the ORN probably reflects a perceptual grouping mechanism rather than a cue-related response; their arguments were based on the highly different nature of the stimuli eliciting ORN (e.g., Hautus & Johnson, 2005; Hautus et al., 2009; Johnson et al., 2007; Lipp et al., 2010; McDonald & Alain, 2005). We add here the argument of subadditivity for multiple congruent cues. Another supporting piece of evidence in the present data is given by the highly similar ORN topographies across different types of cues and listening conditions (cf. Fig. 2).

The P400 component was less tightly related to the perceptual reports. Unlike in previous studies (Alain et al., 2001, 2002; Hautus & Johnson, 2005), P400 failed to reach significance in some conditions despite clear perceptual distinction between one- and two-sound objects. One might speculate that a procedural difference between our and previous studies may account for this (namely, subjects were to withhold their response for several hundred milliseconds). However, a similar dissociation between P400 and perceptual decisions was observed by Hautus and colleagues (2009) with the instruction to respond as quickly as possible. Hence the results are more in line with the view that ORN and P400, as well as P400 and behavior, are not as tightly connected as previously assumed (Johnson et al., 2007). In terms of cue redundancy, there was no clear sub-, super- or fully additive pattern for the P400; the results are thus not informative regarding this question.

Unexpectedly, we observed effects related to the cues of concurrent segregation also in the N1 latency range. Some previous studies reported mistuning-related effects preceding the ORN latency range in MEG (Alain & McDonald, 2007; Lipp et al., 2010). These effects were, however, even earlier than in the N1 range; it remains unclear whether they relate to the present effects. Because our control conditions suggest that the N1 effects were not caused by the acoustic differences between the base and the manipulated stimuli, we tentatively suggest that the N1 differences are related to the automatic processing of the cues of concurrent sound segregation, but not necessarily to the resulting percept (following the interpretations of Alain & McDonald, 2007, as well as Lipp et al., 2010).

Finally, a significant effect of listening condition was found for the ORN amplitude, with larger amplitudes during active than passive listening. This is in accordance with some (e.g. Alain et al., 2001) but not all (e.g. Alain & Izenberg, 2003; Lipp et al., 2010) previous studies. Alain and colleagues (2001) found that attention only affected the ORN amplitude when the same fundamental frequency and number of manipulated partial have been used throughout a stimulus block, but not when the fundamental frequency and/or the manipulated partial were randomly varied. The authors suggested that under constant stimulus conditions, participants start to actively search for the mistuned partial in order to perform the task, and that this search process caused the attention effect. In the present study, although a random variation in perceived location was present, neither the fundamental frequency nor the number of the manipulated partial varied within the stimulus blocks. It is thus possible that the modulation of ORN amplitude by listening condition reflects task-specific preparation during active listening rather than a genuine attention effect on ORN. If this was the case, then the search process assumed by Alain and colleagues (2001) appears to be insensitive to location variation – i.e., location information may not be part of the search template.

In conclusion, we provide evidence for the ORN component reflecting the auditory system's combined assessment as to whether one or more sources contributed to the incoming sound. Our results are not consistent with the view that ORN would directly reflect the processing of the sensory cues that underlie this perceptual decision. This further qualifies the ORN component as an indicator of concurrent sound segregation, and shows that the brain accomplishes this complex operation in a short time (<200 ms).

Acknowledgments

This work was funded by the Hungarian Academy of Sciences (Magyar Tudományos Akadémia [MTA], Lendület project LP2012-36/2012 to I.W.), by the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG Cluster of Excellence 1077 "Hearing4all"), by the German Academic Exchange Service (Deutscher Akademischer Austauschdienst [DAAD], Project 56265741), and by the Hungarian Scholarship Board (Magyar Ösztöndíj Bizottság [MÖB], Project 39589). The experiment was realized using Cogent 2000 developed by the Cogent 2000 team at the FIL and the ICN. EEG data were analyzed with EEGlab (Delorme & Makeig, 2004) and additional plugins written by Andreas Widmann, University of Leipzig. The authors are grateful to Zsuzsanna D'Albini for assistance in data acquisition.

References

Alain, C. (2007). Breaking the wave: Effects of attention and learning on concurrent sound perception. *Hearing Research*, *229*, 225–236.
Alain, C., Arnott, S. R., & Picton, T. W. (2001). Bottom-up and top-down influences on auditory scene analysis: Evidence from brain potentials. *Journal of Experimental Psychology*, *27*, 1072–1089.
Alain, C., & Izenberg, A. (2003). Effects of attentional load on auditory scene analysis. *Journal of Cognitive Neuroscience*, *15*, 1063–1073.
Alain, C., & McDonald, K. L. (2007). Age-related differences in neuromagnetic brain activity underlying concurrent sound perception. *Journal of Neuroscience*, *27*, 1308–1314.
Alain, C., Reinke, K. S., He, Y., Wang, C., & Lobaugh, N. (2005). Hearing two things at once: Neurophysiological indices of speech segregation and identification. *Journal of Cognitive Neuroscience*, *17*, 811–818.
Alain, C., Schuler, B. M., & McDonald, K. L. (2002). Neural activity associated with distinguishing concurrent auditory objects. *Journal of the Acoustical Society of America*, *111*, 990–995.
Alain, C., Theunissen, E. L., Chevalier, H., Batty, M., & Taylor, M. J. (2003). Developmental changes in distinguishing concurrent auditory objects. *Cognitive Brain Research*, *16*, 210–218.
Bendixen, A., Jones, S. J., Klump, G., & Winkler, I. (2010). Probability dependence and functional separation of the object-related and mismatch negativity event-related potential components. *Neuroimage*, *50*, 285–290.
Bregman, A. S. (1990). *Auditory scene analysis: The perceptual organization of sound*. Cambridge, MA: The MIT Press.
Carlyon, R. P. (2004). How the brain separates sounds. *Trends in Cognitive Science*, *8*, 465–471.
Chatrian, G. E., Lettich, E., & Nelson, P. L. (1985). Ten percent electrode system for topographic studies of spontaneous and evoked EEG activity. *American Journal of EEG Technology*, *25*, 83–92.
Clapp, W. C., Johnson, B. W., & Hautus, M. J. (2007). Graded cue information in dichotic pitch: Effects on event-related potentials. *Neuroreport*, *18*, 365–368.
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*, 9–21.
Du, Y., He, Y., Ross, B., Bardouille, T., Wu, X., Li, L., et al. (2011). Human auditory cortex activity shows additive effects of spectral and spatial cues during speech segregation. *Cerebral Cortex*, *21*, 698–707.
Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.
Hautus, M. J., & Johnson, B. W. (2005). Object-related brain potentials associated with the perceptual segregation of a dichotically embedded pitch. *Journal of the Acoustical Society of America*, *117*, 275–280.
Hautus, M. J., Johnson, B. W., & Colling, L. J. (2009). Event-related potentials for interaural time differences and spectral cues. *Neuroreport*, *20*, 951–956.
Haykin, S., & Chen, Z. (2005). The cocktail party problem. *Neural Computation*, *17*, 1875–1902.
Jasper, H. H. (1958). The ten-twenty electrode system of the International Federation. *Electroencephalography and Clinical Neurophysiology*, *10*, 370–375.
Johnson, B. W., Hautus, M., & Clapp, W. C. (2003). Neural activity associated with binaural processes for the perceptual segregation of pitch. *Clinical Neurophysiology*, *114*, 2245–2250.
Johnson, B. W., Hautus, M. J., Duff, D. J., & Clapp, W. C. (2007). Sequential processing of interaural timing differences for sound source segregation and spatial localization: Evidence from event-related cortical potentials. *Psychophysiology*, *44*, 541–551.
Kiesel, A., Miller, J., Jolicœur, P., & Brisson, B. (2008). Measurement of ERP latency differences: A comparison of single-participant and jackknife-based scoring methods. *Psychophysiology*, *45*, 250–274.
Lipp, R., Kitterick, P., Summerfield, Q., Bailey, P. J., & Paul-Jordanov, I. (2010). Concurrent sound segregation based on inharmonicity and onset asynchrony. *Neuropsychologia*, *48*, 1417–1425.
McDonald, K. L., & Alain, C. (2005). Contribution of harmonicity and location to auditory object formation in free field: Evidence from event-related brain potentials. *Journal of the Acoustical Society of America*, *118*, 1593–1604.
Miller, J. O., Ulrich, R., & Schwarz, W. (2009). Why jackknifing yields good latency estimates. *Psychophysiology*, *46*, 300–312.
Sanders, L. D., Joh, A. S., Keen, R. E., & Freyman, R. L. (2008). One sound or two? Object-related negativity indexes echo perception. *Perceptual Psychophysiology*, *70*, 1558–1570.

Sanders, L. D., Zobel, B. H., Freyman, R. L., & Keen, R. (2008). Manipulations of listeners' echo perception are reflected in event-related potentials. *Journal of the Acoustical Society of America*, *129*, 301–309.

Snyder, J. S., & Alain, C. (2005). Age-related changes in neural activity associated with concurrent vowel segregation. *Cognitive Brain Research*, *24*, 492–499.

Snyder, J. S., & Alain, C. (2007). Toward a neurophysiological theory of auditory stream segregation. *Psychological Bulletin*, *133*, 780–799.

Weise, A., Schröger, E., & Bendixen, A. (2012). The processing of concurrent sounds based on inharmonicity and asynchronous onsets: An object-related negativity (ORN) study. *Brain Research*, *1439*, 73–81.

3.3. Study III: Theta oscillations accompanying concurrent auditory stream segregation

Tóth, B., Kocsis, Z., Urbán, G., & Winkler, I. (2016). Theta oscillations accompanying concurrent auditory stream segregation. *International Journal of Psychophysiology*, 106, 141-151. DOI: 10.1016/j.ijpsycho.2016.05.002.

International Journal of Psychophysiology 106 (2016) 141–151

Contents lists available at ScienceDirect

International Journal of Psychophysiology

journal homepage: www.elsevier.com/locate/ijpsycho

Theta oscillations accompanying concurrent auditory stream segregation

Brigitta Tóth [a,b,*], Zsuzsanna Kocsis [a,c], Gábor Urbán [a], István Winkler [a,d]

[a] Institute of Cognitive Neuroscience and Psychology, Research Centre for Natural Sciences, Hungarian Academy of Sciences, Budapest, Hungary
[b] Center for Computational Neuroscience and Neural Technology, Boston University, United States
[c] Department of Cognitive Science, Faculty of Natural Sciences, Budapest University of Technology and Economics, Budapest, Hungary
[d] Department of Cognitive and Neuropsychology, Institute of Psychology, University of Szeged, Szeged, Hungary

ARTICLE INFO

Article history:
Received 2 September 2015
Received in revised form 25 April 2016
Accepted 6 May 2016
Available online 8 May 2016

Keywords:
Concurrent sound segregation
Active and passive listening
Object-related oscillatory activity
Theta oscillation

ABSTRACT

The ability to isolate a single sound source among concurrent sources is crucial for veridical auditory perception. The present study investigated the event-related oscillations evoked by complex tones, which could be perceived as a single sound and tonal complexes with cues promoting the perception of two concurrent sounds by inharmonicity, onset asynchrony, and/or perceived source location difference of the components tones. In separate task conditions, participants performed a visual change detection task (visual control), watched a silent movie (passive listening) or reported for each tone whether they perceived one or two concurrent sounds (active listening). In two time windows, the amplitude of theta oscillation was modulated by the presence vs. absence of the cues: 60–350 ms/6–8 Hz (early) and 350–450 ms/4–8 Hz (late). The early response appeared both in the passive and the active listening conditions; it did not closely match the task performance; and it had a fronto-central scalp distribution. The late response was only elicited in the active listening condition; it closely matched the task performance; and it had a centro-parietal scalp distribution. The neural processes reflected by these responses are probably involved in the processing of concurrent sound segregation cues, in sound categorization, and response preparation and monitoring. The current results are compatible with the notion that theta oscillations mediate some of the processes involved in concurrent sound segregation.

1. Introduction

In natural/everyday situations, multiple sound sources (distal objects) are concurrently active in the environment. Therefore, typically sounds overlapping in both time and spectral contents arrive to our ears (proximal stimulation). Because the primary function of the auditory system is to provide information about the distal objects, it needs to parse the proximal input into signals generated by distinct causes (termed "auditory object", see Griffiths and Warren, 2004; Kubovy and Van Valkenburg, 2001; Winkler et al., 2009). This function has been termed auditory scene analysis (Bregman, 1990; for recent reviews, see Bidet-Caulet and Bertrand, 2009; Gutschalk and Dykstra, 2014; Schnupp et al., 2011; Shamma and Micheyl, 2010; Snyder and Alain, 2007). Bregman (1990) divided the processes of auditory stream segregation into two major categories: sequential (temporal) segregation denotes grouping/segregating sounds over time, whereas simultaneous (concurrent, spectral) segregation denotes grouping/segregating concurrent sound elements. The latter is largely based on the configuration of spectral elements present at the same time, such as grouping together the harmonics (i.e., integer multiples) of the same fundamental frequency. Considerable knowledge has already been compiled about the basic neural mechanisms of pitch perception (e.g., Bendor and Wang, 2005; for a recent review, see Wang and Walker, 2012) and the location of pitch-sensitive areas in the brain (e.g., Penagos et al., 2004; for a recent review, see Griffiths and Hall, 2012). Somewhat less information is available about concurrent sound segregation (Fishman et al., 2014) and the auditory cortical areas involved in it (Alain et al., 2005; Bidet-Caulet et al., 2007). However, much less is known about the neuronal oscillations produced by the large-scale brain network underlying this function. Because communication between the functionally linked areas is assumed to be mediated by slow oscillatory activity (Buzsáki and Draguhn, 2004; Klimesch et al., 2007), the current study was aimed at assessing the large scale brain oscillations associated with concurrent sound segregation.

In the most extensively studied stimulus configuration promoting concurrent sound segregation, one partial of a harmonic complex tone (HCT) is mistuned. Whereas the harmonics in tune with the fundamental are perceptually grouped together into single HCT that has the pitch of the fundamental, the mistuned partial is perceived as a separate pure tone alongside the HCT (Hartmann et al., 1990; Moore et al., 1986; for a mathematical model, see Cheveigne, 1997). Delaying a partial or delivering it from a different location than the rest also helps to segregate it from the rest of the HCT (Bidet-Caulet and Bertrand, 2009; Kocsis et al., 2014; Lipp et al., 2010; McDonald and Alain, 2005; Weise et al.,

* Corresponding author at: Research Centre for Natural Sciences, Hungarian Academy of Sciences, P.O. Box 1519, H-1519 Budapest, Hungary.
E-mail address: toth.brigitta@ttk.mta.hu (B. Tóth).

http://dx.doi.org/10.1016/j.ijpsycho.2016.05.002
0167-8760/

2012). Hearing a harmonic as segregated from the HCT is accompanied by an event-related brain potential (ERP) component termed the object related negativity (ORN; Alain et al., 2001, 2003; Alain and McDonald, 2007). ORN typically peaks between 150 and 180 ms from cue onset, appearing with a fronto-central maximum and with reversed polarity at electrodes placed over the mastoids. Using current source density analysis, Bendixen et al. (2010) found differences between the contributions of the left and right temporal cortices to the ORN response. The right-hemispheric temporal generator was related to spectral analysis (utilizing the superior spectral resolution of the right auditory cortex) while the left-hemispheric temporal activation was associated with auditory stream segregation. The auditory cortical origin of this component (Alain and McDonald, 2007; Alain et al., 2002) has been confirmed by fMRI. When listeners are instructed to tell whether they heard one or two sounds, ORN is followed by the P400 component peaking 300–450 ms from cue onset over centro-parietal areas (Alain et al., 2001, 2002). Evidence from ERP studies showing that ORN is elicited whether or not the sounds are attended (Alain, 2007; Alain and Izenberg, 2003) as well as the finding of behavioral and ERP signs of it in neonates (Bendixen et al., 2015), young infants (Folland et al., 2012), and non-human primates discriminating complex tones with tuned and mistuned partials(Fishman et al., 2014) suggest that concurrent sound segregation occurs automatically, although the task context may modulate some components of the processing network (Alain et al., 2002; Bidet-Caulet et al., 2007; Fishman et al., 2014).

However ERP analysis is not sufficient for characterizing the large-scale functional brain networks underlying sensory information processing functions, such as concurrent sound segregation (Makeig et al., 2004; Pfurtscheller and Lopes da Silva, 1999a,b; Tallon-Baudry and Bertrand, 1999). Time-frequency decomposition of brain activity may complement the information provided by ERPs about the sequence of operations as well as the localization information obtained by fMRI by revealing parallel processes occurring in distinct brain areas. Brain oscillations reflect rhythmic shifting of neuronal excitability over a wide range of spatial and temporal scales (for reviews, see Fell and Axmacher, 2011; Buzsáki and Draguhn, 2004). The brain oscillatory theory (Buzsáki and Draguhn, 2004; Klimesch et al., 2007) proposes that different frequency bands are associated with the neural activity of distinct cell assemblies and therefore, oscillations provide information about distinct cognitive and neuronal processing functions. Changes in the oscillation amplitudes probably reflect changes in the extent of neural activity involved in the associated function. So far, event related spectral response have been mainly utilized to study auditory change detection (Hsiao et al., 2009; Fuentemilla et al., 2008) and the processing of temporal object boundaries (McMullan et al., 2013). Only Bidet-Caulet et al.'s (2008) study has investigated concurrent sound segregation processes. These authors assessed local neuronal processing of amplitude-modulated sounds based on intracranial EEG recordings from the human primary auditory cortex. Thus, whereas the result of the above mentioned studies suggested that the amplitude of neuronal oscillations may represent neurophysiological markers associated with auditory perceptual processes, little is known about how these markers reflect the processing of the various auditory cues promoting concurrent sound segregation.

The present study was aimed at determining the neural oscillatory correlates of concurrent sound segregation by comparing large scale oscillatory activity in frontal, central, temporal and parietal brain regions of both hemispheres, separately 1) between concurrent sound segregation supported by different cues (mistuning a partial, delaying a partial, and delivering a partial with different perceived source location) and 2) between different attentional conditions (active: participants instructed to report after each stimulus whether they perceived one or two concurrent sounds; passive: participants performing a simple visual task while ignoring the sounds; and control: the same visual task is performed in the absence of sounds).

2. Methods

ERP responses extracted from a part of the electroencephalographic (EEG) signals analyzed for the current study have been reported by Kocsis et al. (2014) – see details of the data overlap with the current study in Section 2.4. and in Supplementary Table 1.

2.1. Participants

Twenty healthy volunteers (eight female, mean age 23.5 years, SD = 2.42) participated in the experiment. All participants had pure-tone thresholds within normal limits (<25 dB with <15 dB difference between the two ears) for the frequencies ranging from 250 to 8000 Hz and none of them were taking any medication affecting the central nervous system. Participants signed a written consent after the aims and procedures of the study were explained to them. They received modest financial compensation for participation. The study was approved by the Ethical Committee of the Institute of Cognitive Neuroscience and Psychology, Research Centre for Natural Sciences, HAS. Data of three participants were excluded from the final analysis because of low EEG signal to noise ratio.

2.2. Auditory stimuli

Sequences made up from two types of HCTs ("stimulus type": base vs. manipulated) were delivered to the participants. HCT duration was uniformly 250 ms (including 10 ms rise and 10 ms fall times) and the common intensity was set to 40 dB above hearing threshold, individually adjusted for each participant. Hearing thresholds were determined with the standard staircase method used in clinical practice, delivering the base version tones employed in the current study. The base tone was a HCT comprising the 5 lowest partials (all having the same amplitude and starting in sine phase). Four types of sound manipulations were employed: 1) mistuning the 2nd partial by +8% (mistuning cue), 2) delaying the 2nd partial by 100 ms (but ending at the same time as the other partials; delay cue), 3) delivering the 2nd partial with a different interaural time (ITD) and level difference (ILD) than the other partials (location-difference cue), and 4) combination of all three above cues for both the 2nd and the 4th partial (all cues combined). The location cue was implemented as follows: Half of the HCTs without the location-difference cue were presented with ITD and ILD parameters promoting the perception of a source 45° to the right, while the other half 45° to the left of the midline (ITD of ±200 μs and ILD of ±5 dB, applied congruently to all tonal components). The location-difference cue was then achieved by setting the opposite ITD/ILD combination for the manipulated partial(s) than for the rest of the partials, thus creating a ca. 90° perceived horizontal direction difference between the manipulated and the other partials. Similarly to the no location-difference HCTs, half of the HCTs with the location-difference cue had most partials originating from the left, the other half from the right.

Each different sound manipulation type was delivered separately in a single stimulus block of 280 HCTs (140 base version, 140 manipulated). Base and manipulated HCTs were delivered with a 1400 ms uniform onset-to-onset interval. Their order was randomized with the exception that the blocks commenced with 10 base-version HCTs, which were excluded from the analyses. In each block, all tones had the same fundamental frequency, which, however, changed from block to block. Eleven fundamental frequencies were used with the lowest frequency being 200 Hz, and the rest following in one-semitone steps (i.e., the highest fundamental frequency being 378 Hz). The order of the fundamental frequencies and the order of the different stimulus blocks were randomized independently of each other, separately for each participant. Further, the order of the HCTs with fully/predominantly left- vs. right-45° perceived origin was randomized

separately for each stimulus block, independent of the randomization of the two stimulus types (base vs. manipulated).

2.3. Task conditions

Three task conditions were employed: visual control task, passive listening, and active listening. The task conditions are illustrated on Fig. 1. A fixation cross (the "+" sign most of the time, see below) was continuously present at the center of the computer screen placed at 1.15 m directly in front of the participant (viewing angle 0.4°). Each active/passive listening trial started with the presentation of a HCT (250 ms), while the visual control task trials started with a silent interval of equal duration. Following a randomly varying delay (300–500 ms from the offset of the sound/silence), the fixation cross changed to the letter "X" for 100 ms, informing participants about the onset of the response period. The response period lasted till the onset of the next trial, which commenced 1400 ms from the onset of the sound/silence (compatibly with the timing of the sound sequences as described above above). Thus the length of the response period was between 550 and 750 ms depending on the duration of the delay between the offset of the sound/silence and the onset of the response period (750 ms response period for 300 ms delay, 550 ms response period for 500 ms delay). In the active listening condition, for each sound, participants were instructed to mark whether they heard one or two concurrent sounds by pressing one or the other response key during the response period. In the passive listening and the visual control condition, participants were instructed to press one of the two response keys as soon as they could after the fixation cross turned to "X" (ignoring the sound in the passive listening condition).

Auditory and visual stimuli were generated by an IBM PC computer using the Cogent 2000 stimulus presentation software (developed by the Cogent 2000 team at the FIL and the ICN and Cogent Graphics developed by John Romaya at the LON) under Matlab 2013a (The MathWorks Inc.). Sounds were delivered to participants binaurally via Sennheiser HD600 headphones (Sennheiser electronic GmbH & Co. KG).

Altogether, 6 different stimulus block's data were analyzed for the current study (see Table 1): four active listening task condition blocks with 1) the mistuning cue, 2) with the delay cue, 3) with the location-difference cue, and 4) with all cues combined; 5) one passive listening condition with all cues combined; and 6) one visual control task condition without sound presentation.

2.4. Procedure

The data reported here has been collected in a single experimental session, which also included several conditions reported only in Experiment 2 of Kocsis et al. (2014). The overlap between the conditions analyzed by Kocsis et al. and those analyzed for the current study are shown in Supplementary Table 1. (Note that even for the data reported in both studies, Kocsis et al. extracted ERPs, whereas here we conducted a time-frequency analysis extracting Event-Related Spectral Perturbations (ERSP) from the same EEG traces.) The experimental session is described here in full, noting the position of the stimulus blocks analyzed for the current study.

The experimental session started with two stimulus blocks during which participants watched a silent movie, followed by the passive listening condition stimulus block reported in the current study. These were followed by training for the active listening task. The first training block included 20 base-version and 20 manipulated HTCs (all cues combined for the 2nd and 4th partials). The second training block included 10 base version HTCs and 10 of each of the three single-cue 2nd-partial manipulated ones. The order of the different types of HTCs was separately randomized in the two training blocks. Participants were instructed to press one response key when perceiving one and the other key when perceiving two concurrent sounds. The response key assignment, which remained the same for the rest of the experiment, was counterbalanced across participants. The two training blocks were repeated if the percentage of responses corresponding to the percept promoted by the tone (base version: one sound; manipulated version: two concurrent sounds) was below 65%. No participant needed more than one repetition of the training blocks to clear the threshold. The training phase was followed by six active listening condition blocks (each with a different cue combination). After a longer break, first the visual control condition and finally five more active listening condition blocks (again with different cue combinations) were delivered. Because the order of the blocks with different cue combinations was randomized separately for each participant, the four cue combinations which are reported here (see Table 1) were delivered at random positions within the eleven active listening condition blocks delivered by Kocsis et al. (2014) (see Supplementary Table 1). The experimental session lasted for ca. 3 h. Additional breaks were inserted whenever requested by the participant.

2.5. Electrophysiological recording and data analysis

EEG was recorded with 63 Ag/AgCl electrodes placed on the scalp according to the extended international 10–20 system (Chatrian et al., 1985; Jasper, 1958) with Synamp amplifiers (Neuroscan Inc.). The reference electrode was attached to the tip of the nose, and the electrode AFz was used as the ground. Eye movements were monitored by bipolar recordings from two electrodes placed above and below the left eye (VEOG) and two electrodes placed lateral to the outer canthi of the two eyes (HEOG). The sampling rate was 2000 Hz and the EEG was online filtered with a 70 Hz low pass filter. The continuous EEG was band-pass filtered between 0.5 and 45 Hz (band-bass, Hamming windowed Fast Fourier Transform) by the EEGlab 11.0.3.1.b toolbox (Delorme et al., 2007) and ADJUST v3 plugin (Mognon and Buiatti, 2011) running under Matlab 2013a (Mathworks Inc.).

The EEG signals were segmented into epochs of 1400 ms duration. The epochs started from 387 ms before and lasted 1013 ms after the

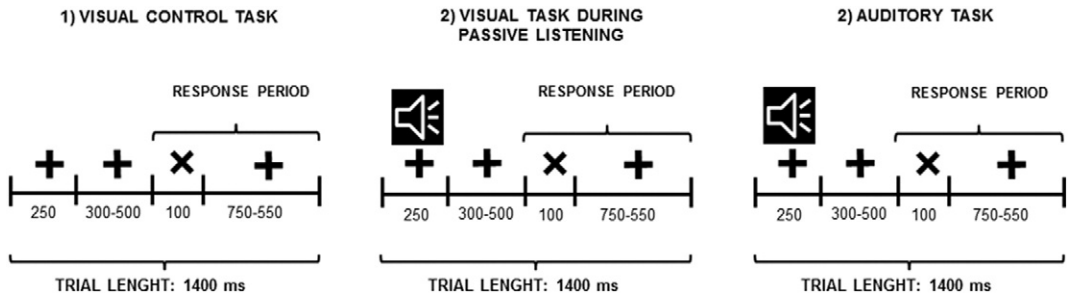

Fig. 1. Schematic representation of the three task conditions: 1) visual control, 2) passive listening, and 3) active listening. The fixation cross ('+') was continuously present, except when it changed to 'X' for 100 ms at the onset of the response period. In the visual control task and the passive listening condition, participants were instructed to respond to the change of the fixation cross. In the active listening condition, participants were instructed to mark during the response period whether they heard one or two sounds. Durations for the different periods of the trials are marked on the x-axis.

Table 1
Summary of the stimulus blocks.

Stimulus block	Task conditions	Type of sound manipulation	Mistuning	Delay	Location
1	Active listening	2nd partial mistuned	+		
2		2nd partial delayed		+	
3		2nd partial with location difference			+
4		2nd and 4th partials mistuned, delayed with location difference – all cues combined	+	+	+
5	Passive listening	2nd and 4th partials mistuned, delayed with location difference – all cues combined	+	+	+
6	Visual control	Fixation cross sign + change to x			

sound/silence onset (see Fig. 1). The pre-sound/silence period was used as the baseline for time frequency analysis. Artefact detection and removal were performed by manual rejection and the ICA function of the EEGLab toolbox (optimized to capture blinks, eye movements and generic discontinuities; for a detailed description of the standard procedure, see Delorme et al., 2007). As was noted in Section 2.1., data of three participants were excluded from the final analysis because of low EEG signal to noise ratio: for these participants, the number of epochs after artefact rejection did not reach the minimum set at 50, separately for each stimulus block. The average number of analyzed epochs per participant and stimulus block was 118.6 (SD: 26.2, 53–258 epochs).

2.5.1. Time frequency analysis

The mean event-related power spectrum changes [Event-related spectral perturbation (ERSP)] were separately calculated for each epoch and electrode by Wavelet Transform time-frequency decomposition (Delorme et al., 2007). Two hundred ERSP data points were calculated from −208 ms to 734 ms (relative to the sound/silence onset) for 84 equal log-spaced frequencies between 4 and 45 Hz by using 2 wavelet cycles at the lowest and 11.25 cycles at the highest frequency. The wavelet width was 1115 samples (equal to 557.5 ms) at the lowest frequency. The mean power in the pre-sound/silence period was used as the baseline. This method involves dividing the value at each time point in the epoch by the baseline, then taking the log10 transform of this quotient and multiplying it by 20, which yields values expressed in decibels (dB).

In order to investigate distributed network of brain oscillations eight regions of interest (ROIs) were defined for further analysis. Electrodes were grouped according to brain lobes and hemispheres with each electrode included in exactly one ROI: left frontal (the AF7, AF3, F7, F5, F3, and F1 electrodes), right frontal (AF8, AF4, F8, F6, F4, F2), left central (FC5, FC3, FC1, C5, C3, C1), right central (FC6, FC4, FC2, C6, C4, C2), left temporal (LM, T7, TP7, P7), right temporal (RM, T8, TP8, P8), left parietal (CP5, CP3, CP1, P5, P3, P1), and right parietal (CP6, CP4, CP2, P6, P4, P2). For the statistical analyses, ERSP power was averaged separately for each ROI. Based on visual inspection of the grand average time-frequency maps of the central ROIs, which showed the most characteristic effects, ERSP power was measured from two time-frequency windows of interest (TFWOI) for each ROI. The first TFWOI was set between 6 and 8 Hz and 60 and 350 ms, whereas the second TFWOI between 4 and 8 Hz and 350 and 450 ms. For the statistical analyses, the ERSP power values of the two TFWOIs were averaged separately for each participant, ROI, stimulus type (base vs. manipulated), sound manipulation type, and task condition.

2.5.2. Statistical analysis

The following statistical analyses were conducted separately for the two TFWOIs:

1) For testing the effects of the different cues on the ERSPs data from the mistuning, delay, and location-difference cues of the active listening task condition (obtained from the 1st, 2nd, and 3th stimulus blocks, as marked in Table 1) were compared by a four-way repeated measures ANOVA with the factors of Region (4 levels, as above) × Laterality (2 levels, as above) × Stimulus type (2 levels, as above) × Sound manipulation type (mistuning vs. delay vs. location-difference cue).
2) For testing the effects of the different tasks on the ERSPs, data from the visual and the passive listening condition and the "all cues combined" manipulation type of the active listening condition (collapsed over the base and manipulated stimulus types, separately for the two auditory conditions) were compared (obtained from the 4th, 5th, and 6th stimulus blocks as marked in Table 1) by a three-way repeated measures ANOVA with the factors of Region (frontal vs. central vs. parietal vs. temporal) × Laterality (left vs. right hemisphere) × Task condition (visual control vs. passive listening vs. active listening).
3) For testing the effects of attention on the ERSPs, data from the passive listening and the "all cues combined" manipulation type of the active listening task condition were compared (obtained from the 4th and the 5th stimulus blocks, as marked in Table 1) by a four-way repeated measures ANOVA with the factors of Region (4 levels, as above) × Laterality (2 levels) × Stimulus type (base vs. manipulated) × Attention (passive vs. active listening).

Greenhouse-Geisser correction for violations of the assumption of sphericity was applied where needed and the ε correction factors are reported. All significant results are reported together with the partial η^2 effect sizes. The *p*-values of post-hoc pair-wise comparisons were adjusted with Bonferoni's correction. All significant results are reported. Statistical analysis was performed with the Statistica version 11.0 (developed by StatSoft).

2.6. Behavioral data analysis

The percentage of responses corresponding to that promoted by the two stimulus types (base vs. manipulated) was separately determined for the two stimulus types and for the three different single-cue manipulation types. A two-way repeated-measures ANOVA was performed with the factors of Stimulus type (base vs. manipulated) × Sound manipulation types (mistuning vs. delay vs. location-difference cue) for assessing the perceptual effects of the different cues on concurrent sound segregation. A more detailed analysis of the behavioral effects of concurrent segregation cues has been conducted by Kocsis et al. (2014).

3. Results

3.1. Effects of different cues on perception

The mistuning and delay manipulations led to the perception of two concurrent sounds most of the time, whereas with the location-difference manipulation, participants indicated two sound objects with a much lower percentage; the base-version sounds were predominantly perceived as a single sound (Fig. 2). The ANOVA yielded significant main effect of Stimulus type ($F[1.16] = 81.37$, $p < 0.001$, $\varepsilon = 0.82$, $\eta^2 = 0.84$) and a significant interaction between Stimulus type and Sound manipulation type ($F[2.32] = 75.12$, $p < 0.001$, $\varepsilon = 0.61$, $\eta^2 = 0.82$). Pairwise comparisons showed that correspondence to the percept promoted by the sound was significantly lower with

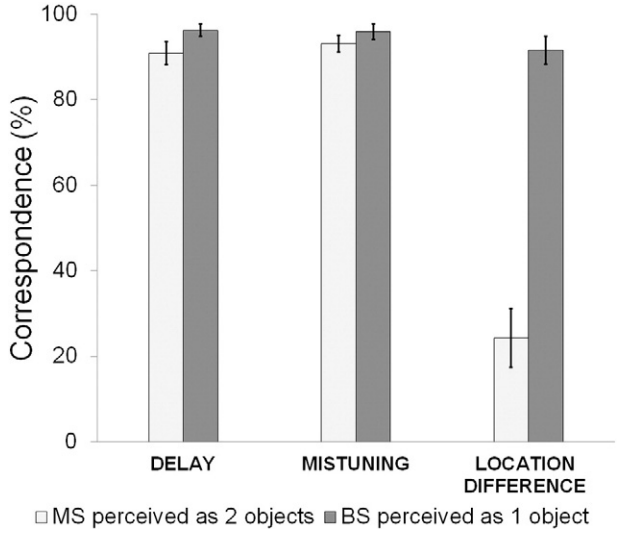

Fig. 2. The effects of different sound manipulations on the perceptual outcome measured as the percentage of correspondence between the perception promoted by a sound (1 sound for the base versions (BS) and 2 concurrent sounds for the manipulated versions (MS)) and actual perception.

the location-difference manipulation than with any other manipulation type ($p < 0.001$ in all cases).

3.2. Auditory vs. visual tasks effects on ERSP power in the theta band

A summary of the task effect data is shown in Fig. 3. The time-frequency plots of the ERSP power averaged for the two central ROIs show two time-frequency regions where the task had clear effects. These were selected as the two TFWOIs (Fig. 3A). In these regions, ERSP power appears to be higher in the conditions with auditory stimuli than for the visual task alone and higher for the active than for the passive listening conditions.

Accordingly, for the early TFWOI (60–350 ms, 6–8 Hz, theta band), a significant Task condition main effect ($F[2.32] = 20.77$, $p < 0.001$, $\eta^2 = 0.56$] was observed. The power of theta oscillations was significantly higher in the active listening compared with the visual control ($p < 0.001$) and the passive listening ($p < 0.002$) condition (Fig. 3A). There was also a significant main effect of Region ($F[3.48] = 11.39$, $p < 0.001$, $\varepsilon = 0.69$, $\eta^2 = 0.42$) with higher theta power in the central compared with the frontal ($p < 0.05$), parietal, and temporal ($p < 0.001$, both) regions. The significant Task condition × Region interaction ($F[6.96] = 11.03$, $p < 0.001$, $\varepsilon = 0.53$, $\eta^2 = 0.41$) was explained by the task-related differences being only significant for the frontal and central regions (passive listening compared with the visual control condition: $p < 0.001$ for both regions; and the active listening compared with the passive listening condition: $p < 0.001$ for both regions). That is, the task only significantly affected theta power in the frontal and central regions (Fig. 3B).

A significant main effect of Task condition ($F[2.32] = 28.37$, $p < 0.001$, $\varepsilon = 0.74$, $\eta^2 = 0.64$) was observed also for the late TFWOI (350–450 ms, 4–8 Hz, theta band). This was caused by the significantly higher theta power in the active listening compared with the visual control and the passive listening conditions ($p < 0.001$, both). There was also a significant main effect of Region ($F[3.48] = 4.89$, $p < 0.01$, $\varepsilon = 0.73$, $\eta^2 = 0.23$) with higher theta power in the central than in the frontal ($p < 0.01$) and the temporal regions ($p < 0.05$). The significant Task condition × Region interaction ($F[6.96] = 4.84$, $p < 0.001$, $\varepsilon = 0.63$, $\eta^2 = 0.23$) was caused by the theta power being higher in the frontal than in the temporal region ($p < 0.05$) and in the central than in the parietal region ($p < 0.001$), but only in the active listening condition. That is, only the active listening condition showed significant theta power scalp distribution effects in the frontal and central regions (Fig. 3C).

3.3. Attention effects on ERSP power in the theta band associated with concurrent sound segregation

A summary of the attention effect data is shown in Fig. 4. The time-frequency plots show clear effects of Attention and Stimulus type in the same two TFWOIs as were described in the previous section (see also Fig. 4A).

For the early TFWOI, the ANOVA yielded a significant main effect of Attention ($F[1.16] = 15.56$, $p < 0.002$, $\eta^2 = 0.49$), which was cause by higher theta power in the active than in the passive listening condition. There was also a significant main effect of Stimulus type ($F[1.16] = 4.90$, $p = 0.042$, $\eta^2 = 0.23$) indicating higher theta power for manipulated than for base sounds. The significant main effect of Region ($F[3.48] = 23.02$, $p < 0.001$, $\varepsilon = 0.47$, $\eta^2 = 0.59$] was caused by higher theta power in the frontal and central than in the temporal and parietal regions ($p < 0.001$ in all comparisons but for that between the frontal and the parietal regions: $p < 0.01$). There was also a significant interaction between Attention and Region ($F[3.48] = 2.90$, $p < 0.01$, $\varepsilon = 0.67$, $\eta^2 = 0.15$). Theta power was higher in the frontal compared with the temporal and the parietal regions, but only during the active listening condition ($p < 0.001$ and $p < 0.002$, respectively). That is, attention to the sounds increased the power of the frontal theta activity (Fig. 4B).

The ANOVA for the late TFWOI yielded a significant main effect of Attention ($F[1.16] = 41.27$, $p < 0.001$, $\eta^2 = 0.72$) showing that theta power was higher in the active than in the passive listening condition. Similarly to the effects found for the first TFWOI, manipulated sounds elicited significantly higher theta power than base sounds ($F[1.16] = 4.71$, $p < 0.05$, $\eta 2 = 0.23$). A significant main effect of Region was also observed ($F[3.48] = 6.55$, $p < 0.001$, $\varepsilon = 0.46$, $\eta^2 = 0.29$), which was caused by the theta power being higher in the central compared with the frontal and the temporal regions ($p < 0.05$ and $p < 0.001$, respectively). Also, a significant interaction was found between Attention and Region ($F(3.48) = 8.53$, $p < 0.001$, $\varepsilon = 0.78$, $\eta^2 = 0.35$). The source of the interaction was that during passive listening theta power was higher in the frontal than in the central and the parietal regions ($p < 0.01$, both), whereas during active listening, theta power in the frontal region tended to be lower than in the central region ($p = 0.054$) and theta power in the central region was higher than in the parietal and the temporal regions ($p < 0.001$, both). That is, with attention to the sounds, the peak of the theta power shifted from the frontal towards the central region (Fig. 4C).

3.4. Cue effect on ERSP power in the theta band associated with concurrent sound segregation during active listening

A summary of the cue effect data is shown in Fig. 5. The time-frequency plots show clear effects of manipulation (cue) type in the same two TFWOIs as were described earlier (see also Fig. 5A).

Theta power in the early TFWOI was significantly higher for manipulated than for base sounds (main effect of Stimulus type: $F[1.16] = 7.99$, $p < 0.05$, $\eta^2 = 0.33$). Sound manipulation type has significantly interacted with Stimulus type ($F[2.32] = 3.83$, $p < 0.05$, $\varepsilon = 0.96$, $\eta^2 = 0.19$), which was caused by the theta power being significantly larger for manipulated relative to base sounds only for the delay cue ($p < 0.005$). That is, tested separately only the delay condition induced a significant theta power difference between the base and the manipulated sounds (Fig. 5B). A significant Region main effect was found ($F[3.48] = 24.10$, $p < 0.001$, $\varepsilon = 0.71$, $\eta^2 = 0.6$). This was explained by the theta power being significantly higher in the frontal than in the parietal ($p < 0.01$) and the temporal ($p < 0.002$) regions and also in the central region compared with the frontal ($p < 0.01$), the parietal, and the temporal regions ($p < 0.001$, both). The ANOVA yielded a significant three-way interaction between Sound manipulation type, Stimulus type, and Region ($F[6.96] = 2.39$, $p < 0.05$, $\varepsilon = 0.68$, $\eta^2 = 0.13$). This was because only the sounds manipulated by the delay cue elicited significantly higher theta power than the corresponding base sounds,

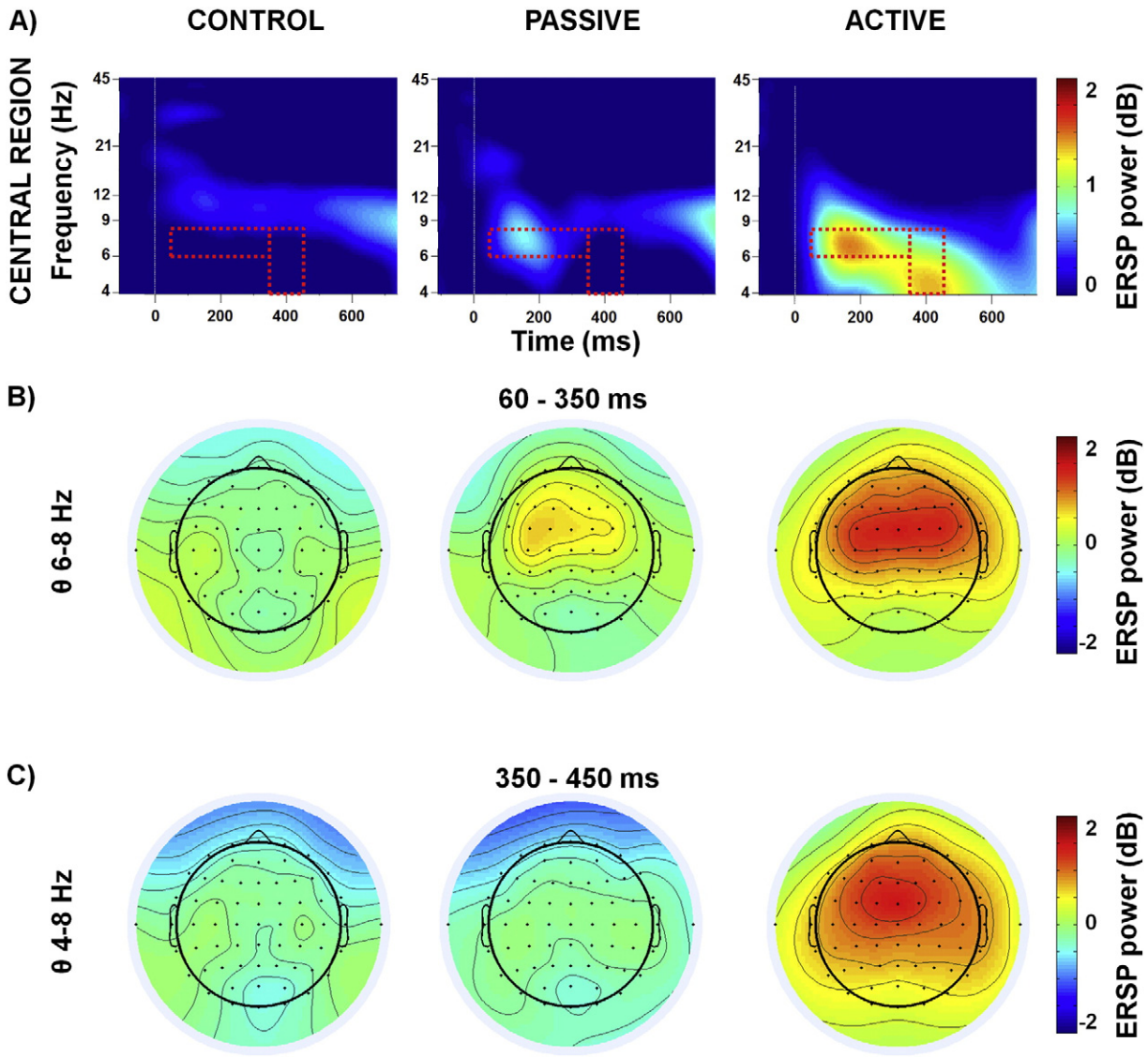

Fig. 3. Task effects on the ERSP (Event-Related Spectral Perturbation). A) Time-frequency plots of the ERSP power averaged across the two central ROIs for the visual control (left), the passive listening (middle), and the active listening condition (right) separately. The two TFWOI (Time-Frequency Windows of Interest): early: 60–350 ms/6–8 Hz; and late: 350–450 ms/4–8 Hz) are marked by rectangles shown with red dashed lines. Tone onset is marked by a white vertical line; B and C) ERPS power scalp distributions in the early (panel B) and late TFWOI (panel C), separately for the three task conditions (left to right as in panel A). ERSP power has been converted to dB for better resolution. Color scales (different for panel A and panels B/C) are shown to the left of each panel. Electrode locations are shown by grey dots on panels B and C.

and only in the frontal, central, and parietal regions ($p < 0.001$, in all cases). That is, the manipulated-base theta power difference was only significant for the delay cue and only in above three regions (Fig. 5B, right column).

For the late TFWOI, manipulated sounds also elicited significantly higher theta power than the base sounds (main effect of Stimulus type: $F[1.16] = 24.13$, $p < 0.001$, $\eta^2 = 0.6$). The main effect of Sound manipulation type ($F[2.32] = 7.39$, $p < 0.005$, $\eta^2 = 0.32$) was caused by the spatial-difference cue eliciting lower theta power than the delay and the mistuning cues ($p < 0.005$, both). The significant interaction between Stimulus type and Sound manipulation type ($F[2.32] = 6.11$, $p < 0.01$, $\eta^2 = 0.28$) was again due to only the delay cue inducing a significant theta power difference between the base and the manipulated tones ($p < 0.001$; Fig. 5C). There was also a significant main effect of Region ($F[3.48] = 8.86$, $p < 0.001$, $\varepsilon = 0.68$, $\eta^2 = 0.36$) with higher theta power in the central compared with the frontal ($p < 0.001$), the parietal ($p < 0.02$), and the temporal regions ($p < 0.001$), and in the parietal compared with the temporal region ($p < 0.02$). There was also a significant interaction between Sound manipulation type and Region ($F[6.96] = 2.35$, $p < 0.05$, $\varepsilon = 0.52$, $\eta^2 = 0.13$), which was due to the spatial-difference cue eliciting lower theta power responses compared with the delay and the mistuning cues in the frontal, central, and parietal regions ($p < 0.001$, all; see Fig. 5C), while in the temporal region, there was no significant difference between the theta power elicited by the mistuning and spatial-difference manipulation. The significant three-way interaction between Stimulus type, Sound manipulation type, and Region ($F[6.96] = 2.28$, $p < 0.05$, $\varepsilon = 0.82$, $\eta^2 = 0.12$), was caused by the theta power being higher for manipulated than for base tones for the delay cue in all regions ($p < 0.001$, all), and for the mistuning cue in the central ($p < 0.02$), the parietal ($p < 0.001$), and the temporal regions ($p = 0.01$). That is, in the above listed regions, the delay and the mistuning, but not the spatial-difference cue induced significant theta power differences between the base and the manipulated tones (Fig. 5C, right column).

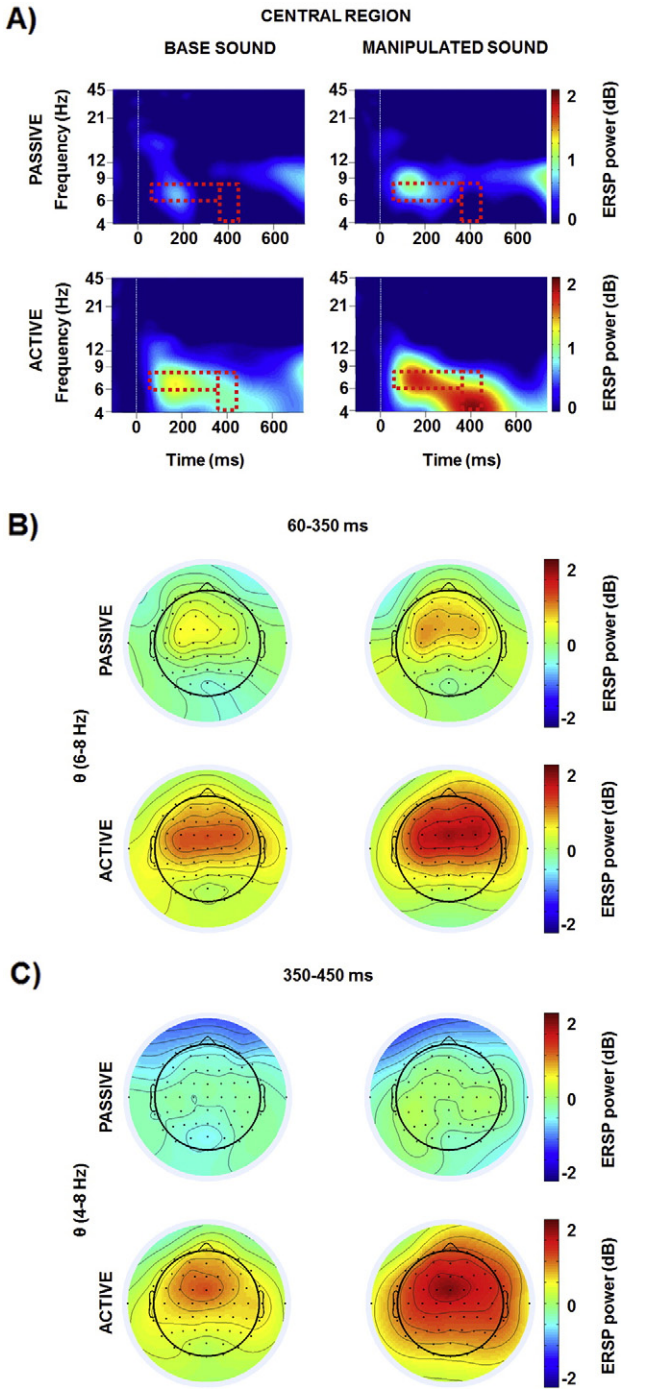

Fig. 4. Concurrent sound segregation related attention effects on the ERSP power in the theta band. A) Time-frequency plots of the ERSP power averaged across the two central ROIs for the base (left) and the manipulated sounds (right), separately for the passive (top) and the active listening conditions (bottom). The two TFWOIs (early: 60–350 ms/6–8 Hz; late: 350–450 ms/4–8 Hz) are marked by rectangles shown with red dashed lines. Tone onset is marked by a white vertical line; B and C) ERPS power scalp distributions in the early (panel B) and late TFWOI (panel C), separately for the passive (top) and the active listening conditions (bottom) and for the base (left) and the manipulated sounds (right). ERSP power has been converted to dB for better resolution. Color scales (different for panel A and panels B/C) are shown to the left of each row. Electrode locations are shown by grey dots on panels B and C.

4. Discussion

The present study, measured event related oscillations during concurrent sound segregation. Three different segregation cues and their combination were used to evoke the perception of two simultaneous sounds. Increased theta activity was found in two (early and late) post-stimulus intervals, which were more pronounced in the presence than in the absence of cues promoting the perception of two concurrent sounds. These ERSP responses were modulated by attention, which enhanced the responses differently in different scalp regions and more in the presence than in the absence of cues promoting the perception of two concurrent sounds. The different auditory cues were shown to differentially activate theta oscillations in different brain regions, and the sensitivity of the different brain regions to the presence of the different concurrent segregation promoting cues differed between the early and the late ERSP intervals: the early response was mostly modulated over fronto-central, whereas the late one over temporo-parietal areas. Thus the analysis of event-related theta oscillation revealed distributed brain networks involved in processing the various cues of concurrent sound segregation.

Comparing between the three task conditions (visual control, passive and active listening) revealed two TFWOIs (early: 60–350 ms/6–8 Hz; and late: 350–450 ms/4–8 Hz), in which the test sounds elicited higher event-related theta activity in the active listening condition than in the two other conditions. The time intervals of the observed post-stimulus theta oscillatory responses are largely overlapping with the two event-related brain potentials known to be elicited by active concurrent sound segregation: the ORN and the P400 response (Alain et al., 2001, 2002). The ORN is assumed to reflect the overall evaluation of the cues promoting the perception of two concurrent sounds (Kocsis et al., 2014), whereas the P400 is elicited only when the listener's task is to mark whether he/she heard one or two separate sounds (Alain et al., 2002).

In the early time window, the effect of sound manipulation on theta power was present both in the passive and the active listening condition. This result is similar to those found for ORN and it is also compatible with the notion that theta oscillations are involved in the neural mechanisms of the formation of perceptual objects. For example McMullan et al. (2013) have shown increased theta oscillation amplitudes at the post-stimulus latency of 100 ms during the automatic registration of spectral boundaries between successive objects. Another related result was obtained through the study of the continuity illusion (Warren et al., 1972; for a review see Warren, 1999). The term "continuity illusion" refers to the perceptual phenomenon when auditory information is restored for gaps occurring in interrupted sounds. Reduced theta-band activity was observed during trials in which the sound was perceived as continuing through an interrupting noise (that is, trials that elicited the continuity illusion) as compared to trials in which the gap was perceived (Riecke et al., 2009).

We found that attention to the sounds increased the frontal midline theta power in the early ERSP response. An similar increase of the ORN amplitude by attention was observed by Kocsis et al. (2014). However, since the auditory N1 amplitude (especially its frontal subcomponent) is also known to be enhanced by attention (Näätänen and Picton, 1987) and no significant interaction was observed between Attention, Region, and Stimulus type in the current study, the current attention effect can also be explained by neural processes involved in the generation of the auditory N1. Manipulation type affected the power of theta oscillation, the delay cue eliciting the largest response. A similar tendency for the ORN amplitude was reported by Kocsis et al. (2014). Because the behavioral results would suggest larger difference between the spatial difference and the other two cues, it is likely that the theta activity observed in the early time window represents a stage whose outcome is further processed before deciding about the overt response. Finally, also the ERSP scalp distribution was significantly affected by the different manipulations. The delay cue elicited stronger theta oscillations over frontal, central, and parietal areas than any of the other cues. However, Kocsis et al. (2014) found no significant ORN scalp distribution differences between the same three cues as were tested here. It is possible that the scalp-distribution differences were caused by neural networks activated during the time of the N1 component. Indeed in

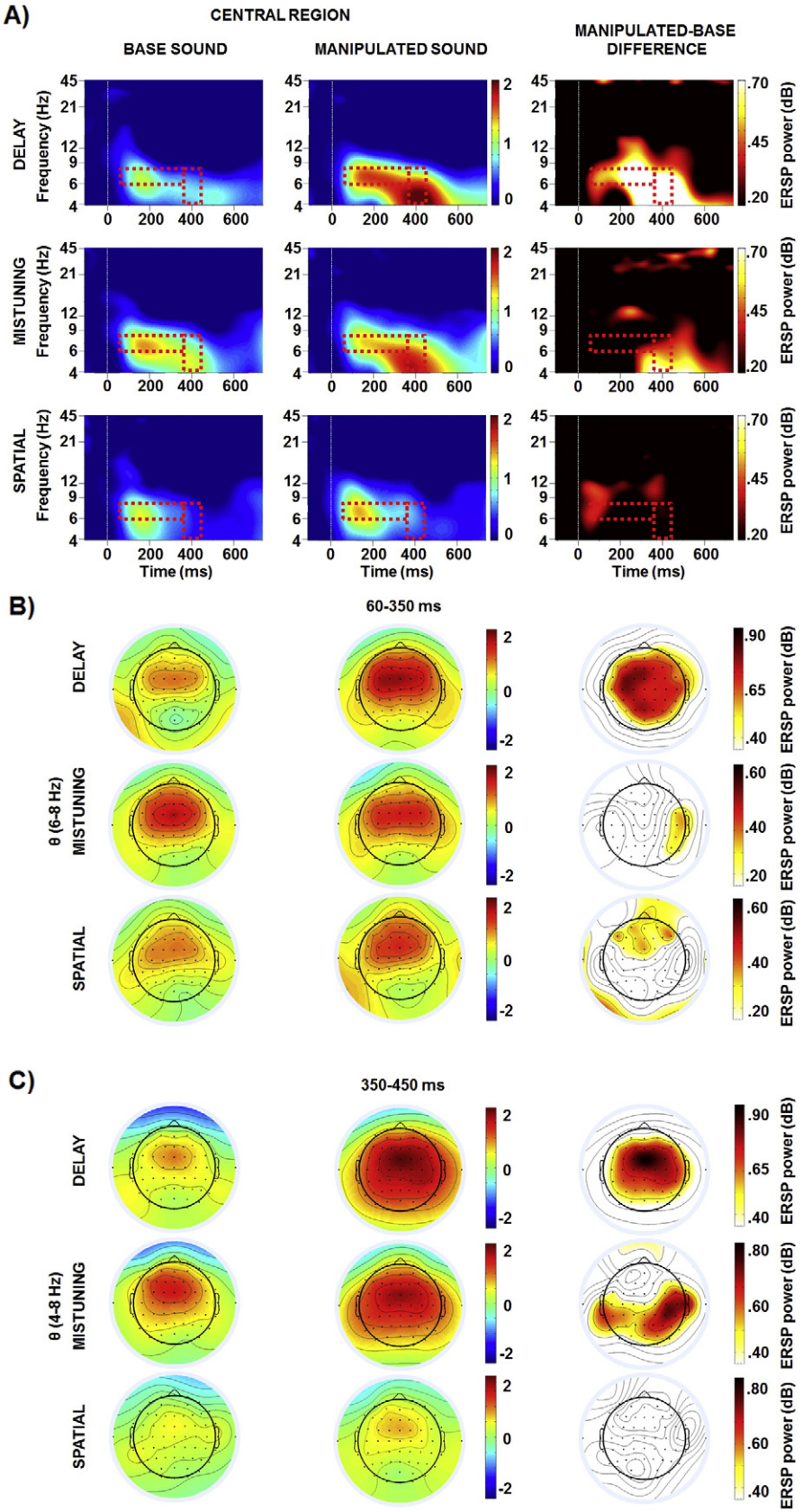

Fig. 5. Cue effects on concurrent sound segregation related ERSP power in the theta band. A) Time-frequency plots of the ERSP power averaged across the two central ROIs for the base (left) and manipulated sounds (middle) as well as their respective difference (right), separately for the the delay (top), the mistuning (middle), and the spatial difference manipulation (bottom). The two TFWOIs (early: 60–350 ms/6–8 Hz; and late: 350–450 ms/4–8 Hz) are marked by rectangles shown with red dashed lines. Tone onset is marked by a white vertical line; B and C) ERPS power scalp distributions in the early (panel B) and late TFWOI (panel C), separately for the three cues (top, middle, and bottom) and for the base (left) and manipulated sounds (middle) together with their respective difference (right). ERSP power has been converted to dB. Color scales (different for panel A and panels B/C) and for ERSP and difference maps are shown to the right of the time-frequency, the scalp distribution, and the difference plots. Electrode locations are shown by grey dots on panels B and C.

their Experiment 2, Kocsis et al. (2014) found an enhancement of the N1 amplitude for some of the manipulated sounds. After controlling for acoustic differences the authors concluded that the observed N1 enhancement was related to concurrent sound processing. If this activity at least partly underlies the current ERSP scalp-distribution differences, then it means that the ERSPs integrate neural activity involved in concurrent sound processing from multiple ERP components. However, the N1 enhancement elicited by manipulated sounds in Kocsis et al.'s (2014) study cannot fully explain the manipulation-type related ERSP scalp distribution differences observed in the current study because 1) N1 is not prominent over parietal scalp areas, one of the regions differentiating between the ERSPs elicited by the different cues and 2) Kocsis et al. (2014) found N1 enhancement for both delay- and location-difference-manipulated sounds, whereas the ERSP enhancement observed in the current study was restricted to the delay manipulation. Thus the current ERSPs do not simply mirror the ERP components extracted from the same EEG signal. Further, whereas no significant ERSP laterality effect emerged in the current study, Bendixen et al. (2010) observed changes in ORN laterality with different sequential probabilities of the base and manipulated sounds. However, the lack of lateralized effects is compatible with the symmetric scalp distribution observed by Bendixen et al. (2010) in their 50–50% (base and manipulated) stimulus block, the one that is compatible with the current stimulus paradigm.

Manipulated sounds also evoked enhanced theta activity relative to the base sounds within the 350–450 ms latency window. This effect was only present when participants were instructed to mark their perception and it was most robust over the centro-parietal cortices. This scalp distribution is compatible with those of both the P400 and P3. Similarly to the early ERSP response, the largest segregation-related response was again evoked by the delay cue. However, unlike the early ERSP response, the effect was significant over centro-parietal regions for both the mistuning and the delay cue, but not for the spatial difference cue. Thus these responses closely follow the perceptual data, suggesting that they reflect processes directly related to the overt response. Compatible results were obtained for the P400 ERP by Kocsis et al. (Kocsis et al., 2014).

The ERSP responses observed in the current study were affected similarly to the ERP responses obtained by Kocsis et al. (2014) by the presence vs. absence of the cues of concurrent sound segregation as well as attention. Therefore we propose that the ORN and P400 components reported by Kocsis et al. (2014) are at least partly driven by time-locked (phase locked) changes of theta oscillations. A growing set of evidence supports the view that sensory stimuli lead to phase-resetting of EEG rhythms in a non-stochastic manner and thus they underlie the averaged ERP responses (Basar-Eroglu et al., 1992; Makeig et al., 2004; Fuentemilla et al., 2006, Hanslmayr et al., 2007). Our ERSP analysis investigated event-related brain oscillations in several distinct frequency bands in parallel. The finding that the effects observed in ERPs were compatible with the effects found in theta-band oscillation with no other band showing effects of the variables tested suggests that the neural networks generating the ORN and also the P400 ERPs communicate via this slow rhythm, indicative of a widespread network spanning several distinct brain regions.

There are, however, also differences between the current ERSP and the previously observed ERP results. We have already noted that the scalp distribution of the ERSPs elicited in the early time window (compatible with N1 and ORN) is sensitive to the type of manipulation (cue) in a way that cannot be explained by either the ORN or a combination of the ORN and N1 effects derived from the same EEG data by Kocsis et al. (2014). Further, in contrast to the P400 ERP response, some theta activity is elicited by the base sounds in the active listening condition (see Fig. 4). This difference may be explained by ERSPs picking up some of the responses that are not time-locked to the onset of the triggering event (in addition to evoked/phase locked activity). These task-induced power changes occurring in a specific frequency band cannot be detected by the averaged ERPs. Therefore, ERSPs provide a more complete picture of the brain networks involved in concurrent sound segregation. In general, there is evidence to suggest that although they may appear to be functionally related, ERP and ERSP are sensitive to distinct aspects of the electrical neural activity and may not always reflect the same underlying cognitive processes (Makeig et al., 2004; Pfurtscheller and Lopes da Silva, 1999a,b; Tallon-Baudry and Bertrand, 1999). For example, ERP activity differences may occur at different scalp locations than the concurrent ERSP differences (Edwards et al., 2009) and averaged ERP and local field potential amplitude differences are more likely to result from phase resetting of neural oscillations than from corresponding changes in the underlying single-trial response amplitudes (Lakatos et al., 2008).

We observed two ERSPs of different properties have been observed to accompany concurrent sound segregation. A variety of previous auditory and visual studies have found two different ERSP responses in the theta band. Deviant stimuli elicit increased midline frontal theta activity relative to standards in the time interval of the mismatch negativity ERP component both in auditory and visual oddball paradigms (Kuo et al., 2011; Grau et al., 2007; Hsiao et al., 2009). Frontal theta amplitudes have also been found to increase during the manipulation period of both visual and auditory working memory tasks (Kawasaki et al., 2010). However in our stimulus paradigm, manipulated and base sounds occurred with same probability (50–50%). Therefore processes of change detection do not apply here, although it is possible that the theta rhythm recorded over the midline frontal cortex indexes early processes of updating the representations held in sensory/immediate memory. Further support for this notion comes from studies showing that negative feedback stimuli (Cohen et al., 2007, 2011; Marco-Pallares et al., 2008; Christie and Tata, 2009) and errors during cognitive tasks (Luu and Tucker, 2001; Yordanova et al., 2012; Cavanagh et al., 2009, 2010) evoke high- amplitude frontal theta oscillations. However, detecting two concurrent sounds is not analogous to sequential change detection. Alain et al. (2002) suggested that it involves detecting the difference between the auditory features extracted from the incoming stimulus (such as spectral makeup) and pre-existing templates for complex sounds. The simplest template is for a harmonic complex tone, which contains a fundamental frequency and one or more harmonics (integer multiples of the base frequency). However, such templates are not episodic, as in temporal deviance detection, where the expected sound is derived from the immediate history of the stimulus sequence and dynamically changes with the properties of the stimulus sequence (e.g., Sussman and Winkler, 2001; Winkler et al., 1996). Rather, these templates are to be understood as schemas helping the auditory system to find elements likely originating from the same source (cf. Ciocca, 2008). The most basic templates must be encoded in the structure of the auditory system, as mistuning a partial elicits an ORN-like response in neonates. The theta oscillation response shown here may be related to accessing this template.

High-amplitude theta oscillations have been observed during the later stages of information processing (approximately during the time interval of the P3 ERP component), for example for task-relevant deviant stimuli in the auditory oddball paradigm (Ishii et al., 2009). In a study aimed at differentiating between top-down and bottom-up attentional effects on event-related oscillations during auditory target detection (Li et al., 2010), target detection primarily requiring bottom-up attention elicited higher-amplitude theta oscillations in the 200–400 ms interval than top-down search. In contrast, target detection in the top-down search condition elicited increased-amplitude theta oscillations in the 350–650 ms interval with respect to pop-out target detection. The theta power effect observed in the current late window is probably closely related to these processes.

In summary, amplitude changes of two different oscillatory responses in the theta band appearing between 60 and 450 ms from

stimulus onset accompanied the presence of cues promoting concurrent sound segregation. Both of these theta-band responses were affected by whether listeners were required to report their perception of the test sounds as well as by the type of the cue promoting concurrent sound segregation. The early response appeared both in the passive and the active listening situation; it did not closely follow the behavioral data; and it had a fronto-central scalp distribution. The late response was only elicited in the active listening condition; it closely matched the behavioral data; and it had a centro-parietal scalp distribution. These results show similarities with the ORN and P400 ERP responses elicited by the same stimuli (Kocsis et al., 2014) suggesting that the early response probably reflects the evaluation of the auditory cues promoting concurrent sound segregation, whereas the late response is involved in performing the task. Theta oscillation generated by fronto-parieto-temporal brain networks have been implicated both in processes relying on memory representations, such as the assumed spectrotemporal templates underlying the processing of the cues of concurrent sound segregation as well as for decision, response preparation, and monitoring processes. Thus theta oscillations may mediate some of the processes involved in concurrent sound segregation.

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.ijpsycho.2016.05.002.

Disclosure statement

The authors declare no actual or potential conflict of interest.

Acknowledgments

This work was funded by the Hungarian Academy of Sciences (Magyar Tudományos Akadémia [MTA] post-doctoral fellowship and the internship of Erasmus Mundus Student Exchange Network in Auditory Cognitive Neuroscience to B.T. and the Lendület project LP2012-36/2012 to I.W.). The authors are grateful to prof. Dr. Alexandra Bendixen for the experimental design and comments on an earlier version of the manuscript, Orsolya Szalárdy for technical assistance and suggestions on the experimental design, Bálint Biczók for assistance in data preprocessing, and Zsuzsanna D'Albini for collecting the data.

References

Alain, C., 2007. Breaking the wave: effects of attention and learning on concurrent sound perception. Hear. Res. 229, 225–236.
Alain, C., Izenberg, A., 2003. Effects of attentional load on auditory scene analysis. J. Cogn. Neurosci. 15, 1063–1073.
Alain, C., McDonald, K.L., 2007. Age-related differences in neuromagnetic brain activity underlying concurrent sound perception. J. Neurosci. 27, 1308–1314.
Alain, C., Arnott, S.R., Hevenor, S., Graham, S., Grady, C.L., 2001. "What" and "where" in the human auditory system. Proc. Natl. Acad. Sci. 98, 12301–12306.
Alain, C., Schuler, B.M., McDonald, K.L., 2002. Neural activity associated with distinguishing concurrent auditory objects. J. Acoust. Soc. Am. 111, 990–995.
Alain, C., Theunissen, E.L., Chevalier, H., Batty, M., Taylor, M.J., 2003. Developmental changes in distinguishing concurrent auditory objects. Cogn. Brain Res. 16, 210–218.
Alain, C., Reinke, K., McDonald, K.L., Chau, W., Tam, F., Pacurar, A., Graham, S., 2005. Left thalamo-cortical network implicated in successful speech separation and identification. NeuroImage 26, 592–599.
Basar-Eroglu, C., Basar, E., Demiralp, T., Schürmann, M., 1992. P300-response: possible psychophysiological correlates in delta and theta frequency channels. Int. J. Psychophysiol. 13, 161–179.
Bendixen, A., Jones, S.J., Klump, G., Winkler, I., 2010. Probability dependence and functional separation of the object-related and mismatch negativity event-related potential components. NeuroImage 50, 285–290.
Bendixen, A., Háden, G.P., Németh, R., Farkas, D., Török, M., Winkler, I., 2015. Newborn infants can disentangle concurrent sounds. Dev. Neurosci. 37, 172–181.
Bendor, D., Wang, X., 2005. The neuronal representation of pitch in primate auditory cortex. Nature 436, 1161–1165.
Bidet-Caulet, A., Bertrand, O., 2009. Neurophysiological mechanisms involved in auditory perceptual organization. Front. Neurosci. 3, 182–191.
Bidet-Caulet, A., Fischer, C., Besle, J., Aguera, P.-E., Giard, M.-H., Bertrand, O., 2007. Effects of selective attention on the electrophysiological representation of concurrent sounds in the human auditory cortex. J. Neurosci. 27, 9252–9261.
Bidet-Caulet, A., Fischer, C., Bauchet, F., Aguera, P.-E., Bertrand, O., 2008. Neural substrate of concurrent sound perception: direct electrophysiological recordings from human auditory cortex. Front. Hum. Neurosci. 1, 5.
Bregman, A.S., 1990. Auditory Scene Analysis: The Perceptual Organization of Sound. The MIT Press, Cambridge, MA.
Buzsáki, G., Draguhn, A., 2004. Neuronal oscillations in cortical networks. Science 304, 1926–1929.
Cavanagh, J.F., Cohen, M.X., Allen, J.J., 2009. Prelude to and resolution of an error: EEG phase synchrony reveals cognitive control dynamics during action monitoring. J. Neurosci. 29, 98–105.
Cavanagh, J.F., Frank, M.J., Klein, T.J., Allen, J.J., 2010. Frontal theta links prediction errors to behavioral adaptation in reinforcement learning. NeuroImage 49, 3198–3209.
Chatrian, G.E., Lettich, E., Nelson, P.L., 1985. Ten percent electrode system fortopographic studies of spontaneous and evoked EEG activity. Am. J. EEG Tech. 25, 83–92.
Cheveigne, A.D., 1997. Harmonic fusion and pitch shifts of mistuned partials 102 pp. 1083–1087.
Christie, G.J., Tata, M.S., 2009. Right frontal cortex generates reward-related theta-band oscillatory activity. NeuroImage 48, 415–422.
Ciocca, V., 2008. The auditory organization of complex sounds. Front. Biosci. 13, 148–169.
Cohen, M.X., Elger, C.E., Ranganath, C., 2007. Reward expectation modulates feedback-related negativity and EEG spectra. NeuroImage 35, 968–978.
Cohen, M.X., Wilmes, K., Vijver, I.V., 2011. Cortical electrophysiological network dynamics of feedback learning. Trends Cogn. Sci. 15, 558–566.
Delorme, A., Sejnowski, T., Makeig, S., 2007. Enhanced detection of artifacts in EEG data using higher-order statistics and independent component analysis. NeuroImage 34, 1443–1449.
Edwards, E., Soltani, M., Kim, W., Dalal, S.S., Nagarajan, S.S., Berger, M.S., Knight, R.T., 2009. Comparisation of time-frequency responses and event-related potential to auditory speech stimuli in human vortex. J. Neurophysiol. 102, 377–386.
Fell, J., Axmacher, N., 2011. The role of phase synchronization in memory processes. Nat. Rev. Neurosci. 12, 105–118.
Fishman, Y.I., Steinschneider, M., Micheyl, C., 2014. Neural Representation of Concurrent Harmonic Sounds in Monkey Primary Auditory Cortex: implications for Models of Auditory Scene Analysis. J. Neurosci. 34, 12425–12443.
Folland, N., Butler, B.E., Smith, N., Trainor, L.J., 2012. Processing simultaneous auditory objects: infants' ability to detect mistuning in harmonic complexes. J. Acoust. Soc. Am. 131, 993–997.
Fuentemilla, L., Marco-Pallarés, J., Münte, T.F., Grau, C., 2008. Theta EEG oscillatory activity and auditory change detection. Brain Res. 1220, 93–101.
Fuentemilla, L., Marco-Pallarés, J., Grau, C., 2006. Modulation of spectral power and of phase resetting of EEG contributes differentially to the generation of auditory event-related potentials. Neuroimage 30, 909–916.
Grau, C., Fuentemilla, L., Marco-Pallarés, J., 2007. Functional neural dynamics underlying auditory event-related N1 and N1 suppression response. NeuroImage 36, 522–531.
Griffiths, T.D., Hall, D., 2012. mapping pitch representation in neural ensembles with fMRI. J. Neurosci. 32, 13343–13347.
Griffiths, T.D., Warren, J.D., 2004. What is an auditory object? Nat. Rev. Neurosci. 5, 887–892.
Gutschalk, A., Dykstra, A.R., 2014. Functional imaging of auditory scene analysis. Hear. Res. 307, 98–110.
Hanslmayr, S., Klimesch, W., Sauseng, P., Gruber, W., Doppelmayr, M., Freunberger, R., Pecherstorfer, T., Birbaumer, N., 2007. Alpha phase reset contributes to the generation of ERPs. Cereb. Cortex 17, 1–8.
Hartmann, W.M., McAdams, S., Smith, B.K., 1990. Hearing a mistuned harmonic in an otherwise periodic complex tone. J. Acoust. Soc. Am. 88, 1712–1724.
Hsiao, F.J., Wu, Z.A., Ho, L.T., Lin, Y.Y., 2009. Theta oscillation during auditory change detection: an MEG study. Biol. Psychol. 81, 58–66.
Ishii, R., Canuet, L., Herdman, A., Gunji, A., Iwase, M., Takahashi, H., Nakahachi, T., Hirata, M., Robinson, S.E., Pantev, C., Takeda, M., 2009. Cortical oscillatory power changes during auditory oddball task revealed by spatially filtered magnetoencephalography. Clin. Neurophysiol. 120, 497–504.
Jasper, H.H., 1958. The ten-twenty electrode system of the International Fed. Electroencephalogr. Clin. Neurophysiol. 10, 370–375.
Kawasaki, M., Kitajo, K., Yamaguchi, Y., 2010. Dynamic links between theta executive functions and alpha storage buffers in auditory and visual working memory. Eur. J. Neurosci. 31, 1683–1689.
Klimesch, W., Sauseng, P., Hanslmayr, S., Gruber, W., Freunberger, R., 2007. Event-related phase reorganization may explain evoked neural dynamics. Neurosci. Biobehav. Rev. 31, 1003–1016.
Kocsis, Z., Winkler, I., Szalárdy, O., Bendixen, A., 2014. Effects of multiple congruent cues on concurrent sound segregation during passive and active listening: an event-related potential (ERP) study. Biol. Psychol. 100, 20–33.
Kubovy, M., Van Valkenburg, D., 2001. Auditory and visual objects. Cognition 80, 97–126.
Kuo, B.-C., Yeh, Y.-Y., Chen, A.J.-W., D'Esposito, M., 2011. Functional connectivity during top-down modulation of visual short-term memory representations. Neuropsychologia 49, 1589–1596.
Lakatos, P., Gy, K., Mehta, A.D., Ulbert, I., Schrodeger, C.E., 2008. Entrainment of neuronal oscillations as a mechanism of attentional selection. Science 320, 110–113.
Li, L., Gratton, C., Yao, D., Knight, R.T., 2010. Role of frontal and parietal cortices in the control of bottom-up and top-down attention in humans. Brain Res. 1344, 173–184.

Lipp, R., Kitterick, P., Summerfield, Q., Bailey, P.J., Paul-Jordanov, I., 2010. Concurrent sound segregation based on inharmonicity and onset asynchrony. Neuropsychology 48, 1417–1425.
Luu, P., Tucker, D.M., 2001. Regulating action: alternating activation of midline frontal and motor cortical networks. Clin. Neurophysiol. 112, 1295–1306.
Makeig, S., Debener, S., Onton, J., Delorme, A., 2004. Mining event-related dynamics. Trends Cogn. Sci. 8, 204–210.
Marco-Pallares, J., Cucurell, D., Cunillera, T., García, R., Andrés-Pueyo, A., Münte, T.F., 2008. Human oscillatory activity associated to reward processing in a gambling task. Neuropsychology 46, 241–248.
McDonald, K.L., Alain, C., 2005. Contribution of harmonicity and location to auditory object formation in free field: evidence from event-related brain potentials. J. Acoust. Soc. Am. 118, 1593–1604.
McMullan, A.R., Hambrook, D., Tata, M.S., 2013. Brain dynamics encode the spectrotemporal boundaries of auditory objects. Hear. Res. 304, 77–90.
Mognon, A., Buiatti, M., 2011. ADJUST Tutorial An Automatic EEG artifact Detector based on the Joint Use of Spatial and Temporal features. Psychophysiology 48, 229–240.
Moore, B.C.J., Glasberg, B.R., Peters, R.W., 1986. Thresholds for hearing mistuned partials as separate tones in harmonic complexes. J. Acoust. Soc. Am. 80 (2), 479–483.
Näätänen, R., Picton, T.W., 1987. The N1 wave of the human electric and magnetic response to sound: a review and an analysis of the component structure. Psychophysiology 24 (375), 425.
Penagos, H., Melcher, J.R., Oxenham, A.J., 2004. A neural representation of pitch salience in nonprimary human auditory cortex revealed with functional magnetic resonance imaging. J. Neurosci. 24, 6810–6815.
Pfurtscheller, G., Lopes da Silva, F.H., 1999a. Event-related EEG/MEG synchronization and desynchronization: basic principles. Clin. Neurophysiol. 110, 1842–1857.
Pfurtscheller, G., Lopes da Silva, F.H., 1999b. Event-related EEG/MEG synchronization and desynchronization: basic principles. Clin. Neurophysiol. 110, 1842–1857.
Riecke, L., Esposito, F., Bonte, M., Formisano, E., 2009. Hearing illusory sounds in noise: the timing of sensory-perceptual transformations in auditory cortex. Neuron 64, 550–561.
Schnupp, J., Nelken, I., King, A.J., 2011. Auditory Neuroscience: Making Sense of Sound. MIT Press, Cambridge.
Shamma, S.a., Micheyl, C., 2010. Behind the scenes of auditory perception. Curr. Opin. Neurobiol. 20, 361–366.
Snyder, J.S., Alain, C., 2007. Toward a neurophysiological theory of auditory stream segregation. Psychol. Bull. 133, 780–799.
Sussman, E., Winkler, I., 2001. Dynamic sensory updating in the auditory system. Cogn. Brain Res. 12, 431–439.
Tallon-Baudry, C., Bertrand, O., 1999. Oscillatory gamma activity in humans and its role in object representation. Trends Cogn. Sci. 3, 151–162.
Wang, X., Walker, K.M.M., 2012. Neural mechanisms for the abstraction and use of pitch information in auditory cortex. J. Neurosci. 32, 13339–13342.
Warren, R.M., 1999. Auditory Perception: A New Analysis and Synthesis. Cambridge University Press, Cambridge.
Warren, R.M., Obusek, C.J., Ackroff, J.M., 1972. Auditory induction: perceptual synthesis of absent sounds. Percept. Psychophys. 20, 380–386.
Weise, A., Schröger, E., Bendixen, A., 2012. The processing of concurrent sounds based on inharmonicity and asynchronous onsets: an object-related negativity (ORN) study. Brain Res. 1439, 73–81.
Winkler, I., Karmos, G., Näätänen, R., 1996. Adaptive modeling of the unattended acoustic environment reflected in the mismatch negativity event related potential. Brain Res. 742, 239–252.
Winkler, I., Denham, S.L., Nelken, I., 2009. Modeling the auditory scene: predictive regularity representations and perceptual objects. Trends Cogn. Sci. 13, 532–540.
Yordanova, J., Kolev, V., Kirov, R., 2012. Brain oscillations and predictive processing. Front. Psychol. 3, 1–2.

3.4. *Study IV: Promoting the perception of two and three concurrent sound objects: an event-related potential study*

Kocsis, Z., Winkler, I., Bendixen, A., & Alain, C. (2016). Promoting the perception of two and three concurrent sound objects: an event-related potential study. *International Journal of Psychophysiology*, 107, 16-28. DOI: 10.1016/j.ijpsycho.2016.06.016.

International Journal of Psychophysiology 107 (2016) 16–28

Contents lists available at ScienceDirect

International Journal of Psychophysiology

journal homepage: www.elsevier.com/locate/ijpsycho

Promoting the perception of two and three concurrent sound objects: An event-related potential study

Zsuzsanna Kocsis [a,b,*], István Winkler [a,c], Alexandra Bendixen [d], Claude Alain [e,f]

[a] Institute of Psychology and Cognitive Neuroscience, Research Centre for Natural Sciences, Hungarian Academy of Sciences, Magyar tudósok körútja 2., Budapest, H-1117, Hungary
[b] Department of Cognitive Science, Faculty of Natural Sciences, Budapest University of Technology and Economics, Egry József u. 1., Budapest, H-1111, Hungary
[c] Institute of Psychology, University of Szeged, Egyetem u. 2., Szeged, H-6722, Hungary
[d] Cognitive Systems Lab, Institute of Physics, Technische Universität Chemnitz, Reichenhainer Str. 70, Chemnitz, D-09126, Germany
[e] Rotman Research Institute, Baycrest Centre, 3560 Bathurst Street, Toronto, Ontario M6A 2E1, Canada
[f] Department of Psychology, University of Toronto, 100 St. George Street, Toronto, Ontario M5S 3G3, Canada

ARTICLE INFO

Article history:
Received 17 February 2016
Received in revised form 24 June 2016
Accepted 29 June 2016
Available online 01 July 2016

Keywords:
Object-related negativity (ORN)
Concurrent sound segregation
Three simultaneous sound objects
Cue combinations

ABSTRACT

The auditory environment typically comprises several simultaneously active sound sources. In contrast to the perceptual segregation of two concurrent sounds, the perception of three simultaneous sound objects has not yet been studied systematically. We conducted two experiments in which participants were presented with complex sounds containing sound segregation cues (mistuning, onset asynchrony, differences in frequency or amplitude modulation or in sound location), which were set up to promote the perceptual organization of the tonal elements into one, two, or three concurrent sounds. In Experiment 1, listeners indicated whether they heard one, two, or three concurrent sounds. In Experiment 2, participants watched a silent subtitled movie while EEG was recorded to extract the object-related negativity (ORN) component of the event-related potential. Listeners predominantly reported hearing two sounds when the segregation promoting manipulations were applied to the same tonal element. When two different tonal elements received manipulations promoting them to be heard as separate auditory objects, participants reported hearing two and three concurrent sounds objects with equal probability. The ORN was elicited in most conditions; sounds that included the amplitude- or the frequency-modulation cue generated the smallest ORN amplitudes. Manipulating two different tonal elements yielded numerically and often significantly smaller ORNs than the sum of the ORNs elicited when the same cues were applied on a single tonal element. These results suggest that ORN reflects the presence of multiple concurrent sounds, but not their number. The ORN results are compatible with the horse-race principle of combining different cues of concurrent sound segregation.

1. Introduction

In everyday situations, we are often surrounded by sounds emanating from multiple sources. Although the acoustic energy from these sources sums into a complex acoustic wave, our auditory system is proficient in parsing this mixture into separate sound sources (i.e., auditory streams [Bregman, 1990] or perceptual objects [e.g., Kubovy and Van Valkenburg, 2001]). Early behavioral studies have identified several cues that contribute to the separation of concurrent sound sources. These include differences in frequency periodicity (Hartmann, 1985; Hartmann et al., 1986; Moore et al., 1986) and in location (Bronkhorst and Plomp, 1988) as well as onset asynchrony (Bregman and Pinker, 1978; Rasch, 1978). Other cues, such as amplitude and frequency modulation, have not been investigated to the same extent, and the results are somewhat inconsistent. For instance, some studies show that slow amplitude or frequency modulations are effective segregators for tones (McAdams, 1984a, 1984b; Dolležal et al., 2012). Along the same lines, other studies showed a lower likelihood for segregation when the tonal elements of a complex sounds are modulated at the same rate than when the rate differs across partials (Bregman et al., 1985; Bregman et al., 1990). In contrast, another study reported no benefit from having different rates of modulation in parsing two different vowels except when the harmonics of one vowel were modulated while harmonics of the other vowel remained stationary (Summerfield et al., 1992). Carlyon (1991) also showed that listeners could not reliably discriminate between coherent and incoherent frequency modulation of complex tones. Thus, further research is needed

* Corresponding author at: Institute of Cognitive Neuroscience and Psychology, Research Centre for Natural Sciences, Hungarian Academy of Sciences, Magyar tudósok körútja 2., Budapest, H-1117, Hungary.
E-mail addresses: kocsis.zsuzsanna@ttk.mta.hu (Z. Kocsis), winkler.istvan@ttk.mta.hu (I. Winkler), alexandra.bendixen@physik.tu-chemnitz.de (A. Bendixen), calain@research.baycrest.org (C. Alain).

http://dx.doi.org/10.1016/j.ijpsycho.2016.06.016
0167-8760/

to better understand the role of frequency and amplitude modulation in concurrent sound segregation.

Many studies have investigated the impact of individual cues on segregating concurrent sounds. In comparison, relatively few studies have considered how the information from different cues is integrated (McDonald and Alain, 2005; Kocsis et al., 2014; Weise et al., 2012). To date, the effects of multiple segregation cues on concurrent sound perception have been investigated by combining two or more *convergent* cues in a way to promote the perception of two concurrent sound objects. That is, either multiple cues were applied to the same tonal element (e.g., having the mistuned tonal element presented at a different location or with a temporal delay) or identically to two different tonal elements, thus promoting the two elements to be grouped together and, again promoting the perception of two concurrent sounds (the complex tone resulting from grouping the two manipulated tonal elements and the complex tone resulting from grouping the unmanipulated tonal elements). An unresolved issue is whether the presence of multiple *divergent* cues, that is, cues promoting different groupings of tonal elements (e.g., mistuning one partial while presenting another partial with temporal delay) could lead to the perception of three (or more) concurrent sound objects. In the present study, divergent cues were operationalized as manipulations of two different tonal elements each promoting the segregation of the target element both from the rest of the harmonic complex and from the other manipulated element (see Fig. 1). Cues acting on the same tonal element were always convergent, whereas cues acting on different tonal elements were always divergent with respect to each other.

The present study aims to investigate whether manipulations of two different tonal elements of a harmonic complex tone could promote the perception of three concurrent sound objects and elicit separate object-related negativity components (ORN; Alain et al., 2001) of the event-related brain potential (ERP). The ORN peaks between 150 and 180 ms from cue onset with maximal amplitude at frontal and frontocentral electrodes. With nose reference, it inverts polarity at the mastoids (Alain et al., 2002), consistent with generators located in the superior temporal gyrus near Heschl's gyrus (Alain et al., 2001; Arnott et al., 2011). ORN has been shown to be larger at the mastoid electrodes during active listening (when listeners were required to judge whether they heard one or two concurrent sounds) than during passive listening (listeners had no task related to the sounds), which indicates attentional modulation of the ORN amplitude (Alain et al., 2001). The ORN can be elicited by many different cues inducing concurrent sound perception including inharmonicity (Alain et al., 2001, 2002; Bendixen et al., 2010), onset asynchrony (Lipp et al., 2010; Weise et al., 2012), dichotic pitch (Johnson et al., 2003; Hautus et al., 2009), differences in the fundamental frequency (Δf_0) of speech sounds (Snyder and Alain, 2005; Alain et al., 2005), and simulated echo (Sanders et al., 2008a, 2008b). There are also reports of ORN being elicited by a combination of some of the above cues, such as inharmonicity and location difference (McDonald and Alain, 2005) or inharmonicity and onset asynchrony (Weise et al., 2012; Kocsis et al., 2014).

Du et al. (2011) showed that the combined effect of location and Δf_0 on the amplitude of the magnetic equivalent of the ORN response closely matched the sum of the ORN responses elicited by the single cues (i.e., location or Δf_0 alone). However, Kocsis et al. (2014) found sub-additive effects of combining inharmonicity, onset asynchrony, and source location difference. Kocsis et al. (2014) used either one of the three single-cue manipulations or combined two or all three cues for segregation. The manipulations affected either one or two tonal elements in a congruent manner (i.e., cues promoting the two tonal elements to be grouped into a single sound object by e.g., same percentage of mistuning or same temporal delay applied to two different tonal elements). In different blocks of trials, participants either watched a subtitled, muted movie (no response required), or were asked to focus on the stimuli and to press a button indicating whether they heard one or two concurrent sound objects. Participants performed generally well (above 87%) in identifying two objects in most conditions. The main finding was that cue combinations always elicited numerically smaller ORN amplitudes than the sum of the ORN amplitudes separately elicited by the comprising cues. That is, the ORN amplitude showed subadditivity to various combinations of different cues promoting the perception of two concurrent sound objects. This suggests that ORN reflects the overall read-out of the auditory system's assessment of the presence of two objects as opposed to indexing the processing of the different cues.

In the present study, we compared the effects of convergent and divergent cues on perceptual and neural (ORN) indicators of concurrent sound segregation. In Experiment 1, we investigated the synergic effect of various cues (i.e., harmonicity, temporal delay, AM, and FM) in conjunction with a location cue (applied to the same tonal element) on the perception of two concurrent sound objects. We also tested whether applying divergent segregation cues on two different tonal elements would promote the perception of three auditory objects. We used four different cues in conjunction with location difference to assess their potential strength and tested how perception of two vs. three sound objects occurs with different cue combinations. In Experiment 2, we recorded the electroencephalogram (EEG) in a passive listening condition to test whether the same stimuli elicit significant ORN responses and to examine whether the ORN amplitudes show additivity (or super/sub-additivity) for cue combinations. This allows us to distinguish two functional interpretations of the ORN component: If ORN elicited by the divergent manipulations (three-objects conditions) is as large as the summed amplitudes of the ORNs elicited separately by the

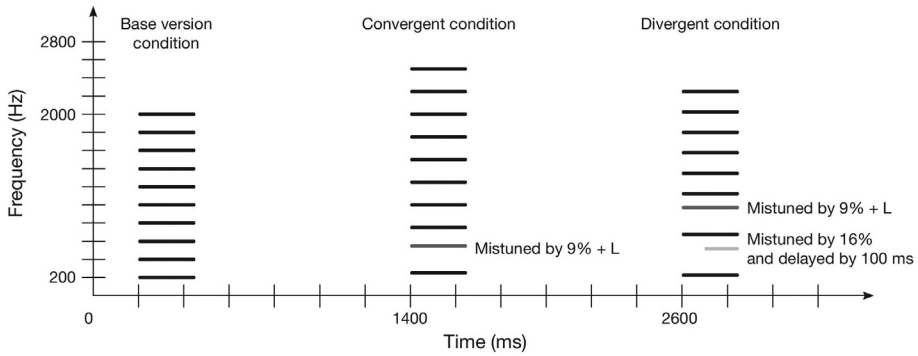

Fig. 1. Schematic depiction of three consecutive trials illustrating stimuli in the base version, convergent, and divergent conditions, which are set up to promote the perception of one, two, and three concurrent objects, respectively. The x axis depicts time, while the y axis depicts frequency. The black horizontal lines depict the frequency components of the harmonic complex. The lighter grey lines depict tonal elements that were manipulated (manipulation type marked on the figure) but presented at the same location as the remaining harmonics. The darker grey lines depict tonal elements that were manipulated and presented from a different location ("+ L" marked by the manipulation type). Note that the *f*0 varied from stimulus to stimulus (200–378 Hz). Therefore, the 10 pure tones of individual stimuli covered different frequency ranges. The frequency range was equalized across conditions. Stimulus timing is compatible with Experiment 2. In Experiment 1, the next stimulus was delivered 300 ms after the listener's response.

comprising cues, this would support the interpretation that ORN reflects the independent evaluation of the various cues during concurrent sound segregation. Alternatively, if ORN generally shows sub-additivity relative to the ORN amplitudes elicited separately by the individual cues, then this would be consistent with the interpretation that ORN reflects the auditory system's overall readout of the presence of multiple auditory objects, regardless of the congruency of the manipulations or the number of objects and different cues.

2. Experiment 1

Experiment 1 investigated whether the combination of different acoustic cues can yield the perception of three sound objects when they are applied as divergent manipulations across two different tonal elements.

2.1. Methods

2.1.1. Participants

Twelve healthy volunteers (six female, mean age 23.16 years, SD = 5.24) participated in the experiment; all of them were right-handed. All participants had pure-tone thresholds within normal limits (<25 dB HL with <15 dB HL difference between the two ears) for the frequencies ranging from 250 to 8000 Hz and none reported taking any medication affecting the central nervous system. Participants were recruited from the Rotman Research Institute participant database, and received modest monetary compensation for their participation. Prior to the beginning of the experiment, written informed consent was obtained from each participant after the experimental procedures and aims of the study were explained to them. The experiment was approved by the Research Ethics Board of the Rotman Research Institute (Toronto, Canada).

2.1.2. Apparatus, stimuli and procedure

Stimuli consisted of complex sounds (250 ms in duration, 10 ms rise and fall times) generated by summing ten pure tones of equal intensity. The sounds were generated digitally at a sampling rate of 48 kHz with MATLAB 2009a. The f_0 of the harmonic complex tones varied randomly within the range of 200 Hz to 378 Hz in one-semitone steps. The f_0 variation was employed to discourage participants from basing their response on a comparison between the incoming sound and the previous one. The complex tones were divided into three conditions (base version, convergent, and divergent) according to the percept they were set up to promote (i.e., one, two, or three objects). Each condition was presented with equal probability (280 stimuli, each). For the base-version condition, all tonal elements were an exact integer multiple of f_0, started at the same time, without amplitude or frequency modulation, and were presented from the same source location. The stimuli in the convergent and divergent conditions were delivered by ten different stimulus types, each (28 trials per stimulus type). Hence, participants were presented with a grand total of 21 stimulus types, which were inter-mixed within each block of trials.

For creating stimuli in the convergent condition, either the 2nd or the 4th partial was manipulated, whereas for the stimuli in the divergent condition, both the 2nd and the 4th partials were manipulated simultaneously in a divergent manner. Five different manipulations and their combinations were used: a) mistuning the partial by either +9 or +16%, b) delaying the onset of the partial by 100 ms (but ending at the same time as the other partials), c) amplitude-modulating (AM) the partial by 5 Hz modulation rate and 50% depth, d) frequency-modulating (FM) the partial by 5 Hz, e) delivering the partial from a different location compared to that of the other harmonics. We chose mistuning and delay because these cues are well-known to promote concurrent sound segregation. We included AM and FM because these cues have not yet been studied extensively behaviorally or with the use of EEG. Location difference has been shown to be a fairly weak cue when standing alone, but to significantly enhance the perception of concurrent sounds when applied together with some other cue (e.g., McDonald and Alain, 2005). Therefore, here the location cue always supplemented another cue in a convergent manner; it was never used alone. For the purpose of adding the location manipulation, the location of each individual complex tone could take one of two possible positions fully crossed with the other conditions. Hence in each condition, half of the tones were presented from the right and the other half from the left speaker; when the location cue was employed, the manipulated partial was presented by the opposite speaker. A summary of the experimental manipulations is given in Table 1, and a visual depiction of trials and stimuli is shown in Fig. 1.

The sounds were presented via speakers (GSI 61, Clinical Audiometer) in a sound-attenuated chamber at an intensity of 45 dB sensation level with respect to the participant's average pure-tone threshold. The speakers were positioned symmetrically to the left and right at 45° from the midline, 200 cm straight from the centre of the room where the participant was sitting. The sounds were delivered in a fully randomized order in six stimulus blocks, each containing 140 stimuli.

The participants sat comfortably and were asked not to move their heads during the experiment. They were instructed to look straight ahead at a fixation cross displayed on a computer screen that was placed approximately 140 cm in front of them. On each trial, one tone complex was presented and participants indicated whether they heard one, two, or three concurrent sounds. They made their response by pressing one of three predefined keys on a standard keyboard using their dominant hand. There was no time constraint to the response, but listeners were instructed to respond as quickly as possible. The next stimulus commenced 300 ms after the button was released. The experiment took about 15–20 min.

We tested whether a) the convergent manipulations resulted in the perception of two concurrent sounds and b) the divergent manipulations resulted in the perception of three concurrent sounds by comparing, with paired-samples *t*-tests, the proportions of reporting the perception of two and three concurrent sound objects, respectively, between the convergent and divergent conditions. We also tested whether the proportion of reporting two concurrent sound objects in the convergent conditions varied as a function of the manipulated cue using a repeated-measures ANOVA with the *Harmonic number* (2 levels: 2nd partial vs. 4th partial) and *Cue* (5 levels: mistuned by 9%, mistuned by 16%, delay, AM, FM) as within-subject factors.

All statistically significant results are reported. ANOVA effects are reported together with the partial η^2 effect size measure. Greenhouse–Geisser correction was applied when the assumption of sphericity was violated; the ε correction factor is reported in these cases. Post-hoc tests for repeated-measures ANOVAs were carried out with the Bonferroni correction of the confidence level for multiple comparisons.

2.2. Results and discussion

Fig. 2 shows the group-average percentages of the responses (i.e., one, two or three concurrent sound objects), separately for each stimulus type (upper panel) as well as the two- and three-object responses collapsed separately for the convergent and the divergent condition (lower panel). As expected, participants often reported hearing one sound object (87.44%) when all partials were in tune, unmodulated, delivered from the same location, and began at the same time (base-version condition). In the convergent condition, participants reported hearing two concurrent sounds on average on 68.21% of the trials, while in the divergent condition they reported hearing 46.01% of the trials as three concurrent sounds. The paired-samples *t*-test for the proportion of reporting two concurrent sound objects showed a significant difference between the convergent and the divergent conditions: $t(11) = 3.835$, $p < 0.005$, due to a higher proportion of reporting two objects in the convergent than in the divergent condition. The proportion of reporting three concurrent sound objects also showed a significant difference between the convergent and divergent condition:

Table 1
Summary of experimental manipulations.

Condition & stimulus type	Mistuning by 9%	Mistuning by 16%	Delay of 100 ms	Frequency modulation	Amplitude modulation	Location difference
Base-version condition						
1	–	–	–	–	–	–
Convergent condition						
2	2nd partial	–	–	–	–	2nd partial
3	–	2nd partial	–	–	–	2nd partial
4	–	–	2nd partial	–	–	2nd partial
5	–	–	–	2nd partial	–	2nd partial
6	–	–	–	–	2nd partial	2nd partial
7	4th partial	–	–	–	–	4th partial
8	–	4th partial	–	–	–	4th partial
9	–	–	4th partial	–	–	4th partial
10	–	–	–	4th partial	–	4th partial
11	–	–	–	–	4th partial	4th partial
Divergent condition						
12	2nd partial	4th partial	–	–	–	2nd partial
13	4th partial	2nd partial	–	–	–	4th partial
14	–	4th partial	–	2nd partial	–	2nd partial
15	–	2nd partial	–	4th partial	–	4th partial
16	–	4th partial	–	–	2nd partial	2nd partial
17	–	2nd partial	–	–	4th partial	4th partial
18	4th partial	2nd partial	2nd partial	–	–	4th partial
19	2nd partial	4th partial	4th partial	–	–	2nd partial
20	4th partial	2nd partial	4th partial	–	–	4th partial
21	2nd partial	4th partial	2nd partial	–	–	2nd partial

$t(11) = -9.123$, $p < 0.001$, due to a higher proportion of reporting three objects in the divergent than in the convergent condition.

However, in the divergent condition, listeners reported hearing two objects with approximately the same probability (50.33%) as hearing three objects (46.01%). This suggests that they could not reliably separate two and three objects when cues promoted the presence of three concurrent objects. Therefore, we pooled the two- and three-objects responses into a joint response category that reflects the proportion of hearing multiple sound objects. This way, participants reported hearing multiple sound objects on average in 82.11% in the convergent condition and in 96.34% in the divergent condition.

The repeated-measures ANOVA yielded a significant interaction between *Harmonic number* and *Cue* ($F(4,44) = 3.95$, $p < 0.008$, $\eta^2 = 0.264$, $\varepsilon = 0.871$). To decompose the interaction, separate paired-sample *t*-tests between the 2nd and 4th harmonics were conducted for each cue. These tests revealed a significant difference between the harmonic numbers only for the mistuning by 9% ($t(11) = 3.335$, $p = 0.007$; all other p values >0.05), with the 9% mistuning cue being identified as more than one object less often with the 4th than the 2nd harmonic (Fig. 3). These results explain the significant main effect of *Harmonic number* ($F(1,11) = 21.779$, $p < 0.001$, $\eta 2 = 0.664$) and the main effect of *Cue* ($F(4,44) = 4.299$, $p < 0.01$, $\eta 2 = 0.281$, $\varepsilon = 0.9$).

Our finding that the proportion of reporting hearing multiple auditory objects was higher when the 2nd than when the 4th harmonic was mistuned by 9% is consistent with previous reports showing greater likelihood of reporting hearing two sound objects for lower than for higher harmonics (Alain et al., 2001). Furthermore, for all stimulus types in the convergent condition, except for the 4th harmonic mistuned by 9%, the perception of multiple concurrent sounds was evoked on the majority of the trials. Hence the results of Experiment 1 also show that, at least with the help of the location difference cue, both the AM and the FM cues evoked concurrent sound segregation.

Although the sounds of the divergent condition yielded a higher proportion of hearing three sound objects than those of the convergent condition, the proportion of reporting hearing three sound objects remained relatively low. This may reflect some limitation of the ability to process more than two sound objects at a time. Evidence from a behavioral and some ERP studies suggest that for qualitatively similar sound streams (Cusack et al., 2004), only one object is brought into the foreground, while the others remain "undiscriminated" in the background (Brochard et al., 1999; Sussman et al., 2005; Leung et al., 2011, 2015; Kulagina et al., 2015). This is consistent with the current findings showing that on average, on >90% of the divergent-condition trials, participants did perceive more than one sound object, but they only reported hearing three concurrent sounds in ca. half of these trials while categorizing the other half as consisting of two concurrent sounds. This suggests that the divergent manipulations may have been mostly identified as multiple objects.

The results from Experiment 1 suggest that participants might be far from perfect in distinguishing between convergent and divergent cues. This would account for the relatively low proportion of trials on which they reported hearing three concurrent sound objects. However, from behavioral data alone, it is difficult to determine to what extent convergent and divergent cues are processed and integrated during concurrent sound segregation. The low probability of reporting three concurrent sound objects could reflect a limitation in early sensory encoding, with the presence of multiple cues interfering with one another. For instance, one could imagine that the presence of divergent cues creates conflict, making it more difficult to reach figure-ground segregation.

In Experiment 2, we measured auditory ERPs during passive listening conditions in order to investigate the encoding of convergent and divergent cues without potential task-related biases. Based on the results of Experiment 1, we expect that an ORN component will be elicited in Experiment 2 in those conditions where participants reported hearing multiple auditory objects — that is, in all conditions except the one in which the 4th partial was mistuned by 9%. ORN has been shown to reflect the presence of two concurrent sound objects. Assuming that ORN is also elicited by more than two concurrent sound objects, we expect that ORN will be elicited by those manipulated sounds for which participants reported hearing more than one sound object. Differences in ORN amplitude will possibly provide insight into the mechanisms supporting the processing of convergent and divergent cues.

3. Experiment 2

3.1. Methods

3.1.1. Participants

Eighteen healthy volunteers, none of whom had taken part in Experiment 1, participated in the second experiment. The data from two

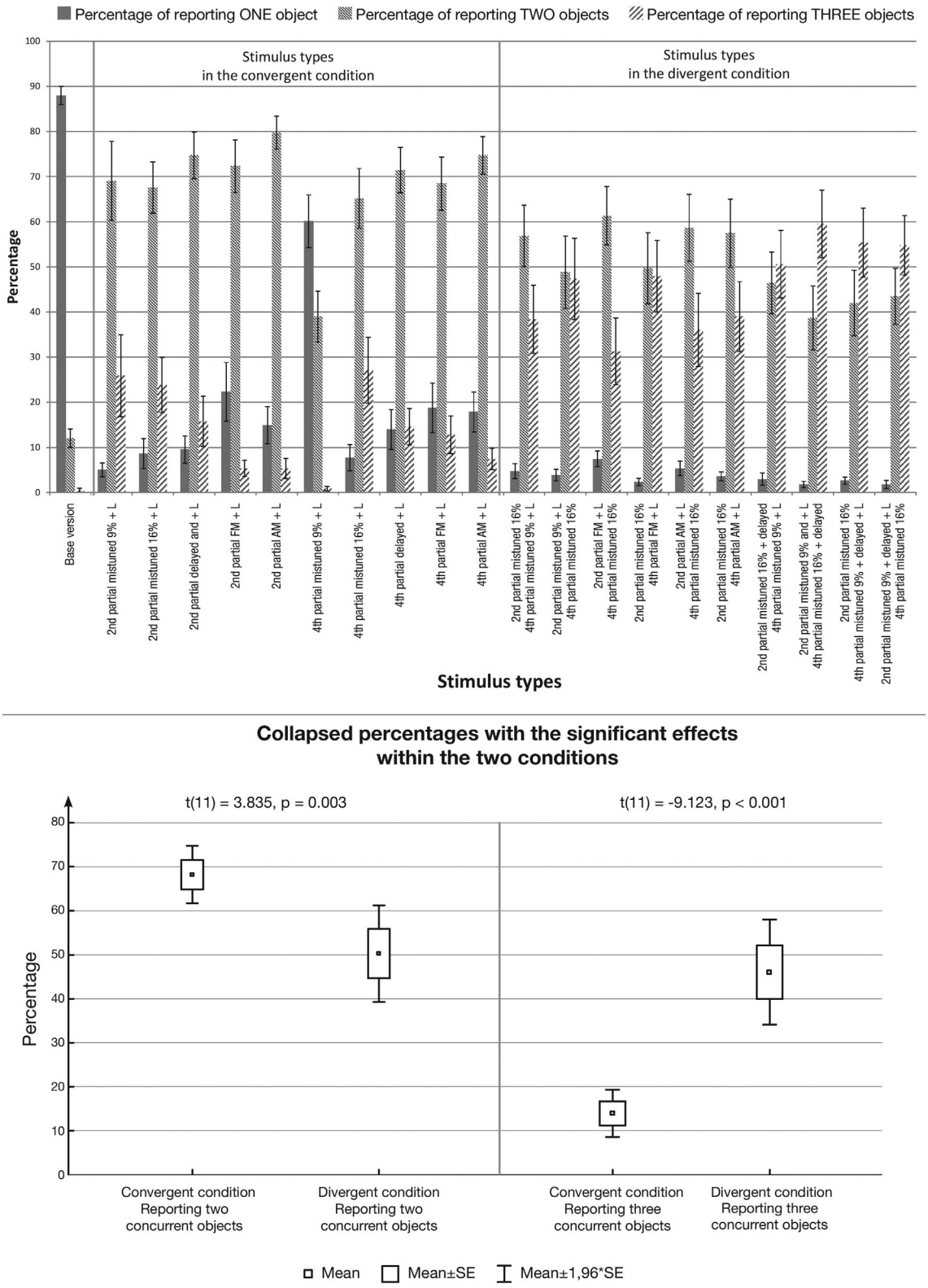

Fig. 2. Upper panel: Group-average ($N = 12$) percentages of reporting one, two, or three concurrent sound objects for each stimulus type in the base version, convergent, and divergent conditions along with the standard error of mean bars. The presence of the location cue is marked by "+ L". Lower panel: Two- and three-object responses collapsed across the different stimulus types separately for the convergent and the divergent conditions together with the results of the between-condition statistical comparisons.

participants were excluded due to poor signal-to-noise ratio in the ERPs (<55% artefact-free trials). Sixteen participants' (fourteen female, mean age 22.05 years, SD = 2.41) data were included in the analysis. All but one participant were right-handed and all had pure-tone thresholds within normal limits (<25 dB HL with <15 dB HL difference between the two ears) for the frequencies ranging from 250 to 8000 Hz. No participant reported taking any medication affecting the central nervous system. Participants were recruited from the Rotman Research Institute participant database, and received modest monetary compensation for their participation. Prior to the beginning of the experiment, written

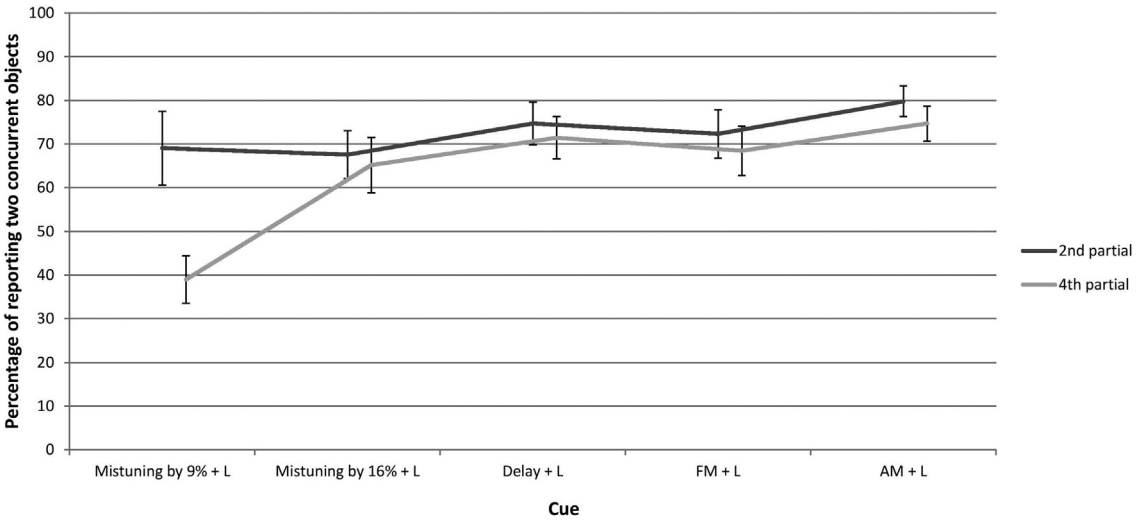

Fig. 3. Interaction between Harmonic number and Cue for the percentage of reporting two or three concurrent objects with the different stimulus types of the convergent condition. The horizontal axis shows the cue manipulations ("+ L" denotes the presence of the location cue), the vertical axis shows the average proportion of reporting multiple concurrent object. The dark grey line marks the 2nd partial, while the lighter grey line marks the 4th partial being manipulated. The standard error of means is shown by bars.

informed consent was obtained from each participant after the experimental procedures and aims of the study were explained to them. The study was approved by the Research Ethics Board of the Rotman Research Institute (Toronto, Canada).

3.1.2. Apparatus, stimuli and procedure

The apparatus and stimuli used in the EEG experiment were the same as in Experiment 1 except for the following: Seven blocks of 450 harmonic complex tones were presented to the participants while they watched a subtitled and muted movie (chosen from the list of available movies at the laboratory) on a computer screen that was placed approximately 140 cm in front of them. The base-version complex was repeated 150 times overall. As in Experiment 1, the convergent and divergent conditions each included 10 different stimulus types. These were repeated 150 times, each. Participants were thus presented with a total of 3150 stimuli. Note that whereas in Experiment 1, the three conditions received the same number of trials, in Experiment 2, each stimulus type received the same number of stimuli. Stimuli were delivered in a fully randomized order throughout the stimulus blocks. Each stimulus block lasted about nine minutes. The stimulus onset asynchrony was 1200 ms. The net duration of the stimulation was approximately 63 min. With breaks and electrode cap mounting and removal, the total time of the session was ca. 3 h.

3.1.3. Electrophysiological recording and data analysis

EEG was continuously recorded with 64 Ag/AgCl electrodes placed on the scalp according to the extended international 10–20 system (Jasper, 1958; Chatrian et al., 1985). Eye movements were monitored by electrooculogram (EOG) recordings from two electrodes placed below the eyes and two placed lateral to the outer canthi of both eyes. The Cz electrode served as online reference for all EEG and EOG electrodes, and during the offline analysis, data were re-referenced to the average of all signals. EEG and EOG signals were amplified (0–40 Hz) by SynAmps amplifiers (Neuroscan Inc.) and sampled at 500 Hz. Data were resampled to 250 Hz and filtered off-line using a 0.1–30 Hz band-pass finite impulse response (FIR) filter (Kaiser windowed, Kaiser $\beta = 5.65$, filter length 4530 points).

For each stimulus, an epoch of 600 ms duration including a 100 ms pre-stimulus baseline was extracted from the continuous EEG record. Epochs with an amplitude change exceeding 100 μV at any electrode were excluded from further analysis, which led to retaining 75.2% of the responses on average. Epochs for the 21 different stimulus types were separately averaged, collapsing over the two possible locations (left vs. right presentation).

For identifying and measuring the ORN component, difference waveforms were calculated separately between averaged ERPs elicited by the different manipulated tones and the average response to the base version tone. Measurements for statistical analysis were taken from the midline frontal site (Fz; based on the studies by Alain et al., 2001, 2002), except for a post-hoc analysis based on visual inspection that was conducted on ERPs recorded at the Pz electrode (see below). The midline frontal electrode was chosen because the ORN is often largest at frontal and frontocentral scalp sites (e.g., Alain et al., 2001, 2002). Average ORN amplitudes were measured from 60-ms wide windows in the 132–192 ms or the 212–272 ms latency range. The use of two different windows result from the fact that stimuli with a 100-ms temporal delay manipulation are known to elicit an ORN response commencing later than that evoked by the inharmonicity cue (e.g., Kocsis et al., 2014); further, the amplitude modulation manipulation appeared also to elicit longer-latency ORN. The 60-ms width of the windows was set to cover the variation of ORN peak latencies elicited by different cue combinations (e.g., the ORN peak latencies elicited by mistuning and FM cues fell into the 150–170 ms range and thus the window was centred on their mean value).

Following visual inspection of the traces, a second negative difference response was also measured, because those stimuli of the divergent conditions that included either the delay or the amplitude-modulation cue on one of the partials (see Table 1, conditions 15 to 20) appeared to have elicited two temporally separate negative ERP difference waveforms, the latter of which was regarded as a second ORN. The second ORN amplitudes were measured on the Pz electrode from 40-ms wide windows between 212 and 252 ms for each condition.

All ERP difference amplitudes (including the second ORN) were tested against zero using one-sample, one-tailed *t*-tests. In addition, the same statistical analyses as conducted in Experiment 1 for the proportions of marking two and/or three concurrent objects were conducted here on the average ORN amplitudes measured from Fz. Additivity between cue effects was tested with paired-samples two-tailed *t*-tests, separately comparing the ORN amplitudes elicited by the different stimulus types in the divergent condition (also adding the second ORN amplitude in conditions 15–20) with the summed amplitudes of the ORN components elicited by the constituting cues when presented in the convergent condition. To further test for the potential integration of the cues, paired-samples two-tailed *t*-tests were conducted, separately comparing the ORN amplitudes elicited by the stimulus types in the

divergent condition with the larger one of the constituting-cue ORNs from the convergent condition. All significant statistical results are reported. ANOVA effects are reported together with the partial η^2 effect size measure. The Greenhouse–Geisser correction was applied when the assumption of sphericity was violated; the ε correction factor is reported in these cases. Post-hoc tests for repeated-measures ANOVAs were carried out by the Bonferroni correction of the confidence level for multiple comparisons.

3.2. Results and discussion

Midline frontal (Fz) ERP responses elicited by the base-version tone and the manipulated harmonic complexes together with the corresponding difference waveforms are shown in Fig. 4, separately for all stimulus types. The corresponding scalp topography maps are presented in Fig. 5. With most manipulations, the ORN response appeared in its typical 130–200-ms latency range (peaks within the 150–170 ms range; amplitudes measured from the 132–192 ms range). However, for the manipulation delaying the onset of one of the tonal elements by 100 ms and for the amplitude-modulation manipulation, the resulting ORN was elicited in the 212–272-ms latency range, since there is not enough evidence yet to refer to a typical latency range. Thus, similarly to the delay cue, the slow amplitude modulation cue was picked up later by the auditory system than the spectral cues.

In the convergent condition, ORN amplitudes were found to be significantly different from zero for all mistuning and delay manipulations, except for when the 4th partial was mistuned by 9%, whereas none of the manipulations with amplitude or frequency modulation yielded significant ORNs. In contrast, in the divergent condition, all stimulus types elicited significant ORNs (see Fig. 4 and Table 2).

The paired *t*-test between the ORN amplitudes pooled across all convergent- versus all divergent-condition stimulus types showed a significant difference ($t(15) = 2.202$, $p < 0.05$). This was due to the ORNs elicited in the convergent condition being significantly smaller than those elicited in the divergent condition.

The repeated-measures ANOVA comparing the ORN amplitudes elicited by different cues in the convergent condition showed a significant interaction between *Harmonic number* and *Cue* ($F(4,60) = 10.898$, $p < 0.001$, $\eta^2 = 0.421$). To decompose the interaction, separate paired-sample *t*-tests between the 2nd and 4th harmonics were conducted for each cue. These tests revealed a significant difference between the harmonic numbers only for the mistuning by 9% ($t(15) = -8.801$, $p < 0.001$, all other p values >0.05), with the 9% mistuning cue eliciting a smaller-amplitude ORN when applied on the 4th than on the 2nd harmonic (Fig. 6). This also explains the significant main effect of *Harmonic number* ($F(1,15) = 20.024$, $p < 0.001$, $\eta 2 = 0.572$). There was also a significant main effect of *Cue* ($F(4,60) = 10.969$, $p < 0.001$, $\eta 2 = 0.422$). Post-hoc pairwise comparisons revealed that mistuning by 16% elicited significantly larger ORN amplitudes than the mistuning by 9% (caused by the lower ORN amplitude elicited when the manipulation was applied to the 4th partial), the FM, and the AM manipulations ($p < 0.05$ in all comparisons), and that the delay manipulation elicited significantly larger ORN amplitudes than the FM and AM manipulations ($p < 0.005$ in both comparisons).

In those six divergent stimulus types that included either delay or amplitude modulation as a cue, it is reasonable to expect that the late ORN elicited by these manipulations could appear separately from the

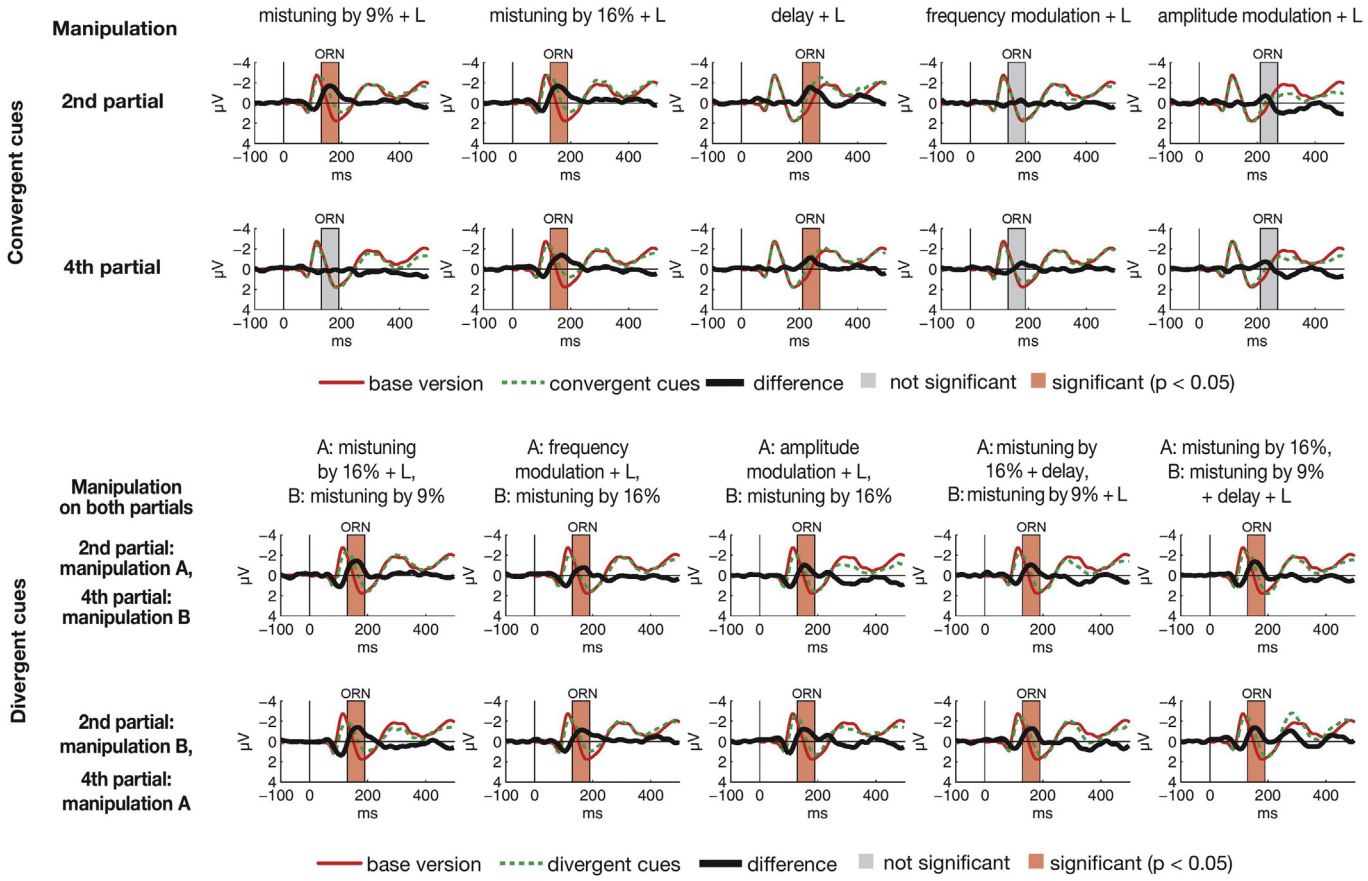

Fig. 4. Group-average ($N = 16$) midline frontal (Fz) ERPs elicited by the 20 different manipulated (green dashed line) and the base-version tones (red solid line) during passive listening, together with their difference waveforms (thick black line). Stimulus onset is at the crossing of the *x* and *y* axes. The location cue is marked by "+ L". Note that for delayed and amplitude-modulated partials ORN appears at a later latency range than for the other manipulations. Measurement time windows are marked as rectangles, red for significant, grey for nonsignificant components. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

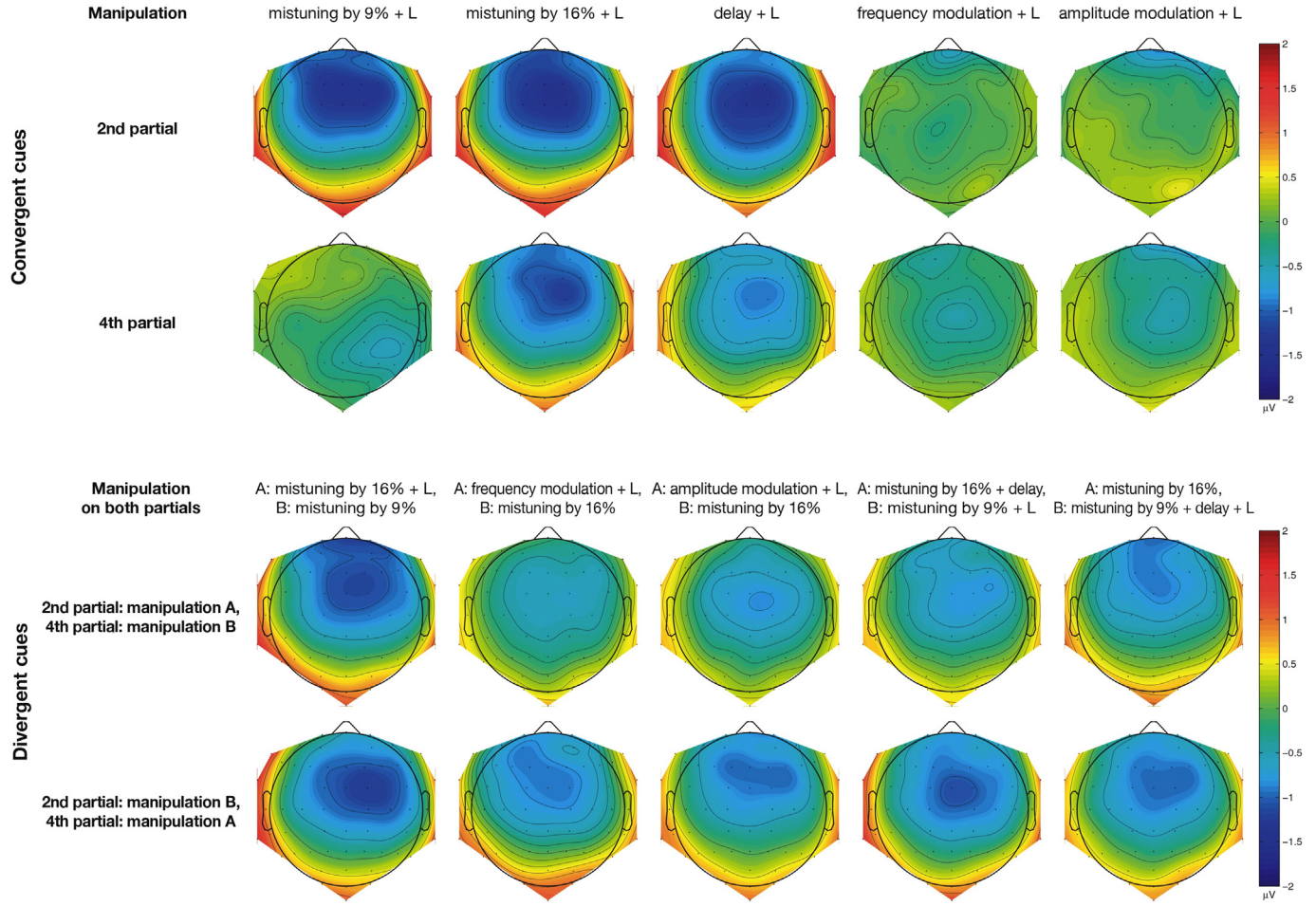

Fig. 5. Group-average ($N = 16$) ORN scalp topographies of the difference amplitudes measured in the 132–192 ms window for all divergent-condition stimulus types and for those convergent condition stimulus types that do not include delay or AM manipulation. For the convergent-condition stimulus types with the delay or the AM cue, the scalp topography is based on the 212–272 ms window. The presence of the location cue is marked by "+ L". The common voltage scale is shown at the right side of the figure.

early ORN elicited by the spectral cues. Indeed, noticeable ERP differences between the manipulated and base-version tones were visible not only in the typical early ORN time window but also later, at about 200 ms from stimulus onset. The time window of the second difference response is compatible with the one observed for the delay and the amplitude-modulation manipulations when presented alone in the convergent conditions. However, unlike the rest of the ORN responses, the second difference response was most prominent over parietal sites (Fig. 7). Parietal (Pz) difference waveforms significantly differed from zero in this second latency range (i.e., 212–252 ms) in all cases except when the 2nd partial was mistuned by 16% while the 4th partial was amplitude-modulated (see Table 3). Fig. 8 shows the scalp topography

Table 2
Group-average ($N = 16$) frontal (Fz) ERP amplitudes measured in the ORN latency range of the manipulated-minus-base difference waveform for the 20 different manipulated stimulus types.

Convergent condition	2nd partial mistuned 9% + L	2nd partial mistuned 16% + L	2nd partial delayed + L	2nd partial FM + L	2nd partial AM + L	4th partial mistuned 9% + L	4th partial mistuned 16% + L	4th partial delayed + L	4th partial FM + L	4th partial AM + L
Mean amplitude at Fz (μV)	−1.323	−1.36	−1.175	−0.084	−0.174	0.146	−1.039	−0.776	−0.189	−0.361
t(15)	−7.242***	−7.404***	−4.705***	−0.552	−0.591	0.696	−4.361***	−3.662**	−1.266	−1.316
Divergent condition	2nd partial mistuned 16% + L and 4th partial mistuned 9%	2nd partial mistuned 9% and 4th partial mistuned 16% + L	2nd partial FM + L and 4th partial mistuned 16%	2nd partial mistuned 16% and 4th partial FM + L	2nd partial AM + L and 4th partial mistuned 16%	2nd partial mistuned 16% and 4th partial AM + L	2nd partial mistuned 16% + delayed and 4th partial mistuned 9% + L	2nd partial mistuned 9% + L and 4th partial mistuned 16% + delayed	2nd partial mistuned 16% and 4th partial mistuned 9% + delayed + L	2nd partial mistuned 9% + delayed + L and 4th partial mistuned 16%
Mean amplitude at Fz (μV)	−1.07	−0.953	−0.493	−0.842	−0.604	−0.942	−0.687	−0.924	−0.907	−0.922
t(15)	−4.634***	−4.809***	−3.101**	−7.223***	−2.36*	−4.127***	−3.361**	−4.484***	−4.7***	−3.801**

Notes: Significant differences from zero are marked with asterisks (* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$); the presence of the location cue is marked by "+ L.".

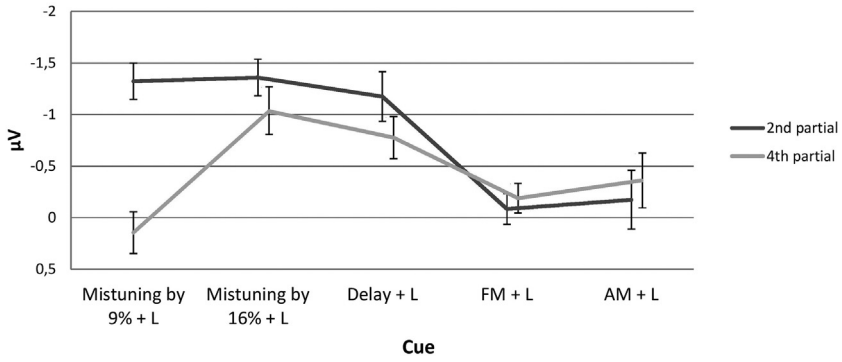

Fig. 6. Interaction between Harmonic number and Cue for the ORN amplitudes elicited by the different stimulus types of the convergent condition. The horizontal axis shows the cue manipulations ("+ L" denotes the presence of the location cue), the vertical axis shows the average ORN amplitude values. The dark grey line marks the amplitudes elicited by manipulations affecting the 2nd, lighter grey line the 4th partial. The standard error of means is shown by bars.

of the second difference responses for the six divergent stimulus types. Although the scalp distribution of this response is different from that of the typical ORN response, the second difference between manipulated and base-version sounds may reflect the additional object-related processing that could lead to the three-object percept. Note, however, that this interpretation cannot be tested from the current data, because participants were not asked to report their perception in Experiment 2.

The additivity analysis showed that summing two or three ORN amplitudes elicited by sounds with convergent-condition cues always produced numerically larger values than the amplitude of the ORN elicited by the corresponding divergent-condition stimulus (also summing the second-ORN amplitude, where applicable), although the differences did not always reach significance (see Table 4).

Comparing the divergent-condition ORN amplitudes with the larger one of the constituting ORN amplitudes, it was found that the larger constituent ORN amplitude is not significantly different from or, in some cases, even significantly larger than the corresponding divergent-condition ORN amplitude (where applicable, the second-ORN amplitude was summed with the early ORN amplitude; see Table 5).

In summary, it was found that most cue combinations evoked a significant ORN response. Exceptions were convergent-condition stimulus types with the amplitude- or frequency-modulation cue and with mistuning the 4th partial by 9%. Further, the ORN amplitude was significantly larger for mistuning by 9% the 2nd than the 4th harmonic, and it was larger for the delay and the mistuning by 16% than for the AM and FM cues. Most divergent manipulations that included the delay or the AM cue elicited a significant second late difference between the manipulated and the base-version sounds in addition to the typical-latency ORN response. Note that observation of this late effect, although reasonable based on the finding of later ORN response for one of the constituent manipulations, was based on post-hoc visual inspection in the present study, thus it requires replication. Subadditivity was observed between the ORN amplitudes for divergent cues relative to the sum of the constituting convergent-cue ORN amplitudes. The larger one of the ORNs elicited by one of the constituent cues was either not significantly different from or even significantly larger than the ORN amplitude elicited in the corresponding divergent condition.

4. General discussion

We aimed 1) to create stimuli allowing one to investigate the effects of perceiving three concurrent auditory objects by divergently manipulating two tonal elements and 2) to test how the ORN event-related potential component reflects the evaluation of such cues. The results of Experiment 1 suggested that listeners were more likely to report hearing three concurrent sounds with the divergent than with the convergent manipulations, which makes it possible that listeners are capable of detecting three concurrent sounds in manipulated complex tones. Still, participants' performance was relatively low in discriminating between two and three concurrent sounds in the divergent condition. The

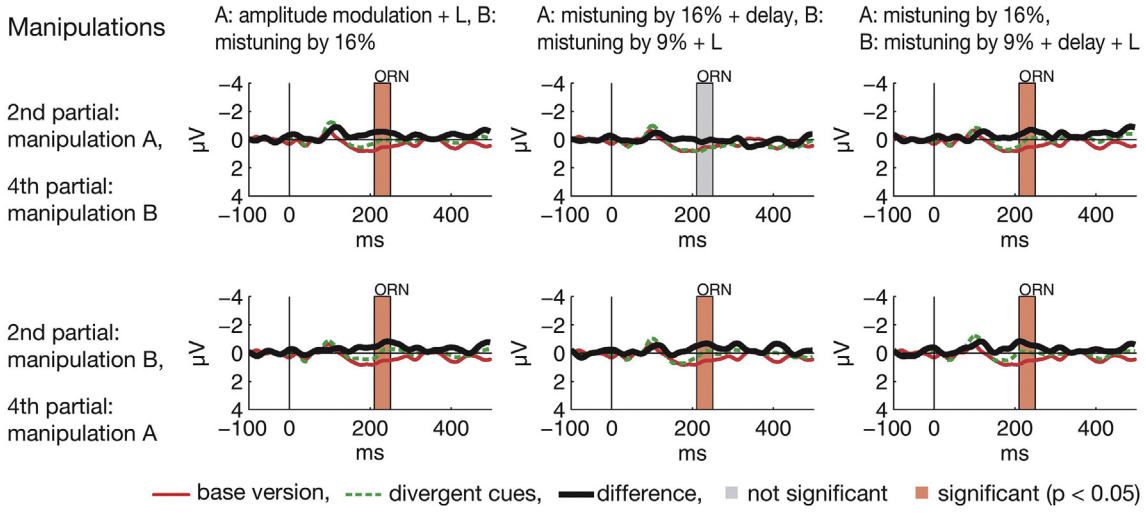

Fig. 7. Group-average ($N = 16$) midline parietal (Pz) ERPs elicited in the 6 divergent stimulus types that included the delay or the AM cue (green dashed line) and the base-version tones (red solid line) during passive listening, together with the corresponding difference waveforms (thick black line). Stimulus onset is at the crossing of the x and y axes. The presence of the location cue is marked by "+ L". Measurement time windows are marked as rectangles, red for significant, grey for nonsignificant components. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 3
Group-average ($N = 16$) midline parietal (Pz) ERP amplitudes measured in the 212–252 ms latency range of the manipulated-minus-base difference waveform for the 6 divergent condition stimulus types in which the delay or the AM cue was used.

Stimulus type	2nd partial AM + L and 4th partial mistuned 16%	2nd partial mistuned 16% and 4th partial AM + L	2nd partial mistuned 16% + delayed and 4th partial mistuned 9% + L	2nd partial mistuned 9% + L and 4th partial mistuned 16% + delayed	2nd partial mistuned 16% and 4th partial mistuned 9% + delayed + L	2nd partial mistuned 9% + delayed + L and 4th partial mistuned 16%
Mean amplitude at Pz (μV)	−0.527	0.098	−0.538	−0.644	−0.605	−0.678
t(15)	−2.611*	0.507	−2.796*	−3.85**	−3.15**	−3.054**

Notes: Significant differences from zero are marked with asterisks ($p < 0.05$, ** $p < 0.01$, *** $p < 0.001$); the presence of the location cue is marked by "+ L".*

ERP results of Experiment 2 are generally compatible with the behavioral ones: In Experiment 1, participants were more likely to perceive three concurrent sounds in the divergent condition than in the convergent condition; on this basis, divergent manipulations could be expected to elicit larger ORNs than convergent ones, which was the case in Experiment 2. Yet the ORN amplitudes were subadditive when comparing the divergent-condition stimulus types with the constituting convergent-condition ones. One explanation is that the ORN reflects only the likelihood of the presence of multiple concurrent sounds without carrying information regarding the number of concurrent sounds. With more cues, the likelihood of the presence of multiple concurrent sounds increases, which results in larger ORN amplitudes and larger probability of the listener reporting hearing more than one sound. Thus this interpretation of the results is consistent with the notion that ORN reflects the auditory system's overall readout of the likelihood that the sound input carries contributions from multiple sound sources. This is in agreement with previous findings on the functional significance of ORN (e.g., Kocsis et al., 2014) and extends it to situations with divergent cues. One possible implementation of this type of function is the horse-race model (Mordkoff and Yantis, 1991), which allows fast evaluation of the evidence. If concurrent sound segregation operated on the principle of the horse-race model, then separate decisions would be formed in parallel for each cue, and the outcome (leading to perception) would be based on the first (if any) cue suggesting the likely presence of multiple sound sources. That is, the most salient cue would drive the perceptual decision. This type of processing is compatible with the foreground-background organization of perception in which one object is selected for the focus of attention while the remaining ones are not distinguished. In our experiment, we found that the cue eliciting the largest ORN produced approximately equal or, in some cases, significantly larger ORN amplitude than that obtained in the corresponding divergent condition. This result is fully consistent with the horse-race model. There is also further behavioral and electrophysiological evidence supporting this view (Brochard et al., 1999; Sussman et al., 2005; Leung et al., 2011, 2015; Kulagina et al., 2015) when the concurrently active sound streams are qualitatively similar to each other (Cusack et al., 2004). The sound objects that could be extracted from the current stimuli were all tonal, and thus indeed qualitatively similar. However, it should be noted that the additivity tests are slightly confounded by the location cue being present in all constituting convergent cues, whereas in the divergent cues, location difference was only added to one of the manipulated partials.

In the behavioral experiment, listeners reported perceiving three distinct sounds in approximately half of the divergent-condition trials, significantly more often than in the convergent trials. This is in line with the fact that a significant difference between the ORNs elicited by convergent and divergent cues was found, although it may have been caused by the divergent cues always eliciting significant ORN

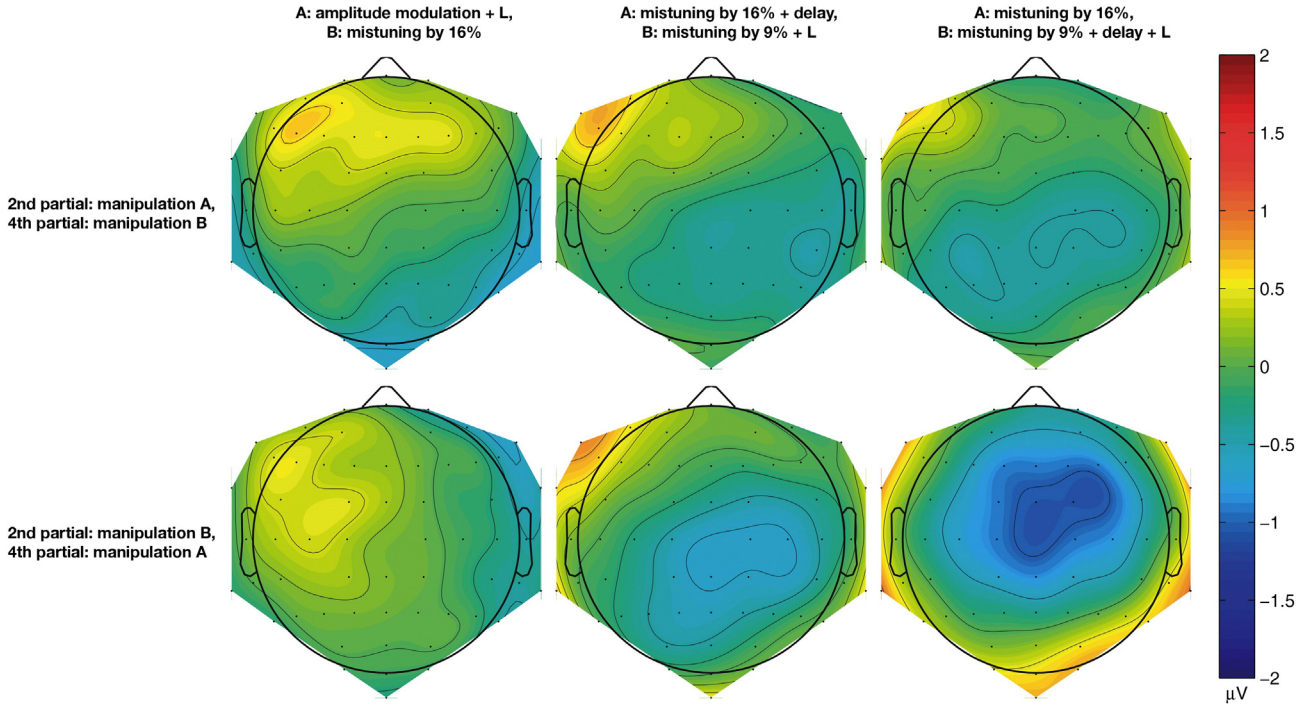

Fig. 8. Group-average ($N = 16$) scalp topographies of the difference amplitudes measured in the 212–252 ms window for divergent-condition stimulus types that included either the delay or the AM cue. The presence of the location cue is marked by "+ L". The common voltage scale is shown at the right side of the figure.

Table 4
Testing the additivity of ORN amplitudes for multiple divergent cues of concurrent sound segregation. For the stimulus types including neither the AM nor the delay cue, only the early time window measurements are used. For the stimulus types including either the AM or the delay cue, the convergent condition's corresponding time window is used and the divergent condition amplitudes are summed from both the early and late time window. Amplitudes are measured from the midline frontal (Fz) electrode. The ORN amplitudes elicited by the contributing convergent cues and the ORN amplitude elicited by the corresponding divergent-condition stimuli are shown together with the t and *p* values for the paired two-tailed t tests.

Stimulus types including neither AM nor delay	Mean amplitude at Fz (μV)	t(15)
2nd partial mistuned by 16% + L + 4th partial mistuned by 9% + L (condition 2 + 6)	−1.214	−0.621
2nd partial mistuned 16% + L and 4th partial mistuned 9% (condition 11)	−1.071	
2nd partial mistuned 9% + 4th partial mistuned 16% (condition 1 + 7)	−2.362	−4.977***
2nd partial mistuned 9% and 4th partial mistuned 16% + L (condition 12)	−0.953	
2nd partial with frequency modulation (5 Hz) + 4th partial mistuned 16% (condition 4 + 7)	−1.122	−2.211**
2nd partial frequency modulated (5 Hz) + L and 4th partial mistuned 16% (condition 13)	−0.493	
2nd partial mistuned 16% + 4th partial with frequency modulation (5 Hz) (condition 2 + 9)	−1.549	−3.075**
2nd partial mistuned 16% and 4th partial frequency modulated (5 Hz) + L (condition 14)	−0.842	
Stimulus types including either delay or AM		
2nd partial with amplitude modulation (5 Hz) + L + 4th partial mistuned 16% + L (condition 5 + 7)	−1.213	−1.489
2nd partial amplitude modulated (5 Hz) + L and 4th partial mistuned 16% (condition 15)	−0.632	
2nd partial mistuned 16% + L + 4th partial with amplitude modulation (5 Hz) + L (condition 2 + 10)	−1.721	−1.626
2nd partial mistuned 16% 4th partial amplitude modulated (5 Hz) + L (condition 16)	−1.245	
2nd partial mistuned 16% + L + 2nd partial delayed + L + 4th partial with frequency modulation (5 Hz) + L (condition 2 + 3 + 9)	−2.724	−8.407***
2nd partial mistuned 16% and delayed and 4th partial mistuned 9% + L (condition 17)	−0.587	
2nd partial mistuned 9% + L + 4th partial mistuned 16% + L + 4th partial delayed + L (condition 1 + 7 + 8)	−3.138	−5.998***
2nd partial mistuned 9% + L and 4th partial mistuned 16% and delayed (condition 18)	−0.805	
2nd partial mistuned 16% + L + 4th partial mistuned 9% + L + 4th partial delayed + L (condition 2 + 6 + 8)	−1.99	−4.196***
2nd partial mistuned 16% and 4th partial mistuned 9% and delayed + L (condition 19)	−0.727	
2nd partial mistuned 9% + L + 2nd partial delayed + L + 4th partial mistuned 16% + L (condition 1 + 3 + 7)	−3.537	−5.808***
2nd partial mistuned 9% and delayed + L and 4th partial mistuned 16% (condition 20)	−1.285	

Notes: Significant differences are marked with asterisks (* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$); the presence of the location cue is marked by "+ L".

Table 5
Comparing between the larger one of the constituting ORN amplitudes with the corresponding divergent-condition ORN amplitude (where applicable, the second-ORN amplitude was summed with the early ORN amplitude).

Stimulus types to compare	Mean amplitude at Fz (μV)	t(15)
2nd partial mistuned 16% + L (condition 2)	−1.359	−2.047
2nd partial mistuned 16% + L and 4th partial mistuned 9% (condition 11)	−1.071	
2nd partial mistuned 9% + L (condition 1)	−1.323	−1.935
2nd partial mistuned 9% and 4th partial mistuned 16% + L (condition 12)	−0.953	
4th partial mistuned 16% + L (condition 7)	−1.039	−2.267*
2nd partial frequency modulated (5 Hz) + L and 4th partial mistuned 16% (condition 13)	−0.493	
2nd partial mistuned 16% + L (condition 2)	−1.359	−2.809*
2nd partial mistuned 16% and 4th partial frequency modulated (5 Hz) + L (condition 14)	−0.842	
4th partial mistuned 16% + L (condition 7)	−1.039	−0.758
2nd partial amplitude modulated (5 Hz) + L and 4th partial mistuned 16% (condition 15)	−0.632	
2nd partial mistuned 16% + L (condition 2)	−1.359	−0.327
2nd partial mistuned 16% 4th partial amplitude modulated (5 Hz) + L (condition 16)	−1.245	
2nd partial mistuned 16% + L (condition 2)	−1.359	−2.307*
2nd partial mistuned 16% and delayed and 4th partial mistuned 9% + L (condition 17)	−0.587	
2nd partial mistuned 9% + L (condition 1)	−1.323	−1.466
2nd partial mistuned 9% + L and 4th partial mistuned 16% and delayed (condition 18)	−0.805	
2nd partial mistuned 16% + L (condition 2)	−1.359	−1.537
2nd partial mistuned 16% and 4th partial mistuned 9% and delayed + L (condition 19)	−0.727	
2nd partial mistuned 9% + L (condition 1)	−1.323	−0.105
2nd partial mistuned 9% and delayed + L and 4th partial mistuned 16% (condition 20)	−1.285	

Notes: Significant differences are marked with asterisks (* $p < 0.05$); the presence of the location cue is marked by "+ L".

amplitudes, whereas in the convergent condition, weaker cues did not elicit a significant response (in accordance with Kocsis et al., 2014). Thus it is possible that the ORNs elicited by divergent cues were produced by an integration process of the two or three cue's ORN responses. The (post-hoc) finding of a second negative response appearing in those divergent trials that included either the delay or the AM as one of the manipulations is compatible with this view. Because these manipulations delay the emergence of ORN, the second, late negativity observed in response to these sounds possibly reflects the later ORN activity in connection with the processing of these manipulations or, taking into account that the scalp distribution of this response is different from that of ORN, perhaps additional processing aimed at further specifying the auditory scene. On this account, the observed subadditivity of ORN amplitudes and the less-than-reliable behavioral discrimination of hearing two or three concurrent sounds can be explained as an effect of insufficient salience of the concurrent sounds. That is, although separately the cues may be sufficient for segregating the manipulated partials from the complex tone, they are not sufficient for making all three separations for a clear perception of three concurrent sounds. Because in Experiment 2, listeners were not asked to report the number of the objects they perceived, we could not compare responses between those trials when listeners perceived three as opposed to two objects. This prevents us from verifying the hypothetical role of the process reflected by the second ORN. Thus there is no definite answer to the question whether ORN can reflect the number of concurrent objects.

In the present study, we employed several different types of cues in order to avoid basing the conclusions on one type of cue, exclusively. The cue effects found both for the probability of perceiving multiple concurrent sounds and for the ORN amplitudes suggest that the various cues are not equally effective in promoting the segregation of concurrent sounds. Mistuning the 4th partial by 9% led to concurrent sound segregation with significantly lower probability than any other manipulation, and it failed to elicit a significant ORN response. This finding contrasts those of some previous studies that found significant ORN elicited by even smaller amounts of mistuning on higher partials. Alain et al. (2001) obtained reliable ORN responses for passive and also for active listening with 8% mistuning of the 2nd partial of a 12-partial complex tone in one condition. The probability of the fully harmonic tone was ca. 17% in the stimulus blocks of this study. Kocsis et al. (2014) found significant ORN for 8% mistuning of the 2nd and also of the 4th partial in a 5-partial complex tone. In each stimulus block of Kocsis et al.

(2014), half of the tones were fully harmonic and the other half was manipulated. In contrast to both of these studies, the proportion of the base-version sounds was very low in the current experiment (5% across the blocks). This could have led to lower ORN amplitudes, perhaps through refractoriness. This explanation is supported by the finding of lower ORN amplitudes with higher manipulated-sound probabilities (Bendixen et al., 2010).

The harmonic complexes with an amplitude- or frequency-modulated partial did not elicit significant ORN components. In contrast, in Experiment 1, the same harmonic complexes mostly yielded perception of two concurrent sound objects, although the probability of this was numerically lower than that of the most effective cues. There are two possible explanations for this inconsistency. It is possible that the inconsistent results point to lower sensitivity of the ERP than that of the behavioral measures of concurrent sound segregation. However, contrasting evidence has been obtained in Kocsis et al. (2014) where, for the location-difference cue, behavioral measures did not reflect the perception of two concurrent sounds, whereas a significant ORN was elicited. Alternatively, the inconsistent results obtained in the two experiments may stem from the different stimulus probabilities employed. In Experiment 1, sounds promoting the perception of a single object were presented on 33.3% of the trials. In contrast, in Experiment 2, the proportion of base-version trials amounted only to about 5% of the trials. Thus ORN was expected to be elicited in 95% of the trials, which may have attenuated ORN amplitudes, as was shown by Bendixen et al. (2010). Finally, AM and FM may be weaker cues of concurrent sound segregation, therefore requiring attention to segregate sounds separated by these cues (which was devoted to the stimuli in Experiment 1 but not 2). Charbonneau (1981) found that if the slightly different modulation functions of partials in natural instrument tones are replaced by the same modulation function, the difference is undetectable for human listeners. Along the same lines, when competing vowels were modulated (whether in the same phase or not), concurrent segregation did not occur. Segregation was only likely to be successful if one vowel was modulated while the other was not (Summerfield et al., 1992). Summerfield and colleagues employed parameters similar to our experiment (base frequencies of 100–141 Hz, modulation frequency of 2.5 and 8 Hz for AM and FM, respectively). However they presented vowels, not complex tones. Furthermore, the AM cue was found to be effective in a stream-segregation task with tones of higher frequencies than the current ones (i.e., 1000 and 4000 Hz with 30, 100, 300 Hz modulation frequency; Dolležal et al., 2012). Further, Bregman et al. (1985, 1990) showed that the fusion of tones was strongest when two tones had the same AM frequency, and this may overwrite even a non-harmonic relation between the frequencies of the two tones. That is, Bregman and colleagues, basing on the common fate principle, employed common AM to integrate two tones that would otherwise have been more likely segregated. However, they employed different parameters than the previously discussed studies: 500–1500 Hz carrier frequencies with 100 Hz modulation frequency, and 1600–3000 Hz carrier frequencies with 125 Hz modulation frequency, respectively. The variance of these findings suggest that the size of the effects of AM (and probably also FM) difference on concurrent sound segregation may depend on the paradigm and the parameters employed. Therefore, a systematic study of the effects of these cues is needed to clarify their role and efficacy in auditory scene analysis.

Finally, we found that mistuning of the 4th partial by 9% was less effective in promoting the perception of two concurrent sounds and elicited ORNs with lower amplitude than the same manipulation occurring on the 2nd partial. This is generally in line with the results reported by Alain et al. (2001). It thus appears that for some sound manipulations (here: a moderate amount of mistuning), higher partials provide less salient cues for concurrent sound segregation.

In conclusion, the current results complement previous findings in that it is possible to create the perception of three concurrent sounds in a complex tone paradigm. Although some of the ORN evidence is compatible with the notion that the ORN response reflects the overall readout of the auditory system regarding the presence of multiple concurrent objects (thus extending the finding of Kocsis et al., 2014), there are signs of possible additional processing when divergent cues promote the presence of three concurrent objects. To maximize the possibility of detecting the ERP correlates of these putative additional processes, further research should a) employ qualitatively different concurrent sounds and b) measure ERPs while asking listeners to mark the number of concurrent sounds they have perceived.

Acknowledgments

This work was funded by the Erasmus Mundus Student Exchange Network in Auditory Cognitive Neuroscience to Z.K., by the Hungarian Academy of Sciences (Magyar Tudományos Akadémia [MTA], Lendület project LP2012-36/2012) to I.W., the Canadian Institute for Health Research (MOP106619, C.A.), and the Natural Sciences and Engineering Research Council of Canada (C.A.). The experiment was realized using Cogent 2000 developed by the Cogent 2000 team at the FIL and the ICN. EEG data were analysed with EEGlab (Delorme & Makeig, 2004) and additional plugins written by Andreas Widmann, University of Leipzig. The authors are grateful to Yu He for assistance in data acquisition.

References

Alain, C., Arnott, S.R., Picton, T.W., 2001. Bottom-up and top-down influences on auditory scene analysis: evidence from brain potentials. J. Exp. Psychol. 27, 1072–1089.
Alain, C., Schuler, B.M., McDonald, K.L., 2002. Neural activity associated with distinguishing concurrent auditory objects. J. Acoust. Soc. Am. 111, 990–995.
Alain, C., Reinke, K.S., He, Y., Wang, C., Lobaugh, N., 2005. Hearing two things at once: neurophysiological indices of speech segregation and identification. J. Cogn. Neurosci. 17, 811–818.
Arnott, S.A., Bardouille, T., Ross, B., Alain, C., 2011. Neural generators underlying concurrent sound segregation. Brain Res. 1387, 116–124.
Bendixen, A., Jones, S.J., Klump, G., Winkler, I., 2010. Probability dependence and functional separation of the object-related and mismatch negativity event-related potential components. NeuroImage 50, 285–290.
Bregman, A.S., 1990. Auditory Scene Analysis: The Perceptual Organization of Sound. The MIT Press, Cambridge, Massachusetts.
Bregman, A.S., Pinker, S., 1978. Auditory streaming and the building of timbre. Can. J. Psychol. 32, 19–31.
Bregman, A.S., Abramson, J., Doehring, P., Darwin, C.J., 1985. Spectral integration based on common amplitude modulation. Percept. Psychophys. 37, 483–493.
Bregman, A.S., Levitan, R., Liao, C., 1990. Fusion of auditory components: effects of the frequency of amplitude modulation. Percept. Psychophys. 47, 68–73.
Brochard, R., Drake, C., Botte, M.C., McAdams, S., 1999. Perceptual organization of complex auditory sequences: effect of number of simultaneous subsequences and frequency separation. J. Exp. Psychol. Hum. Percept. Perform. 25, 1742–1759.
Bronkhorst, A.W., Plomp, R., 1988. The effect of head-induced interaural time and level differences on speech intelligibility in noise. J. Acoust. Soc. Am. 83, 1508–1516.
Carlyon, R.P., 1991. Discriminating between coherent and incoherent frequency modulation of complex tones. J. Acoust. Soc. Am. 89 (1), 329–340.
Charbonneau, G.R., 1981. Timbre and the perceptual effects of three types of data reduction. Comput. Music. J. 5, 10–19.
Chatrian, G.E., Lettich, E., Nelson, P.L., 1985. Ten percent electrode system for topographic studies of spontaneous and evoked EEG activity. Am. J. EEG Technol. 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. J. Exp. Psychol. Hum. Percept. Perform. 30, 643–656.
Delorme, A., Makeig, S., 2004. EEGLAB: an open source toolbox for analysis of singletrial EEG dynamics including independent component analysis. J. of Neurosci. Methods 134, 9–21.
Dolležal, L.V., Beutelmann, R., Klump, G.M., 2012. Stream segregation in the perception of sinusoidally amplitude-modulated tones. PLoS One 7 (9), e43615.
Du, Y., He, Y., Ross, B., Bardouille, T., Wu, X., Li, L., Alain, C., 2011. Human auditory cortex activity shows additive effects of spectral and spatial cues during speech segregation. Cereb. Cortex 21, 698–707.
Hartmann, W.M., 1985. Perceptual Entities from Complex Inharmonic Tones. Association for Research in Otolaryngology 8th Meeting, p. 168 abs.
Hartmann, W.M., McAdams, S., Smith, B.K., 1986. Matching the Pitch of a Mistuned Harmonic in a Complex Sound. IRCAM Annual Report, pp. 54–63.
Hautus, M.J., Johnson, B.W., Colling, L.J., 2009. Event-related potentials for interaural time differences and spectral cues. Neuroreport 20, 951–956.
Jasper, H.H., 1958. The ten-twenty electrode system of the international federation. Electroencephalogr. Clin. Neurophysiol. 10, 370–375.
Johnson, B.W., Hautus, M., Clapp, W.C., 2003. Neural activity associated with binaural processes for the perceptual segregation of pitch. Clin. Neurophysiol. 114, 2245–2250.

Kocsis, Z., Winkler, I., Szalárdy, O., Bendixen, A., 2014. Effects of multiple congruent cues on concurrent sound segregation during passive and active listening: an event-related potential (ERP) study. Biol. Psychol. 100, 20–33.
Kubovy, M., Van Valkenburg, D., 2001. Auditory and visual objects. Cognition 80 (1–2), 97–126.
Kulagina, E., Drisdelle, B.L., Alain, C., Grimault, S., Eck, D., Vachon, F., Jolicoeur, P., 2015. The perception of concurrent sound objects through the use of harmonic enhancement: a study of auditory attention. Atten. Percept. Psychophys. 77 (3), 922–929.
Leung, A.W.S., Jolicoeur, P., Vachon, F., Alain, C., 2011. Concurrent sound perception impairs gap detection: implication for object-based account of auditory attention. J. Exp. Psychol. Hum. Percept. Perform. 37, 727–736.
Leung, A.W.S., Jolicoeur, P., Alain, C., 2015. Neuroelectric evidence of competition for attention during concurrent sound perception. J. Cogn. Neurosci. 27, 2186–2196.
Lipp, R., Kitterick, P., Summerfield, Q., Bailey, P.J., Paul-Jordanov, I., 2010. Concurrent sound segregation based on inharmonicity and onset asynchrony. Neuropsychologia 48, 1417–1425.
McAdams, S., 1984a. Spectral Fusions, Spectral Parsing and the Formation of Auditory Images Ph.D. thesis Stanford University.
McAdams, S., 1984b. The Auditory Image: A Metaphor for Musical and Psychological Research on Auditory Organization, in: Cognitive Processes in the Perception of Art, Crozier, R. and Chapman, A. (eds), pp. 183–187, (North Holland, Amsterdam)
McDonald, K.L., Alain, C., 2005. Contribution of harmonicity and location to auditory object formation in free field: evidence from event-related brain potentials. J. Acoust. Soc. Am. 118, 1593–1604.
Moore, B.C., Glasberg, B.R., Peters, R.W., 1986. Thresholds for hearing mistuned partials as separate tones in harmonic complexes. J. Acoust. Soc. Am. 80, 479–483.
Mordkoff, J.T., Yantis, S., 1991. An interactive race model of divided attention. J. Exp. Psychol. Hum. Percept. Perform. 17, 520–538.
Rasch, R.A., 1978. The perception of simultaneous notes as in polyphonic music. Acustica 40, 21–33.
Sanders, L.D., Joh, A.S., Keen, R.E., Freyman, R.L., 2008a. One sound or two? Object-related negativity indexes echo perception. Percept. Psychophys. 70, 1558–1570.
Sanders, L.D., Zobel, B.H., Freyman, R.L., Keen, R., 2008b. Manipulations of listeners' echo perception are reflected in event-related potentials. J. Acoust. Soc. Am. 129, 301–309.
Snyder, J.S., Alain, C., 2005. Age-related changes in neural activity associated with concurrent vowel segregation. Cognit. Brain Res. 24, 492–499.
Summerfield, Q., Culling, J.F., Fourcin, A.J., 1992. Auditory segregation of competing voices: absence of effects of FM and AM coherence. Philos. Trans. R. Soc. Lond., B. Biol. Sci. 336, 357–366.
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. Cogn. Affect. Behav. Neurosci. 5 (1), 93–110.
Weise, A., Schröger, E., Bendixen, A., 2012. The processing of concurrent sounds based on inharmonicity and asynchronous onsets: an object-related negativity (ORN) study. Brain Res. 1439, 73–81.

4. General discussion

The goal of this thesis was to answer three main questions regarding the EEG correlates of concurrent segregation: 1) whether the auditory figure-ground segregation elicited the electrophysiological correlates of concurrent sound segregation; 2) how different cues and combinations of cues of concurrent sound segregation affect the ORN component; 3) whether ORN and/or P400 sum together the outputs of independent cue detectors or they reflect the overall readout of the auditory system's assessment of the likelihood of the presence of multiple concurrent objects. To answer these questions, we used the stochastic figure-ground stimuli to study the effects of integrating between spectral and temporal cues of sound segregation on the ERP (*Thesis I*), while we used the more well-known mistuned harmonic paradigm to assess the effects of different cues and the congruent combinations on the ERP (*Thesis II*) and on event-related oscillations (*Thesis III*). Finally, we introduced a novel method for creating the perception of more than two concurrent sounds and studied their effects on the ERP (*Thesis IV*). In all studies, EEG was recorded. For assessing the effects of concurrent sound segregation, the event-related potentials or event-related theta oscillations were studied.

With respect to the first question we found that the ORN signals the presence of a figure, an object popping out of the background, even when it requires integrating sounds both sequentially and simultaneously. In answer to the second question, we found that different cues are not equally effective in eliciting the ORN and that these ORN responses are not additive when concurrent sound segregation cues are combined. The latter result indicates that the ORN component does not reflect the evaluation of individual sensory cues. Theta oscillations have also been modulated by the presence of the cues and therefore mediate some processes involved in the segregation of concurrent sounds. Finally, to the third question we

found that congruent combinations of two or three concurrent sound segregation cues elicits an ORN that is always smaller than the sum of the ORN amplitudes elicited by the cues separately. Perception of three versus two concurrent objects did not significantly modulate the ORN. These two results indicate that the ORN shows an overall readout of the auditory system regarding the presence of more than one objects appearing at the same time. A more detailed discussion of the results follows.

Auditory figure-ground segregation

The ERP correlate of concurrent sound segregation has mostly been studied with the mistuned harmonic paradigm (e.g., Alain et al., 2001). Recently, a novel stimulus paradigm has been introduced (Teki et al., 2011, 2013), the stochastic figure-ground stimuli which promotes integration between simultaneous and sequential sound grouping. Study I investigated the effects of the coherence (the number of repeating concurrent tones) and duration (the number of consecutive repetitions) within this paradigm, in a psychophysical and a subsequent electrophysiological experiment. In the psychophysical experiment, the optimal parameter ranges were established for the electrophysiological experiment and it was tested whether location difference between the figure and the ground would aid listeners in detecting the figure component. In accordance with Teki and colleagues' (2011) results we found that the detection of the figure increases with increasing coherence and duration values. However the location difference manipulation did not yield the expected results. We predicted that with increasing location difference (from 0 up to 90°) the spatial separation would increasingly help the figure-ground segregation. However, we found that large spatial separation interfered with the figure's detection. The lack of evidence for the parametrical effect of the location cue is in accordance with previous reports of concurrent sound segregation (e.g., Kocsis et al., 2014) suggesting that location is a weak cue of sound segregation.

Correct identification of the figure led to the elicitation of an ORN and a P400 component providing evidence that these components are also observed when concurrent sound segregation requires the integration of spectral cues over time. The amplitude of ORN increased with increasing duration and coherence values. Teki and colleagues (2016) obtained similar ORN effects with their subjects engaged in a visual task. This result confirms that the ERP effect observed in our experiment is in fact an ORN and not an awareness related negativity (ARN) or an N2 component. We also found that the P400 amplitude was correlated with detection performance. This result is compatible with the notion that the P400 either reflects the perceptual recognition of the presence of multiple concurrent auditory objects and/or the preparation for reporting the detection of multiple objects. As for the source localization of the components, both the ORN and P400 had generators in the temporal cortex, which is also in line with previous reports (e.g., Alain and McDonald, 2007; Snyder, Alain and Picton, 2006).

Effects of multiple congruent cues on concurrent sound segregation

ORN has been extensively studied with the help of the mistuned harmonic paradigm (Alain et al., 2001) with several cues of concurrent sound segregation and some of their combinations. However, no systematic research has combined several cues. Study II reports a systematic investigation of the combinations of three cues (inharmonicity, onset asynchrony, location difference) under passive and active listening situations. A similar pattern of ORN elicitation was found under both listening conditions, most cues elicited ORN alone while their combinations always elicited a significant ORN. The location cue alone, however, seemed to be a weak cue for ORN elicitation in the passive condition. Similar results were reported by Deutsch (1975) who found that discrepancy in location may be ignored when other cues favour the fusion of partials. Despite this finding, the long-standing view was that spatial separation was the most effective dimension along which a message could be located

for selection (Moray, 1969). Indeed, McDonald and Alain (2005) found that location difference alone helped the segregation of the harmonic even when the harmonic was in tune in an active listening task. Currently, there is no agreement whether or not location difference is sufficient for simultaneous sound segregation. It is possible that utilization of the location difference cue depends on the stimulus paradigm and the task used. Another possibility is that location difference is more effective when subjects are allowed to pay attention to one ear for a longer time than when they have to alternate between ears (Deutsch, 1975). In our study, in one condition the tones were randomly presented either to the left or the right ear which might explain the lack of effect of location difference.

ORN amplitude has been shown to increase with increasing cue saliency (Alain et al., 2001; Clapp, Johnson and Hautus, 2007). Study II showed that the ORN amplitude remains the same when not one, but two partials are manipulated congruently. Furthermore, we used the combinations of different cues and we found that the sum of the separate cues' amplitudes was always numerically smaller than the combined ORN amplitude. Based on these results, we claim that ORN does not reflect a cue-based response, but an overall response that is based on the most salient cue(s) and once detection based on this cue reaches a threshold, the auditory system signals the presence of multiple concurrent auditory objects (the "horse-race" mode of operation).

The harmonic sieve theory (Goldstein, 1973; Gerson and Goldstein, 1978) was introduced in connection with the complex tones and their partials, namely that harmonics that are multiple integers of the base frequency are passed through the sieve and form an in-tune complex sound, whereas when one partial is mistuned that very partial will not pass through the sieve and thus it represents a separate object. In Study II, other cues than mistuning was used as well, and it is possible that those cues can be integrated into the harmonic sieve theory so that each "hole" of the sieve incorporates not only the frequency

information regarding the certain partial but its onset, source location and possibly other features as well and if the partial cannot pass through the sieve, the partial will be regarded as a separate sound object.

The P400 component was also observed in most but not all conditions during active listening, and it was less sensitive than the ORN. In the location difference alone condition ORN was elicited, but not the P400. The lack of P400 might have been related to the fact that listeners were not able to reliably tell one from two objects in that condition. When two partials were manipulated, a positive deflection was observed in the P400 component latency that did not reach significance. This might have been due to the preceding negative components overlapping the P400.

Theta oscillations accompanying concurrent sound segregation

Concurrent sound segregation has been studied by event-related potentials in detail, but much less is known about event-related oscillations produced by brain networks underlying this process. Slow oscillatory activity has been suggested to underlie communication between functionally linked areas (Buzsáki and Draguhn, 2004). Study III describes a report aimed at investigating large scale brain oscillations accompanying concurrent sound segregation based on the data of Study II. The aim of the study was twofold: we wanted to compare the oscillatory activity across brain regions a) between concurrent sound segregation elicited by different cues (inharmonicity, onset asynchrony and location difference); and b) between passive and active situations. In Study III, we found increased theta activity in two intervals - an early and late time window - which correspond to the time windows of the ORN and P400 described in Study II. The ERSP responses were larger in the active than in the passive situation. The manipulated tones in the active situation elicited

higher theta activity than the other two conditions. Finally, different cues differently modulated the theta oscillations.

The ERSP analysis examined all EEG frequency bands. However, none but the theta band showed any significant results. The ERPs results corresponded to the effects found for the theta band oscillation. This suggests that the neural networks underlying the generation of the ERP components ORN and P400 may communicate via theta rhythms. There are some inconsistencies between the ERSP and the ERP data: the scalp distribution of the ERPs did not show any significant differences across cues, whereas the scalp distribution of the ERSPs did show cue-specific differences. Also, in the later time window of the active listening condition, some theta activity is elicited by fully harmonic sounds, whereas they did not elicit the P400 ERP response. Based on these results, it is possible that even though the ERPs and ERSPs appear to be related they are sensitive to different aspects of the neural activity or the underlying perceptual processes.

Two and three concurrent sound objects

Although a number of studies showed that humans are able to hear out a mistuned partial from and otherwise harmonic complex (e.g., Moore et al., 1986; Alain et al., 2001), no previous study have investigated whether it is possible to create the perception of three concurrent sound objects by applying concurrent sound segregation cues to a harmonic complex sound.

In Study IV, we used a modified version of the mistuned harmonic paradigm by manipulating two partials of a complex tone either in a convergent manner (i.e., that the two partials can create a single harmonic complex tone) or divergently (so that each manipulated partial should be perceived as a separate tone). The latter variant allowed listeners to perceive up to three concurrent sound objects. In a behavioural pilot experiment, we tested whether it

was possible for listeners to separately hear out two partials from a complex tone, thus creating three concurrent objects (with the remaining partials forming the third object). We found that they were reliably capable of reporting two sound objects when the manipulations were applied in the convergent manner. However, stimuli promoting the perception of three separate objects were heard as either two or three objects with equal probability. This shows that although it is possible to hear three sound objects in this paradigm, listeners cannot reliably tell apart two- vs. three-objects. Note that listeners reported significantly more instances of three concurrent sounds for the three-object sounds than for two-object sounds. Thus, although the divergent manipulation indeed promoted the perception of three concurrent objects, only this effect was not sufficient for discriminating the two vs. three concurrent objects. In the EEG study, we used the same stimuli as in the pilot, but no active response was required from the subjects. We found that ORN was elicited in almost all conditions. Sounds with AM or FM difference between the manipulated and the base harmonics yielded the smallest ORN amplitudes. This indicated that these two manipulations are also weak cues of concurrent sound segregation. No ORN was apparent when the 4th partial was mistuned by 9%. This is in line with the finding of Alain and colleagues (2001) who reported that listeners were more likely to report hearing two objects when the mistuned harmonic in the complex tone was lower in frequency than when it was higher and that the amplitudes of both the ORN and P400 was smaller for higher harmonics. In Study II, we found a significant ORN for mistuning the 4th partial by 8%. The main difference between study IV and the other two studies was that whereas in Study II, each stimulus block contained 50% manipulated and 50% base version tones, in Study IV, the probability of the base version tone was only 5% and all the different manipulations were intermixed within the stimulus blocks. This could be the reason for lower ORN amplitudes in Study IV compared to

Study II, as Bendixen and colleagues (2010) found lower amplitudes when the number of the manipulated sounds was higher.

When two partials were divergently manipulated, numerically larger ORN amplitudes were elicited as compared to the convergently manipulated tones. This is consistent with the horse-race model, namely that the most salient cue drives the perception and once the accumulated evidence reaches a threshold the auditory system registers the presence of multiple concurrent object.

5. Conclusions and further directions

The results show that the ORN event-related component reflects the emergence of an auditory object from the background in a stochastic figure-ground segregation paradigm. Segregation of simultaneous sound objects is also indexed by the ORN when single or multiple cues are used on complex tones. ORN is only followed by the P400 component in situations in which the sounds are task-relevant. ORN peaks around 150-250 ms after stimulus onset reflecting an earlier, attention-independent processing stage, whereas P400 peaks after 400 ms indexing a later, top-down controlled processing stage.

The picture emerging from studies II and IV is that ORN reflects the overall readout of the auditory system regarding the presence of multiple concurrent objects. P400 could then represent the outcome of the perceptual decision. Thus, whereas ORN reflects a bottom-up primitive grouping mechanism, P400 likely reflects a process which incorporates top-down effects on perceptual decisions.

The properties of ORN and P400 are compatible with Bregman's two-stage theoretical framework. In the first stage of processing, in which the acoustic input is decomposed into putative perceptual groupings (i.e., proto-objects are formed) ORN reflects an on/off flag signalling whether one or more sound objects are likely to be present. If no cue or cue combination reaches a certain threshold, the incoming signal is represented as a single sound object. Once a threshold is reached (regardless of the number of congruent cues, using the horse-race principle) the presence of more than one object is signalled, and competition between the alternatives commences. This competition can be biased by attention, thus choosing which alternative is consciously perceived.

The processes of concurrent sound segregation probably precede those of sequential sound segregation (see Winkler and Schröger's (2015) model of auditory event formation).

This allows sequential grouping processes to integrate the outcome of concurrent sound segregation thus forming proto-objects from a subset of the simultaneously encountered sounds.

However, the current results are also compatible with the temporal coherence model of auditory scene analysis. Within this framework, the ORN would appear as a temporal coherence detector (described in *Section 1.4*). Teki and colleagues (2013) suggest that in the first stage of processing, feature analysis is done, whereas in the second stage, the output of the previous stage is grouped according to temporal modulation. Since ORN reflects the aggregate readout of cues, it can be seen as temporal coherence detector. Teki and colleagues (2013) also suggest that in addition to the first two stages, attention plays a key role in the formation of the streams as attention biases the auditory system toward a particular grouping of sound source attributes that depend on the listener's current behavioural and perceptual goals. In this light, P400 reflects the outcome of the biasing process: the detection of the target features. It is important to note that there are neurophysiological signals sensitive to temporal coherence (O'Sullivan et al., 2015). Thus it is likely that temporal coherence is indeed computed in the human brain. However, this does not decide between Bregman's theoretical framework and temporal coherence models, because temporal coherence based grouping could occur during the first phase of Bregman's auditory scene analysis (i.e., it would be regarded as the heuristic implementation of the Gestalt principle of "common fate").

Deciding between Bregman's framework and the temporal coherence model requires further research. Insights could be gained from stream segregation experiments using stimuli that can be grouped by mechanisms outside temporal coherence, such as, inserting predictable patterns separately into the putative streams (e.g., a familiar melody within an unfamiliar background; cf. Szalárdy and colleagues, 2014), salient or unexpected stimuli that are inconsistent with the listening situations (a dog bark in the bathroom), or presenting speech

streams in which the semantic congruity can also help segregation. Another possible piece of evidence could come from finding representations of alternative (non-dominant) sound organizations, as these are not computed by temporal coherence models. The existence of neuronal populations encoding sound objects that are currently not perceived would support Bregman's model of auditory scene analysis.

Other open issues are listed below. First, not much is known about the neural substrates of concurrent sound segregation. fMRI measures could shed further light on the brain regions involved in these processes. Second, the stochastic figure-ground stimuli bring the research closer to more realistic sound environments, but there is still space to improve: concurrent sound segregation could be studied by the use of streams of natural sounds (not random noise background); for example, environmental sounds containing a meaningful target. Third, so far the paradigms used have been employing mostly short sounds, future studies should be conducted to investigate the differences if continuous sounds were used instead, and the mistuning happened during listening and not only in a very short time window. Finally, integrating concurrent sound segregation into segregating speech streams would result in better understanding of how our auditory system copes with the cocktail party problem.

References

Akram, S., Englitz, B., Elhiali, M., Simon, J. Z., & Shamma, S. A. (2014). Investigating the neural correlates of a streaming percept in an informational masking paradigm. *PLoS ONE,* 9(12): e114427.

Alain, C. (2007). Breaking the wave: Effects of attention and learning on concurrent sound perception. *Hearing Research*, 229, 225-236.

Alain, C., & Arnott, S. R. (2000). Selectively attending to auditory objects. *Frontiers in Bioscience,* 5, 202-212.

Alain, C., Arnott, S. R., & Picton, T. W. (2001). Bottom-up and top-down influences on auditory scene analysis: Evidence from brain potentials. *Journal of Experimental Psychology,* 27, 1072-1089.

Alain, C., & Izenberg, A. (2003). Effects of attentional load on auditory scene analysis. *Journal of Cognitive Neuroscience*, 15, 1063-1073.

Alain, C., & McDonald, K. L. (2007). Age-related differences in neuromagnetic brain activity underlying concurrent sound perception. *Journal of Neuroscience*, 27, 1308-1314.

Alain, C., Schuler, B. M., & McDonald, K. L. (2002). Neural activity associated with distinguishing concurrent auditory objects. *Journal of the Acoustical Society of America,* 111(2), 990-995.

Alain, C., Reinke, K. S., He, Y., Wang, C., & Lobaugh, N. (2005). Hearing two things at once: Neurophysiological indices of speech segregation and identification. *Journal of Cognitive Neuroscience,* 17, 811-818.

Alain, C., Theunissen, E. L., Chevalier, H., Batty, M., & Taylor, M. J. (2003). Developmental changes in distinguishing concurrent auditory objects. *Cognitive Brain Research,* 16, 210-218.

Arnott, S. A., Bardouille, T., Ross, B., & Alain, C. (2011). Neural generators underlying concurrent sound segregation. *Brain Research*, 1387, 116-124.

Barascud, N., Pearce, M. T., Griffiths, T. D., Friston, K. J., & Chait, M. (2016). Brain responses in humans reveal ideal observer-like sensitivity to complex acoustic patterns. *Proceeding of the National Academy of Sciences .U.S.A.* 113, E616–E625.

Basar, E., Basar-Eroglu, C., Karakas, S., & Schürmann, M. (1999). Oscillatory brain theory: A new trend in neuroscience. *IEEE Engineering in Medicine and Biology Magazine,* 18(3), 56-66.

Békésy, G. von. (1963). Three experiments concerned with speech perception. *Journal of the Acoustical Society of America*, 35, 602-606.

Bendixen, A., Jones, S. J., Klump, G., & Winkler, I. (2010). Probability dependence and functional separation of the object-related and mismatch negativity event-related potential components. *Neuroimage*, 50, 285-290.

Bendixen, A., Háden, G. P., Németh, R., Farkas, D., Török, M., & Winkler, I. (2015). Newborn infants detect cues of concurrent sound segregation. *Developmental Neuroscience,* 37(2), 172-181.

Berger, H. (1929). Über das Elektroenkephalogramm des Menschen. *Archiv für Psychiatrie und Nervenkrankenheiten,* 87, 527-570.

Bertrand, O., & Tallon-Baudry, C. (2000). Oscillatory gamma activity in humans: a possible role for object representation. *International Journal of Psychophysiology*, 38, 211-223.

Bidet-Caulet, A., & Bertrand, O. (2009). Neurophysiological mechanisms involved in auditory perceptual organization. *Frontiers in Neuroscience*, 3, 182-191.

Bizley, J. K., & Cohen, Y. E. (2013). The what, where and how of auditory object perception. *Nature Reviews Neuroscience*, 14, 693-707.

Bizley, J. K., Walker, K. M., Silverman, B. W., King, A. J., & Schnupp, J. W. (2009). Interdependent encoding of pitch, timbre, and spatial location in auditory cortex. *Journal of Neuroscience*, 29, 2064–2075.

Bőhm, T. M., Shestopalova, L., Bendixen, A., Andreou, A. G., Georgiou, J., Garreau, G., Poliquen, P., Cassidy, A., Denham, S. L., & Winkler, I. (2013). The role of perceived source location in auditory stream segregation: Separation affects sound organization, common fate does not. *Learning & Perception,* 5, Supplement 2, 55-72

Bregman, A. S. (1990). *Auditory scene analysis: The perceptual organization of sound.* Cambridge, MA: The MIT Press.

Bregman, A. S., & Pinker, S. (1978). Auditory streaming and the building of timbre. *Canadian Journal of Psychology*, 32(1), 19-31.

Brunswik, E. (1955). Representative Design and Probabilistic Theory in a Functional Psychology. *Psychological Review*, 62(3), 193–217.

Buell, T. N., & Hafter, E. R. (1991). Combination of binaural information across frequency bands. *Journal of the Acoustical Society of America*, 90(4), 1894-1900.

Buzsáki, G., & Draguhn, A. (2004). Neuronal oscillations in cortical networks. *Science,* 304, 1926-1929.

Carlyon, R. P. (2004). How the brain separates sounds. *Trends in Cognitive Science*, 8(10), 465-471.

Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with two ears. *Journal of the Acoustical Society of America*, 25(5), 975-979.

Ciocca, V. (2008). The auditory organization of complex sounds. *Frontiers in Bioscience*, 13, 148-169.

Clapp, W. C., Johnson, B. W., & Hautus, M. J. (2007). Graded cue information in dichotic pitch: Effects on event-related potentials. *Neuroreport*, 18, 365-368.

Culling, J. F., & Summerfield, Q. (1995). Effects of contralateral presentation and of interaural time differences in segregating a harmonic from a vowel. *Journal of the Acoustical Society of America,* 98, 1380-1387.

Darwin, C. J., Ciocca, V., & Sandell, G. J. (1994). Effects of frequency and amplitude modulation on the pitch of a complex tone with a mistuned harmonic. *Journal of the Acoustical Society of America,* 95, 2631-2636.

Darwin, C. J., & Hukin, R. W. (1999). Auditory objects of attention: the role of interaural time differences. *Journal of Experimental Psychology: Human Perception and Performance,* 25(3), 617-629.

Darwin, C. J., & Sandell, G. J. (1995). Absence of effect of coherent frequency modulation on grouping a mistuned harmonic with a vowel. *Journal of the Acoustical Society of America,* 97, 3135-3138.

Davis, H. (1976). Principles of electric response audiometry. *Annals of Otology, Rhinology and Laryngology,* 85(Suppl. 28), 1-96.

Denham, S. L., & Winkler, I. (2015). Auditory perceptual organization. In Wagemans, J. ed.: *The Oxford Handbook of Perceptual Organization*, Oxford University Press.

Deutsch, D. (1975). Two-channel listening to musical scales. *Journal of the Acoustical Society of America,* 57, 1156-1160.

Deutsch, D. (1979). Binaural integration of melodic patterns. *Perception and Psychophysics,* 25, 399-405.

Duifhuis, H., Willems, L. F., & Sluyter, R. J. (1982). Measurement of pitch in speech: An implementation of Goldstein's theory of pitch perception. *Journal of the Acoustical Society of America,* 71, 1568-1580.

Duncan, J., & Humphreys, G. W. (1989). Visual search and stimulus similarity. *Psychological Review*, 96(3), 433-458.

Duncan, J., & Humphreys, G. W. (1992). Beyond the search surface: visual search and attentional engagement. *Journal of Experimental Psychology: Human Perception and Performance,* 18(2), 578-588.

Elhiali, M., Xiang, 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.

Fell, J., Hinricks, H., & Röschke, J. (1997). Time course of human 40 Hz EEG activity accompanying P3 responses in an auditory oddball task. *Neuroscience Letters*, 235(3), 121-124.

Folstein, J. R., & Van Patten, C. (2008). Influence of cognitive control and mismatch on the N2 component of the ERP: A review. *Psychophysiology,* 45(1), 152–170.

Galambos, R. (1992). A comparison of certain gamma band (40 Hz) brain rhythms in cat and man. In: Basar, E., Bullock, T. H. (Eds.), *Induced Rhythms in the Brain*. Birkhauser.

Gerson, A., & Goldstein, J. L. (1978). Evidence for a general template in central optima processing for pitch of complex tones. *Journal of the Acoustical Society of America,* 63, 498-510.

Godey, B., Schwartz, D., de Graaf, J. B., Chauvel, P., Liegeois-Chauvel, C. (2001). Neuromagnetic source localization of auditory evoked fields and intracerebral evoked potentials: a comparison of data in the same patients. *Clinical Neurophysiology*, 112, 1850–1859.

Goldstein, J. L. (1973). An optimum processor theory for the central formation of the pitch of complex tones. *Journal of the Acoustical Society of America,* 54, 1496-1516.

Gordon, I. E. (2004). *Theories of Visual Perception*, 3rd edition. Psychology Press: Hove and New York.

Griffiths, T. D., & Warren, J. D. (2004). What is an auditory object? *Nature Reviews Neuroscience*, 5, 887-892.

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, 5382-5388

Halberda, J., Simons, D.J., & Wetherhold, J. (submitted). Overcoming the three-item limit: Gestalt grouping principles explain increases in change detection capacity.

Handel, S. (1988). Space is to time as vision is to audition: Seductive but misleading. *Journal of Experimental Psychology: Human Perception and Performance,* 14(2), 315-317.

Hartmann, W. M., McAdams, S., & Smith, B. K. (1990). Hearing a mistuned harmonic in an otherwise periodic complex tone. *Journal of the Acoustical Society of America*, 88, 1712-1724.

Hautus, M. J., & Johnson, B. W. (2005). Object-related brain potentials associated with the perceptual segregation of a dichotically embedded pitch. *Journal of the Acoustical Society of America,* 117, 275-280.

Hautus, M. J., Johnson, B. W., & Colling, L. J. (2009). Event-related potentials for interaural time differences and spectral cues. *Neuroreport*, 20, 951-956.

Haykin, S., & Chen, Z. (2005). The cocktail party problem. *Neural Computation*, 17(9), 1875-1902.

Hillyard, S. A., Squires, K. C., Bauer, J. W., & Lindsay, P. H. (1971). Evoked potential correlates of auditory signal detection. *Science*, 172(3990), 1357-1360.

Hiraumi, H., Nagamine, T., Morita, T., Naito, Y., Fukuyama, H., & Ito, J. (2005). Right hemispheric predominance in the segregation of mistuned partials. *European Journal of Neuroscience,* 22, 1821-1824.

Huotilainen**,** M**.,** Winkler**,** I**.,** Alho**,** K**.,** Escera**,** C**.**, Virtanen**,** J**.**, Ilmoniemi**,** R**.** J**.**, Jääskeläinen**,** I**.** P**.**, Pekkonen**,** E**.**, & Näätänen**,** R.(1998). Combined mapping of human auditory EEG and MEG responses. *Electroencephalography and Clinical Neurophysiology,* 108(4)**,** 370-379.

Jack, C. E., & Thurlow, W. R. (1973). Effects of degree of visual association and angle of displacement on the "ventriloquism" effect. *Perceptual and Motor Skills,* 37(3), 967-979.

Johnson, B. W., & Hautus, M. (2010). Processing of binaural information in human auditory cortex: Neuromagnetic responses to interaural timing and level differences. *Neuropsychologia*, 48, 2610-2619.

Johnson, B. W., Hautus, M., & Clapp, W. C. (2003). Neural activity associated with binaural processes for the perceptual segregation of pitch. *Clinical Neurophysiology*, 114, 2245-2250.

Johnson, B. W., Hautus, M. J., Duff, D. J., & Clapp, W. C. (2007). Sequential processing of interaural timing differences for sound source segregation and spatial localization: Evidence from event-related cortical potentials. *Psychophysiology*, 44, 541-551.

Julesz, B. (1971). *Foundations of Cyclopean Perception.* Chicago: The University of Chicago Press.

Kaiser, J., Lutzenberger, W., Preissl, H., Ackermann, H., & Birnbaumer, N. (2000). Right hemisphere dominance for the processing of sound source lateralization. *The Journal of Neuroscience,* 20(17), 6631-6639.

Kidd, G. Jr., Mason, C. R., Deliwala, P. S., Woods, W. S., & Colburn, H. S. (1994). Reducing informational masking by sound segregation. *Journal of the Acoustical Society of America,* 95, 3475–3480.

Klimesch, W., Sauseng, P., & Hanslmayr, S. (2007). EEG alpha oscillations: the inhibition-timing hypothesis. *Brain Research Reviews,* 53, 63–88.

Kocsis, Z., Winkler, I., Bendixen, A., & Alain, C. (2016). Promoting the perception of two and three concurrent sound objects: an event-related potential study. *International Journal of Psychophysiology*, 107, 16-28.

Kocsis, Z., Winkler, I., Szalárdy, O., & Bendixen, A. (2014). Effects of cue redundancy on concurrent sound segregation during passive and active listening: An event-related potential (ERP) study. *Biological Psychology*, 100, 20-33.

Koffka, K. (1935). *Principles of Gestalt psychology.* New York: Harcourt Brace.

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

Kraus, N., & Nicol, T. (2009). Auditory evoked potentials. In Binder, M. D., Hirokawa N., & Windhorst U., (eds.), *Encyclopedia of Neuroscience*. Springer Science+Business Media, Berlin, Germany.

Krishnan, L, Elhiali, M., & Shamma, S. A. (2014). Segregating complex sound sources through temporal coherence. *PLoS Computational Biology*, 10(12), e1003985.

Kubovy, M. (1981). Concurrent-pitch segregation and the theory of indispensable attributes. In Kubovy, M., & Pomerantz, J. R. (eds.), *Perceptual Organization.* Hillsdale, N.J.: Erlbaum.

Kubovy, M., Cutting, J. E., & McGuire, M. R. (1974). Hearing with the third ear: Dichotic perception of a melody without monaural familiarity cues. *Science*, 186, 272-274.

Kubovy, M., & Van Valkenburg, D. (2001). Auditory and visual objects. *Cognition*, 80, 97-126.

Kumar, S., Bonnici, H. M., Teki, S., Agus, T. R., Pressnitzer, D., Maguire, A., & Griffiths, T. D. (2014). Representations of specific acoustic patterns in the auditory cortex and hippocampus. *Proceeding of the Royal Society B Biological Sciences,* 281:20141000.

Lehmann, D., & Julesz, B. (1978). Lateralized cortical potentials evoked in humans by dynamic random-dot stereograms. *Vision Research,* 18, 1265-1271.

Lerea, L. (1961). An investigation of auditory figure-ground perception. *The Journal of Genetic Psychology,* 98, 229-237.

Lipp, R., Kitterick, P., Summerfield, Q., Bailey, P. J., & Paul-Jordanov, I. (2010). Concurrent sound segregation based on inharmonicity and onset asynchrony. *Neuropsychologia*, 48, 1417-1425.

Lu, K., Xu, Y., Yin, P., Oxenham, A. J., Fritz, J. B., & Shamma, S. A. (2017). Temporal coherence structure rapidly shapes neuronal interactions. *Nature Communications*, 8, 13900.

Luck, S. J. (2005). *An introduction to the 3event-related potential technique*. MIT Press; Cambridge, MA.

Mäkelä, J. P., Hari, R., & Leinonen, L. (1988). Magnetic responses of the human auditory cortex to noise/square wave transitions. *Electroencephalography and Clinical Neurophysiology,* 69, 423-430.

Marshall, L., Mölle, M., & Bartsch, P. (1996). Event-related gamma band activity during passive and active oddball tasks. *NeuroReport*, 7, 1517-1520.

McAdams, S. (1984). *Spectral fusion, spectral parsing and the formation of auditory images.* PhD thesis, Stanford University.

McDonald, K. L., & Alain, C. (2005). Contribution of harmonicity and location to auditory object formation in free field: Evidence from event-related brain potentials. *Journal of the Acoustical Society of America,* 118, 1593-1604.

Micheyl, C., Shamma, S., & Oxenham, A. J. (2007). Hearing out repeating elements in randomly varying multitone sequences: a case of streaming? In Kollmeier, B., Klump, G., Hohmann, V., Langemann, U., Mauermann, M., Uppenkamp, S., & Verhey, J., (eds.) *Hearing - from basic research to application*. Berlin: Springer.

Micheyl, C., Kreft, H., Shamma, S. A., & Oxenham, A. J. (2013). Temporal coherence versus harmonicity in auditory stream formation. *Journal of the Acoustical Society of America,* 133(3), 188-194.

Miller, L. M., & Recanzone, G. H. (2009). Populations of auditory cortical neurons can accurately encode acoustic space across stimulus intensity. *Proceedings of the National Academy of Sciences,* 106, 5931-5935.

Moore, B. C. J. (1982). *An Introduction to the Psychology of Hearing. (Second edition)* London: Academic Press.

Moore, B. C. J., Glasberg, B. R., & Peters, R. W. (1985). Relative dominance of individual partials in determining the pitch of complex tones. *Journal of the Acoustical Society of America,* 77, 1853-1860.

Moore, B. C. J., Glasberg, B. R., & Peters, R. W. (1986). Thresholds for hearing mistuned partials as separate tones in harmonic complexes. *Journal of the Acoustical Society of America,* 80(2), 479-483.

Moore, B. C. J., & Gockel, H. (2002) Factors influencing sequential stream segregation. *Acta Acustica united with Acustica,* 88(3), 320-333.

Moray, N. (1969). *Listening and attention*. Penguin Books; Harmondsworth.

Näätänen, R., Gaillard, A. W. K., & Mäntysalo, S. (1978). Early selective attention effect on evoked potential reinterpreted. *Acta Psychologica*, 42, 313–329.

Näätänen, R., & Picton, T. W. (1987). The N1 wave of the human electric and magnetic response to sound: a review and an analysis of the component structure. *Psychophysiology*, 24, 375-425.

Näätänen, R., & Winkler, I. (1999). The concept of auditory stimulus representation in cognitive neuroscience. *Psychological Bulletin,* 125, 826-859

O'Callaghan, C. (2008). Object perception: Vision and audition. *Philosophy Compass,* 3/4, 803-829.

O'Sullivan, J. A., Shamma, S. A., & Lalor, E. C. (2015). Evidence for neural computations of temporal coherence in an auditory scene and their enhancement during active listening. *Journal of Neuroscience,* 35(18), 7256-7263.

Pastore, R. E., Harris, L. B., & Kaplan, J. K. (1982). Temporal order identification: Some parameter dependencies. *Journal of the Acoustical Society of America,* 71(2), 430-436.

Perrott, D. R., & Barry, S. H. (1969). Binaural fusion. *The Journal of Auditory Research*, 3, 263-269.

Peterson, D. J. & Berryhill, M. E. (2013). The Gestalt Principle of Similarity Benefits Visual Working Memory. *Psychonomic Bulletin & Review,* 20(6), 1282–1289.

Polich, J. (2007). Updating P300: an integrative theory of P3a and P3b. *Clinical Neurophysiology,* 118, 2128–2148.

Rauschecker, J. P. (1997). Processing of complex sounds in the auditory cortex of cat, monkey, and man. *Acta Otolaryngolica Suppl.,* 532, 34–38.

Reite, M., Teale, P., Zimmermann, J., Davis, K., & Whalen, J. (1988). Source location of a 50 msec latency auditory evoked field component. *Clinical Neurophysiology*, 70(6), 490-498.

Roberts, B., & Bailey, P. J. (1993). Spectral pattern and the perceptual fusion of harmonics. I. The role of temporal factors. *Journal of the Acoustical Society of America,* 94, 3153-3164.

Roberts, B., & Bregman, A. S. (1991). Effects of the pattern of spectral spacing on the perceptual fusion of harmonics. *Journal of the Acoustical Society of America,* 90, 3050-3060.

Roberts, B., & Brunstrom, J. M. (1998). Perceptual segregation and pitch shifts of mistuned components in harmonic complexes and in regular inharmonic complexes. *Journal of the Acoustical Society of America,* 104(4), 2326-2338.

Roberts, B., & Brunstrom, J. M. (2001). Perceptual fusion and fragmentation of complex tones made inharmonic by applying different degrees of frequency shift and spectral stretch. *Journal of the Acoustical Society of America,* 110(5), 2479-2490.

Rubin, E. (1915). *Synsoplevede Figurer: Studier i psykologisk Analyse. Første Del* [Visually experienced figures: Studies in psychological analysis. Part one]. Copenhagen and Christiania: Gyldendalske Boghandel, Nordisk Forlag.

Sanders, L. D., Joh, A. S., Keen, R. E., & Freyman, R. L. (2008). One sound or two? Object-related negativity indexes echo perception. *Perceptual Psychophysiology,* 70, 1558-1570.

Sanders, L. D., Zobel, B. H., Freyman, R. L., & Keen, R. (2008). Manipulations of listeners' echo perception are reflected in event-related potentials. *Journal of the Acoustical Society of America,* 129, 301-309.

Shamma, S. A., & Micheyl, C. (2010). Behind the scenes of auditory perception. *Current Opinion in Neurobiology,* 20, 361-366.

Shamma, S. A., Elhiali, M., & Micheyl, C. (2011). Temporal coherence and attention in auditory scene analysis. *Trends in Neuroscience*, 34(3), 114–123.

Simson, R., Vaughan, H. G., & Ritter, W. (1977). The scalp topography of potentials associated with missing visual or auditory stimuli. *Electroencephalography and Clinical Neuropsychology,* 40, 33-42.

Snyder, J. S., & Alain, C. (2005). Age-related changes in neural activity associated with concurrent vowel segregation. *Cognitive Brain Research*, 24, 492-499.

Snyder, J. S., & Alain, C. (2007). Towards a neuropsychological theory of auditory stream segregation. *Psychological Bulletin,* 133(5), 780-799.

Snyder, J. S., Alain, C., Picton, T. W, (2006). Effects of attention on neuroelectric correlates of auditory stream segregation. *Journal of Cognitive Neuroscience*, 18, 1-13

Szalárdy, O., Bendixen, A., Bőhm, T. M., Davies, L. A., Denham, S. L., & Winkler, I. (2014). The effects of rhythm and melody on auditory stream segregation. *Journal of the Acoustical Society of America,* 135(3), 1392-1405.

Teki, S., Chait, M., Kumar, S., von Kriegstein, K., & Griffiths, T. D. (2011). Brain bases for auditory stimulus-driven figure-ground segregation. *The Journal of Neuroscience,* 31, 164-171.

Teki, S., Chait, M., Kumar, S., Shamma, S., & Griffiths, T. D. (2013). Segregation of complex acoustic scenes based on temporal coherence. *eLife,* 2: e00699.

Teki, S., Barascud, N., Picard, S., Payne, C., Griffiths, T. D., & Chait, M. (2016). Neural correlates of auditory figure-ground segregation based on temporal coherence. *Cerebral Cortex,* 26(9), 1-12.

Titchener, E.B. (1901). *Experimental psychology: A manual of laboratory practice.* New York: Macmillan.

Tóth, B., Kocsis, Z., Háden, G.P., Szerafin, Á., Shinn-Cunningham, B., & Winkler, I. (2016). EEG signatures accompanying auditory figure-ground segregation. *Neuroimage*, 141, 108-119.

Tóth, B., Kocsis, Z., Urbán, G., & Winkler, I. (2016). Theta oscillations accompanying concurrent auditory stream segregation. *International Journal of Psychophysiology*, 106, 141-151.

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology* 12, 97-136.

Treisman, A. M. (1998). Feature binding, attention and object perception. *Philosophical Transaction of the Royal Society London B*, 353, 1295-1306.

Tremblay, K. L., & Kraus, N. (2002). Auditory training induces asymmetrical changes in cortical neural activity. *Journal of Speech, Language and Hearing Research,* 45, 564-572.

Van Noorden, L. P. A. S. (1975). *Temporal coherence in the perception of tone sequences*. Doctoral dissertation, Eindhoven University of Technology, Eindhoven, The Netherlands.

Warren, D. H., Welch, R. B., & McCarthy, T. J. *(1981)*. The role of visual-auditory "compellingness" in the ventriloquism effect: implications for transitivity among the spatial senses. *Perceptual Psychophysiology,* 30*,* 557-564.

Weise, A., Schröger, E., & Bendixen, A. (2012). The processing of concurrent sounds based on inharmonicity and asynchronous onsets: An object-related negativity (ORN) study. *Brain Research*, 1439, 73-81.

Weisz, N., Hartmann, T., Müller, N., Lorenz, I., & Obleser, J. (2011). Alpha rhythms in audition: cognitive and clinical perspectives. *Frontiers in Psychology,* 2:73.

Winkler, I. (2007). Interpreting the mismatch negativity. *Journal of Psychophysiology,* 21(3–4), 147–163.

Winkler, I., Denham, S. L., & Nelken, I. (2009). Modeling the auditory scene: Predictive regularity representations and perceptual objects. *Trends In Cognitive Sciences,* 13(12), 532-540.

Winkler, I., Denham, S. L., & Escera, C. (2013). *Auditory event-related potentials*. In *Encyclopedia of Computational Neuroscience.* SpringerReference. Springer-Verlag Berlin Heidelberg.

Winkler, I., & Schröger, E. (2015). Auditory perceptual objects as generative models: Setting the stage for communication by sound. *Brain and Language*, 148, 1-22.

Wightman, F. L., & Jenison, R. (1995). Auditory spatial layout. In Epstein, W., & Rogers, S.J. (eds.), *Perception of space and motion*. San Diego, CA: Academic Press.

Woods, D. L. (1995). The component structure of the N1 wave of the human auditory evoked potential. *Electroencephalography and Clinical Neurophysiology Supplement* 44:102-9.

Yordanova, J., Kolev, V., & Polich, J. (2001). P300 and alpha event-related desynchronization (ERD). *Psychophysiology*, 38, 143-152.

Yost, W. A. (1991). Auditory image perception and analysis: The basis for hearing. *Hearing Research,* 56, 8-18.