Please cite this article in press as: Gandras K et al. Electrophysiological correlates of speaker segregation and foreground-background selection in ambiguous listening situations. Neuroscience (2017), http://dx.doi.org/10.1016/j.neuroscience.2017.07.021

Neuroscience xxx (2017) xxx–xxx

ELECTROPHYSIOLOGICAL CORRELATES OF SPEAKER SEGREGATION AND FOREGROUND-BACKGROUND SELECTION IN AMBIGUOUS LISTENING SITUATIONS

KATHARINA GANDRAS, [a*] **SABINE GRIMM** [b] **AND ALEXANDRA BENDIXEN** [a,b]

[a] *Department of Psychology, Cluster of Excellence "Hearing4all", European Medical School, Carl von Ossietzky University of Oldenburg, D-26111 Oldenburg, Germany*

[b] *Department of Physics, School of Natural Sciences, Chemnitz University of Technology, D-09126 Chemnitz, Germany*

Abstract—In everyday listening environments, a main task for our auditory system is to follow one out of multiple speakers talking simultaneously. The present study was designed to find electrophysiological indicators of two central processes involved – segregating the speech mixture into distinct speech sequences corresponding to the two speakers, and then attending to one of the speech sequences. We generated multistable speech stimuli that were set up to create ambiguity as to whether only one or two speakers are talking. Thereby we were able to investigate three perceptual alternatives (no segregation, segregated – speakerA in the foreground, segregated – speakerB in the foreground) without any confounding stimulus changes. Participants listened to a continuously repeating sequence of syllables, which were uttered alternately by two human speakers, and indicated whether they perceived the sequence as an inseparable mixture or as originating from two separate speakers. In the latter case, they distinguished which speaker was in their attentional foreground. Our data show a long-lasting event-related potential (ERP) modulation starting at 130 ms after stimulus onset, which can be explained by the perceptual organization of the two speech sequences into attended foreground and ignored background streams. Our paradigm extends previous work with pure-tone sequences toward speech stimuli and adds the possibility to obtain neural correlates of the difficulty to segregate a speech mixture into distinct streams.

This article is part of a Special Issue entitled: Sequence Processing. **© 2017 IBRO. Published by Elsevier Ltd. All rights reserved.**

Key words: auditory scene analysis, attention, perceptual multistability, sequence processing, speech, event-related potential (ERP).

*Corresponding author. Address: Department of Psychology, Carl von Ossietzky University of Oldenburg, Ammerländer Heerstraße 114-118, D-26129 Oldenburg, Germany.
E-mail addresses: katharina.gandras@uni-oldenburg.de (K. Gandras), sabine.grimm@physik.tu-chemnitz.de (S. Grimm), alexandra.bendixen@physik.tu-chemnitz.de (A. Bendixen).

http://dx.doi.org/10.1016/j.neuroscience.2017.07.021

INTRODUCTION

Disentangling two or more speakers in a complex auditory scene, such as in a busy cafeteria, poses a challenge to every listener. Our auditory system is able to make sense of such scenes by separating them into meaningful perceptual streams; an ability and branch of research termed *auditory scene analysis* (Bregman, 1990). To accomplish this task, target sound elements need to be separated from other sound elements (*segregation*) by using frequency, location or other cues; and they need to be bound together over time (*integration*) to form a coherent stream, such as sentences expressed by a conversational partner (Shamma et al., 2011; Snyder et al., 2012). For solving the challenging problem of auditory scene analysis, it is assumed that the auditory system is constantly exploring the acoustic environment and testing alternative ways of structuring the sensory input into perceptual units (Gregory, 1980; Winkler et al., 2012). This process can be captured with deliberately ambiguous sound sequences, where listeners report random switches between different perceptual interpretations (Denham and Winkler, 2006; Pressnitzer and Hupé, 2006). Such perceptual fluctuations despite unchanged stimulus parameters constitute a case of perceptual *bi-* or *multistability*. Auditory multistability has most frequently been studied with pure-tone (ABA-ABA-…) streaming paradigms (Gutschalk et al., 2005; Pressnitzer and Hupé, 2006; Denham et al., 2014). Their value lies in the fact that they permit research on perception without the confounding influence of differences in stimulus parameters. As such, multistability has become a popular research tool (cf. Pressnitzer et al., 2011).

Most studies on auditory multistability have focused on having listeners distinguish between integrated and segregated percepts (e.g., 'ABA' triplet sequences versus separate sequences of 'A' and 'B' tones). A few studies (e.g. Gutschalk et al., 2005) have additionally included the distinction between perceptual foreground and background in the segregated case (e.g., whether the 'A' or the 'B' tone sequence appears in the focus of attention). This distinction is important as it provides methodological and conceptual links between the research fields of auditory multistability (Pressnitzer et al., 2011; Snyder et al., 2012) and auditory-selective attention in situations with more than one sound source (Fritz et al., 2007; Zion Golumbic et al., 2013; Bronkhorst, 2015).

Because the foreground-background distinction is not very salient in the classic 'ABA' auditory streaming paradigm, Szalárdy and colleagues (2013b) developed a more complex variant of this paradigm that allowed listeners to clearly differentiate between perceptual foreground and background. Based on Wessel (1979), they employed rising 3-tone pitch patterns ('123123123...') with alternating timbre (creating a repeating six-tone sequence 'A1B2A3B1A2B3', with 'A' and 'B' denoting the different timbres). With this stimulus, sequential integration leads to the percept of a rising tone pattern (123123), whereas sequential segregation leads to the percept of a falling tone pattern with either timbre A (A3A2A1...) or timbre B (B3B2B1) in the perceptual foreground. Those three perceptual options allowed Szalárdy and colleagues (2013b) to investigate event-related potential (ERP) correlates of sequential integration versus segregation as well as of perceptual foreground-background formation with the same stimulus. They observed an early difference between foreground and background ERPs in the P1 latency range (around 70 ms after stimulus onset).

In the present study, we aimed to assess the replicability of the findings of Szalárdy et al. (2013b) when transferring the paradigm to speech signals. Finding similar effects for speech material would permit clearer links between research on ERP correlates of multistability and auditory-selective attention (e.g. O'Sullivan et al., 2015), since studies in the latter field often investigate speaker selection (as opposed to the tone sequences in auditory multistability). Previous research manipulating statistical structure in single sound streams has demonstrated that material type (speech versus non-speech) plays a role for segmenting a continuous stream into its constituent units (e.g., a speech stream into the constituent words or syllables) (Tremblay et al., 2013). It is thus conceivable that the decomposition of sound mixtures into their sound sources is likewise affected by the type of auditory material; be it as a consequence of increased stimulus familiarity for speech material or of genuine domain-specific processing. Hence we designed an experiment that transferred the approach of Szalárdy et al. (2013b) to speech material. We developed a sequence of interleaved consonant–vowel syllables uttered alternately by two human speakers, making sure that this sequence is able to evoke multistable perception. Participants listened to the syllable sequences and continuously indicated via button presses whether they perceived the sequence as an inseparable mixture of both speakers (*integrated percept*) or as two separate streams. In the latter case, they distinguished which speaker was in the attentional foreground (*segregated-speakerA percept* or *segregated-speakerB percept*). Thus, our paradigm captures the two key features of the classic cocktail party situation (Cherry, 1953), in which the listener tries to recognize what one person is saying (attended foreground) while others are speaking at the same time (ignored background) whose voices are sometimes difficult to tell apart from that of the conversation partner (stream segregation). The perceptual organization alternatives are readily relatable to those at the core of current selective auditory attention research (a critical difference of our paradigm to classical multistable speech stimuli such as those causing the verbal transformation effect, cf. Warren, 1968). By combining perceptual reports with electroencephalography (EEG), we aimed at identifying neural correlates of the attentional foreground and background representations. A difference between the ERP markers of foreground and background percept should reflect only the effect of the listener's perception, because the stimulus stays constant.

We expect to find ERP modulations that are governed by the perceptual organization of the speech sequences. If the ERP effects mainly reflect sound source selection as observed in Szalárdy et al. (2013b), we should find that the ERPs elicited by background tones differ from those elicited by foreground tones. Since each integrated syllable is part of the perceptual foreground, we hypothesize a difference between the ERPs elicited during integrated and segregated-background percepts, but not between the ERPs elicited during integrated and segregated-foreground percepts (cf. Szalárdy et al., 2013b). If, in contrast, the ERP effects also reflect sound source segregation, we should find time ranges in which the ERP elicited during integrated percepts differs from the ERPs in both segregated percepts regardless of foreground/background selection.

EXPERIMENTAL PROCEDURES

Participants

Twenty-seven healthy adult volunteers (15 male) between the age of 19 and 30 years (average age 23.7 ± 3.2 years) participated in the experiment. All participants were native German speakers, right-handed, and reported normal hearing as well as normal or corrected-to-normal vision. They had no history of neurological disorder and did not take any medication acting on the central nervous system. All experimental procedures were approved by the ethics committee of the University of Oldenburg and conducted in accordance with the principles laid out in the Declaration of Helsinki (World Medical Association, 2013).

Data of seven participants were excluded post hoc due to complications with the behavioral measures or to highly unbalanced perceptual reports. One participant showed an inconsistent response pattern (the oral description given after each block did not match the actual response behavior), indicating difficulties in understanding or following the instructions. Two participants performed poorly (more than two standard deviations below the mean performance) in the unambiguous control conditions (see section Experimental Procedure for details). Another four participants were excluded from the EEG analysis due to a strong imbalance of the proportions of the three different percepts: Each of them experienced at least one of the percepts of interest in less than 5% of the time. While this is a valid response pattern, it implies that not enough trials were available for achieving a sufficient signal-to-noise ratio in the EEG analysis.

Please cite this article in press as: Gandras K et al. Electrophysiological correlates of speaker segregation and foreground-background selection in ambiguous listening situations. Neuroscience (2017), http://dx.doi.org/10.1016/j.neuroscience.2017.07.021