PROCEEDINGS of the
23rd International Congress on Acoustics

9 to 13 September 2019 in Aachen, Germany

Foreground-background decomposition in complex auditory scenes

Sabine THOMASSEN[1]; Alexandra BENDIXEN[1]

[1] Chemnitz University of Technology, Germany

ABSTRACT

Tremendous insight into auditory scene analysis has been gained by using two-tone sequences in auditory streaming paradigms as a simplified model scenario. Since natural acoustic scenes usually contain more than two potential sound sources, recent efforts strive towards studying scene analysis with more complex source configurations. Our current work along these lines focuses on the question of how perceptual foreground and background are decomposed. In a subjective-reporting procedure, we show that listeners spontaneously report hearing more than one sound source in the perceptual foreground at once. The perceptual background is addressed by combining behavioral measures with electroencephalography (EEG). Our results suggest that integrated and segregated representations of the perceptual background are held in parallel. We discuss these findings in the light of current theories and outline some methodological challenges. To face one of these challenges, namely the considerable effort in training and instructions in subjective-reporting procedures, we merged a visual multistability procedure with an auditory multistable stimulus and collected eye-tracking data alongside the auditory reports. Results show a good agreement between reported auditory perception and visual foreground formation as determined via eye movements. We will discuss whether eye-tracking is an adequate *en passant* method for continuous measurements in auditory perception.

Keywords: auditory scene analysis, auditory stream segregation, electroencephalography (EEG)

1. INTRODUCTION

In everyday life, we continuously analyze our auditory environment to extract the sounds belonging to a sound source of interest. These sounds need to be segregated from the background noise and integrated into a coherent perceptual foreground. Various cues such as spatial location and frequency of individual sounds support this decomposition process and have been studied extensively with the help of alternating two-tone sequences (1,2). These sequences consist of repeatedly presented tones in, for instance, an 'AB' pattern, where A denotes a tone of a given feature and B denotes a tone that differs from A in the respective feature. Depending on the presentation rate and the amount of feature separation between A and B, they can either be perceived as integrated into one coherent stream or as segregated into two streams (1,2). The former alternative represents the auditory system's interpretation that A and B were emitted by the same sound source, whereas the latter case represents separate sound sources (1).

Perceptual organization alternatives in two-tone sequences are usually measured via subjective-reporting procedures, where participants indicate their perception continuously (3) or at the end of a sequence (4). Their perception can switch back and forth between various alternatives while listening to the sequence, a phenomenon called bi- or multistability (1,3,5,6). The beauty of multistability paradigms is that they enable investigations on perception without changing the physical input.

With two-tone sequences, the number of perceptual organization alternatives is quite limited: the sounds are either integrated into one stream or segregated into two streams, one of which can be selected as the perceptual foreground (2). There are additional organization alternatives (6,7), but none of them implies more than two auditory sources. A natural environment, however, mostly consists of more than two auditory sources. Investigations on rather complex auditory scenes must handle a very high number of perceptual alternatives. The addition of just one more tone to the two-tone sequence (i.e., 'ABC') leads to a substantial increase of the number of perceptual organization alternatives. One advantage of these three-tone sequences is that not only foreground but also background formation can

[1] sabine.thomassen@physik.tu-chemnitz.de

be investigated, since there are perceptual organization alternatives with more than one tone in the background. We will present our recent findings on the foreground (8) and background (9) formation in these complex auditory scenes and discuss our approach for an *en passant* method for continuous measurements in auditory streaming paradigms (10).

2. EMPIRICAL WORK

2.1 Foreground formation in complex auditory scenes

In a study on auditory multistability in complex scenes (8), we investigated whether the basic regularities of sound organization derived from two-tone experiments (1) can be extended to a three-tone arrangement ('ABC'), where perception is confronted with many more organization alternatives. We expected increasing frequency and spatial separation to cause an increase of segregated percepts and a decrease of the integrated percept, as known from previous two-tone studies (1,3).

Participants listened to sound sequences of repeatedly presented 'ABC' tone patterns (with A as a low-frequency sinusoidal tone, B as a middle-frequency tone and C as a high-frequency tone). The sequences varied in frequency separation (Δf: 2, 4, 6, or 8) between A, B and C, as well as in the presence or absence of an interaural time difference (ITD). The ITD was a 500 μs onset delay to the right ear during A tone presentation and the same delay to the left ear during C tone presentation, creating the impression of the A tones coming from the left and the C tones coming from the right. Participants listened to the sequence for 5 minutes and continuously reported their subjective perception. They chose one out of 12 pictograms, each of which represented an auditory organization alternative, and switched to another pictogram whenever their perception changed.

Participants were able to handle the procedure, and, in particular, the high number of report alternatives, as evidenced by their reports being systematic and plausible. Specifically, their reports were in line with findings on two-tone sequences: Spatial as well as frequency separation raised the proportion of stream segregation and reduced the proportion of the integrated percept (1,3). In addition, we found that tones of intermediate feature values (the B tones) were rarely perceived in the foreground, which is in line with findings from objective measurements showing difficulties in focusing on inner subsequences (11). Surprisingly, it was quite common for participants to report perceptual organization alternatives where two streams were in the foreground at once, e.g. A tones and C tones segregated into two streams but both equally in the foreground. One of the most often reported organization alternatives was one with two streams in the foreground, namely the percept with B and C tones integrated into one auditory stream and the A tones segregated in another stream, but both streams being perceived in the foreground. While studies on task-based procedures indicate that it is possible but difficult to pay attention to more than one stream (12,13), our results show that there are circumstances or tasks where the auditory system tracks multiple foreground streams in parallel without being forced to do so.

2.2 Background formation in complex auditory scenes

In a second study (9), we addressed the processing of sounds in the background of complex auditory scenes after one stream has been established as the foreground. The question was whether the sounds in the background would be further disentangled into separate streams or not. As already mentioned, there are at least three different tones needed to address this question, since there must be two tones (creating two subsequences) in the background after one tone has formed the foreground. We combined a task-based procedure for foreground formation with an event-related brain potential (ERP) approach to infer on background processing. In detail, tone presentations in the two background subsequences followed specified timing and location regularities that were occasionally violated (deviant). We expected these deviants to elicit distinct MMN (mismatch negativity) patterns depending on whether the background subsequences were processed as segregated into two streams or as integrated into one stream. The MMN is a component in the auditory ERP that is elicited by rare deviations in an otherwise regular sequence of sounds (14). Since MMN elicitation does not require attention, it is a common tool for accessing auditory background processing (15,16).

The tone sequences consisted of three sinusoidal tones, with about 4.5 semitones between two successive tones, repeatedly presented in an ascending order. Participants performed an intensity deviance detection task on one of the tones (A, B, or C), while the two remaining tones contained randomly varying intensities. This task requires perceptual separation of the task-relevant

(foreground) and the task-irrelevant (background) subsequences in order to avoid false alarms based on high-intensity tones from the task-irrelevant subsequences.

To address processing of the background subsequences, two additional parameters were systematically manipulated: location and timing. Locations (ITD: 500 μs onset delay to the right or left ear) in the two background subsequences changed randomly between left and right. In order to establish regularities, four identical locations in a row were introduced occasionally into one of the two background subsequences, while locations in the other background subsequence kept randomly changing. The fifth tone presentation within the location-regular subsequence contained a location change (deviant). These location deviants were expected to elicit an MMN only if the background subsequences were processed as segregated into two streams because, in an integrated stream, the location regularity in one subsequence would be hidden by the randomly varying locations in the other subsequence. This approach is a useful tool for indirect measures of segregation, although it is mostly implemented with intensity deviations (15,17).

The second background parameter was timing. Timing deviants were tones with a 50-ms earlier onset than standards, creating a reduced SOA of 100 ms instead of 150 ms between the previous tone and the deviant tone (for compensation, the following SOA was prolonged by 50 ms). The preceding tone of a timing deviant was either a task-irrelevant tone or a task-relevant tone. Since foreground and background tones were segregated, the actual background-tone SOA in the latter case was twice the standard SOA (300ms). Hence, there were long-SOA timing deviants (250 ms) for tones following a task-relevant tone and short-SOA timing deviants (100 ms) for tones following task-irrelevant tones. Only short-SOA timing deviants were expected to elicit an MMN if the task-irrelevant subsequences were processed as integrated into one coherent stream. This assumption is based on two findings. First, it is known that temporal and order relations are easier to identify within a stream than between streams (18,19). In line with this, there should be no MMN following timing deviants, neither short- nor long-SOA, if the task-irrelevant subsequences were processed as two segregated streams. Second, an MMN following timing deviants or complete stimulus omissions is elicited only when the stimulus onset asynchrony (SOA) is shorter than 150 to 200 ms (20,21). According to this, there should be no MMN following long-SOA timing deviants but only short-SOA timing deviants if the task-irrelevant subsequences were processed as one coherent stream.

Two hypotheses were contrasted: If the tones in the background were processed as two separate streams, we expected no MMN to either of the timing deviants, but an MMN to location deviants. On the contrary, if the tones in the background were processed as one coherent stream, we expected an MMN following short-SOA timing deviants, but no MMN to long-SOA timing deviants and no MMN to location deviants.

Participants' performance in the intensity deviation task indicated that a foreground stream was successfully established for each task tone (A, B, and C) and that the assumption of the remaining tones creating the perceptual background was valid. Regarding background processing, an MMN was elicited by location deviants in each task-tone condition, providing evidence in favor of the background-segregation hypothesis. However, short-SOA timing deviants also elicited an MMN-like component when A and C tones were task-relevant, whereas long-SOA timing deviants did not elicit an MMN. This finding is in line with the background-integration hypothesis.

Thus, we found positive evidence for background integration and segregation in parallel. After ruling out a number of alternative interpretations for this finding (9), we suggest that both mental representations were concurrently active in the background. Our findings also imply that the MMN does not directly measure the selection of a given perceptual alternative, but the availability of a mental representation. This interpretation is supported by previous findings on two-tone sequences (22,23).

2.3 Verifying auditory subjective reports with eye-tracking

From a methodological point of view, there are several challenges in measuring auditory processing, regardless of whether the foreground or the background is addressed. Objective methods such as task- or EEG-based procedures are indirect measures of the current perceptual organization since one has to deduce the percept from a given behavior or the elicitation of a component in the ERP (see the background formation study (9)). These methods are additionally limited by sparse sampling (especially when based on deviance detection as a performance measure) and by the signal-to-noise ratio in the case of physiological measurements. Sparse sampling implies that the percept is only queried at discrete time points, such as at every tenth stimulus on average. This prolongs data

collection by a factor of ten, and might fail to capture fast changes in perception.

On the contrary, there are methods where participants are asked to directly report their subjective perception. These methods are easily implemented as continuous measures (see the foreground formation study (8)), whose sampling rate is only limited by the sampling rate of the employed response device. This makes them very economical and highly accurate regarding temporal resolution. They also correspond well with performance (4) and MMN-based methods (24) in most cases. However, report measures are difficult to verify, and their validity is more sensitive to variations in instructions, training and categorization abilities than objective measurements. The described foreground study (8), for instance, entailed a long training where participants practiced not only identifying their current percept but also handling the high number of response options.

Instructions and training would be less effortful if there were methods to capture the participants' perception *en passant* – in other words, without asking them. This is called a no-report paradigm and has already been established in vision (25,26). Visual no-report methods benefit from the observability of the eyes: Reflexive behaviors such as pupil dilation or the optokinetic nystagmus (OKN) are used to identify the perceptual state of the observer (25,27). For instance, when two different images that cannot be fused are presented, each to one eye, only one of the images is perceived at any given moment, and the perceptually dominant image alternates over time (binocular rivalry (28)). A no-report version of this binocular-rivalry paradigm can be created by having gratings drift in opposite directions – the slow phase of the OKN reliably follows the perceptually dominant grating in this case (25). We combined this visual no-report paradigm with auditory multistability to develop an *en passant* measure of the perceptual state in audition. Our aim was to create a binocular-rivalry stimulus whose dominant percept would be controlled by the currently dominant auditory percept. We associated each subsequence of a two-tone sequence with one of two binocularly presented gratings and hypothesized that the OKN slow phase follows the grating corresponding to the subsequence that was perceived in the foreground.

The auditory stimuli were two subsequences of sine tones, 400 Hz (A) and 1008 Hz (B), that were presented with an inter-tone interval of 400 ms and 600 ms, respectively. Frequency separation was rather high (16 semitones) in order to provoke a large proportion of segregated percepts. Participants listened to the sequences and continuously indicated whether they perceived the A tones, the B tones, or both tones equally in the foreground. Visual stimuli were two horizontally-drifting gratings one of which was presented to each eye. To induce binocular rivalry, the gratings were distinct in color (blue or red bars on a grey background) and motion direction (left or right). Each grating was associated with one subsequence by a pattern of square holes on the colored bars that jumped vertically in temporal synchrony with the respective tone onsets.

While participants continuously reported their auditory perception, they watched the gratings without paying attention to them, and their eye-movements were recorded. Since each tone-onset was associated with one grating drifting either to the left or the right, we expected eye movements in the corresponding direction, as extracted from the OKN slow phase.

Blinks, OKN fast phases (reorientation of the eye position), and phases in which the integrated percept (or both subsequences segregated but equally in the foreground) was reported were removed from the data. The correspondence of auditory foreground stream and OKN slow-phase direction was quantified by two measures: the signed gain and the d-prime. These measures were positive whenever the OKN slow-phase direction matched the motion direction of the grating that was associated with the reported foreground tone, and they were negative when the OKN matched the motion direction of the grating associated with the other tone. The signed gain, in addition, contained the velocity of the eye movements relative to the velocity of the gratings.

Three experiments were performed with this paradigm, and all of them showed an above-chance correspondence between eye movements and auditory perception, as evidenced by the signed gains and d-primes significantly exceeding zero. In the first experiment, this correspondence emerged only across the course of the experiment, while there were no significant effects in early blocks. To test whether this was caused by training effects on the auditory task or by a growth of the audio-visual coupling, we performed a second experiment with a prolonged auditory training phase. This time, both measures differed significantly from 0 during early and late blocks. A third experiment confirmed this observation by simplifying the reporting task via changing the timing of the tone presentations. Instead of a fixed interval within each subsequence, the tone-onsets were randomly placed in a 100-ms time window, with 200-ms intervals between successive tones. This increased the probability of stream segregation, and likely had effects on the strength of the audiovisual coupling as well (see (9) for

discussion).

To sum up, we showed the OKN slow phase induced by a visual multistable stimulus to match the reported subjective perception of an auditory multistable stimulus. Although the association was far from perfect, it shows the principle feasibility of tagging the perception of two-tone sequences by visual rivalry.

3. CONCLUSIONS

We outlined three studies on auditory scene analysis. Two of them addressed the organization of perceptual foreground and background in an auditory scene of higher complexity than the scenes typically chosen for investigations of auditory stream segregation and multistability. The third study proposed an *en passant* measure for auditory perception that might facilitate investigations on complex auditory scenes in the future.

Our first study indicates that the formation of the auditory foreground in three-tone sequences (complex scene) follows the same rules as in two-tone sequences (rather simple scene). We also found a surprisingly high proportion of perceptual organization alternatives with more than one (even complex) foreground stream at once. This was not expected because task-based studies show that attending to more than one stream requires a lot of training or effort (12,13). Nevertheless, it seems that the auditory system forms perceptual organizations comprising more than one sound source in the foreground without being forced to do this. Hence, subjective-reporting procedures should not implicitly assume a certain foreground-background distinction, but directly examine this distinction. In our second study we found that the formation of the auditory background is also surprisingly complex. It seems that the mutual exclusivity of auditory integration and segregation (2) does not hold when it comes to background formation. Auditory background processing might be more complex and more flexible than assumed previously.

In summary, the seemingly simple addition of one more tone to the 'AB' stimulus pattern that is mostly used for studying auditory scene analysis, be it foreground or background (1–3,6,7,16,24), has the potential to uncover novel aspects of how sounds are mentally organized into sound source configurations. The three-tone 'ABC' sequence enables the verification of knowledge from two-tone sequences, gives insights into perceptual background formation, and shows that there might be implicit and explicit assumptions on auditory scene analysis that do not generalize to complex scenes.

We are convinced that three-tone sequences, like the ones we employed in the current work, have the potential to uncover further important properties of auditory scene analysis. For building models of auditory stream formation, it is, for instance, important to know whether all possible ways of grouping the stimulus input are considered, or whether some organizations are excluded a priori. Along these lines, we found some evidence in our work (8) for background integration being impossible when the foreground stream contained the intermediate tones (B tones). Follow-up studies will examine whether this finding generalizes, and if so, what the underlying mechanism might be. It will also be important to study how the auditory system avoids combinatorial explosion when faced with an overwhelming number of grouping alternatives.

Extensions of the three-tone sequences towards situations with even more possible sound source configurations will unavoidably face challenges in training listeners to report their perception. These challenges could partly be relieved if perception could be queried without directly asking the listener. In our third study (10), we developed a new methodological approach where we used eye-movements linked to a visual multistable stimulus as a measurement tool for auditory perception. Since eye-movements and the reported auditory foreground formations were in good agreement, we believe that this approach has the potential to become a first *en passant* measure for auditory perception. At the very least, it can be used to verify that listeners' perceptual reports are plausible – in a research field that still debates about the validity of subjective and objective measurements (and validity of the resulting findings), this is a highly beneficial methodological addition.

REFERENCES

1. Moore BCJ, Gockel HE. Properties of auditory stream formation. Phil Trans R Soc B. 2012;367(1591):919–31.
2. van Noorden LPAS. Temporal coherence in the perception of tone sequences. Doctoral dissertation, Eindhoven University of Technology, Eindhoven, The Netherlands; 1975.
3. Denham SL, Gyimesi K, Stefanics G, Winkler I. Perceptual bistability in auditory streaming: How much do stimulus features matter? Learn Percept. 2013;5(Suppl. 2):73–100.
4. Micheyl C, Oxenham AJ. Objective and Subjective Psychophysical Measures of Auditory Stream Integration and Segregation. J Assoc Res Otolaryngol. 2010;11(4):709–24.
5. Pressnitzer D, Hupé J-M. Temporal dynamics of auditory and visual bistability reveal common principles of perceptual organization. Curr Biol. 2006;16(13):1351–7.
6. Denham SL, Bőhm TM, Bendixen A, Szalárdy O, Kocsis Z, Mill R, et al. Stable individual characteristics in the perception of multiple embedded patterns in multistable auditory stimuli. Front Neurosci. 2014;8.
7. Bendixen A, Denham SL, Gyimesi K, Winkler I, Gymesi K, Winkler I. Regular patterns stabilize auditory streams. J Acous Soc Am. 2010;128(6):3658–66.
8. Thomassen S, Bendixen A. Subjective perceptual organization of a complex auditory scene. J Acous Soc Am. 2017;141(1):265–76.
9. Thomassen S, Bendixen A. Assessing the background decomposition of a complex auditory scene with event-related brain potentials. Hear Res. 2018;370:120–9.
10. Einhäuser W, Thomassen S, Bendixen A. Using binocular rivalry to tag foreground sounds: towards an objective visual measure for auditory multistability. J Vis. 2017;17(1):1–19.
11. Brochard R, Drake C, Botte M-C, McAdams S. Perceptual organization of complex auditory sequences: Effect of number of simultaneous subsequences and frequency separation. J Exp Psychol - Hum Percept Perform. 1999;25(6):1742–59.
12. Demany L, Erviti M, Semal C. Auditory attention is divisible: Segregated tone streams can be tracked simultaneously. J Exp Psychol - Hum Percept Perform. 2015;41(2):356–63.
13. Gallun FJ, Mason CR, Kidd G. Task-dependent costs in processing two simultaneous auditory stimuli. Percept Psychophys. 2007;69(5):757–71.
14. Näätänen R, Paavilainen P, Rinne T, Alho K. The mismatch negativity (MMN) in basic research of central auditory processing: A review. Clin Neurophysiol. 2007;118(12):2544–90.
15. Rahne T, Sussman ES. Neural representations of auditory input accomodate to the context in a dynamically changing acoustic environment. Eur J Neurosci. 2009;29(1):205–11.
16. Sussman ES, Steinschneider M. Neurophysiological evidence for context-dependent encoding of sensory input in human auditory cortex. Brain Res. 2006;1075(1):165–74.
17. Spielmann MI, Schröger E, Kotz SA, Bendixen A. Attention effects on auditory scene analysis: insights from event-related brain potentials. Psychol Res. 2014;78(3):361–78.
18. Bregman AS, Campbell J. Primary auditory stream segregation and perception of order in rapid sequences of tones. J Exp Psychol. 1971;89(2):244–9.
19. Vliegen J, Moore BCJ, Oxenham AJ. The role of spectral and periodicity cues in auditory stream segregation, measured using a temporal discrimination task. J Acous Soc Am. 1999;106(2):938–45.
20. Yabe H, Tervaniemi M, Reinikainen K, Näätänen RN. Temporal window of integration revealed by MMN to sound omission. Neuroreport. 1997;8(8):1971–4.
21. Yabe H, Winkler I, Czigler I, Koyama S, Kakigi R, Sutoh T, et al. Organizing sound sequences in the human brain: The interplay of auditory streaming and temporal integration. Brain Res. 2001;897(1–2):222–7.
22. Horváth J, Czigler I, Sussman ES, Winkler I. Simultaneously active pre-attentive representations of local and global rules for sound sequences in the human brain. Cogn Brain Res. 2001;12(1):131–44.
23. Sussman ES, Bregman AS, Lee W-W. Effects of task-switching on neural representations of ambiguous sound input. Neuropsychologica. 2014;64:218–29.
24. Sussman ES, Ceponiene R, Shestakova A, Näätänen R, Winkler I. Auditory stream segregation processes operate similarly in school-aged children and adults. Hear Res. 2001;153(1–2):108–14.
25. Naber M, Frassle S, Einhäuser W. Perceptual Rivalry: Reflexes Reveal the Gradual Nature of Visual Awareness. Curr Sci. 2011;101(11):1435–9.
26. Tsuchiya N, Wilke M, Frässle S, Lamme VAF. No-Report Paradigms: Extracting the True Neural Correlates of Consciousness. Trends Cogn Sci. 2015;19(12):757–70.
27. Fox R, Todd S, Bettinger L. Optokinetic nystagmus as an objective indicator of binocular rivalry. Vis

Res. 1974;15:849–53.
28. Brascamp JW, Klink PC, Levelt JM. The ‘ laws ’ of binocular rivalry : 50 years of Levelt ’ s propositions. Vis Res. 2015;109:20–37.