Received: 22 December 2017 | Revised: 25 June 2018 | Accepted: 26 June 2018

DOI: 10.1002/hbm.24310



WILEY

RESEARCH ARTICLE

# Context-dependent role of selective attention for change detection in multi-speaker scenes

Christian Starzynski | Alexander Gutschalk

Department of Neurology, Ruprecht-Karls-Universität Heidelberg, Heidelberg, Germany

**Correspondence**
Alexander Gutschalk, Department of Neurology, Ruprecht-Karls-University Heidelberg, Im Neuenheimer Feld 400, Heidelberg 69120, Germany.
Email: alexander.gutschalk@med.uni-heidelberg.de

**Funding information**
Deutsche Forschungsgemeinschaft, Grant/Award Number: GU593/3-2

## Abstract

Disappearance of a voice or other sound source may often go unnoticed when the auditory scene is crowded. We explored the role of selective attention for this change deafness with magnetoencephalography in multi-speaker scenes. Each scene was presented two times in direct succession, and one target speaker was frequently omitted in Scene 2. When listeners were previously cued to the target speaker, activity in auditory cortex time locked to the target speaker's sound envelope was selectively enhanced in Scene 1, as was determined by a cross-correlation analysis. Moreover, the response was stronger for hit trials than for miss trials, confirming that selective attention played a role for subsequent change detection. If selective attention to the streams where the change occurred was generally required for successful change detection, neural enhancement of this stream would also be expected without cue in hit compared to miss trials. However, when listeners were not previously cued to the target, no enhanced activity for the target speaker was observed for hit trials, and there was no significant difference between hit and miss trials. These results, first, confirm a role for attention in change detection for situations where the target source is known. Second, they suggest that the omission of a speaker, or more generally an auditory stream, can alternatively be detected without selective attentional enhancement of the target stream. Several models and strategies could be envisaged for change detection in this case, including global comparison of the subsequent scenes.

**KEYWORDS**

attention, auditory scene analysis, change detection, magnetoencephalography, perceptual awareness

## 1 | INTRODUCTION

Auditory change detection is important for the effortless monitoring of the appearance, modification, or disappearance of relevant sound sources from the auditory scene. In complex auditory scenes, however, our change-detection capacity is limited, such that acoustic changes that are well above the sensory threshold may nevertheless be missed (Eramudugolla, Irvine, McAnally, Martin, & Mattingley, 2005), a phenomenon that has been labeled change deafness.

One hypothesis to explain change deafness is limited capacity of selective attention (or working memory). This hypothesis is supported by behavioral experiments demonstrating that cuing attention to the auditory stream where the change may be expected reduces change deafness (Eramudugolla et al., 2005; Irsik, Vanden Bosch der Nederlanden, & Snyder, 2016). An alternative hypothesis is that change detection in complex auditory scenes may not necessarily require focal attention or segregation of sound sources, but may instead be related to more basic mechanisms such as the coding of transient signals in combination with stimulus-specific adaptation (Cervantes Constantino, Pinggera, Paranamana, Kashino, & Chait, 2012). The latter hypothesis is supported by the asymmetry of change detection for appearing and disappearing streams and by the finding that participants were not generally able to identify the stream where the change occurred retrospectively (Cervantes Constantino et al., 2012).

Physiological studies of change deafness revealed a late negative response (Gregg & Snyder, 2012; Puschmann et al., 2013; Sohoglu & Chait, 2016) that is only evoked at the onset of detected changes, but it remains unresolved if this response is related to a change within a single auditory stream or operates on the whole scene. A related change response, the mismatch negativity (Näätänen, Gaillard, &

  

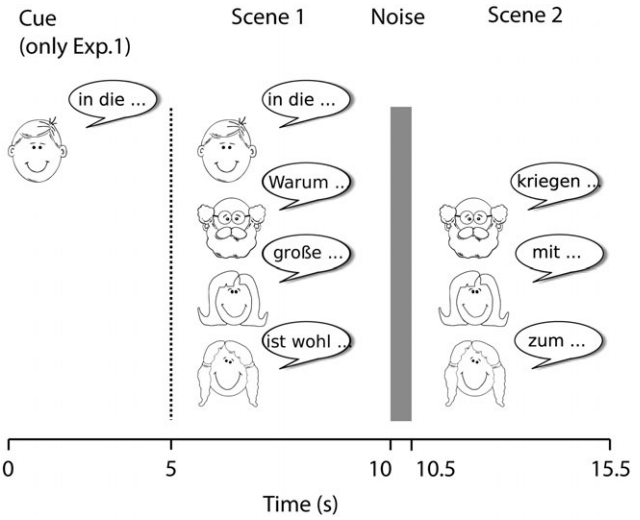


**FIGURE 1** Experimental design. Scene 1 comprised four speakers with two female and two male voices. After 5 s, a white noise burst indicated the beginning of Scene 2. In 80% of trials, one speaker—the target speaker—was omitted. In the other 20%, Scene 2 comprised the same four speakers as Scene 1 (not shown). In Experiment 1, participants were cued to the target speaker by playing only the target speaker ahead of Scene 1. Experiment 2 had the same setup, but no cue was provided

Mantysalo, 1978), has been shown to operate on the level of consciously perceived streams (Dykstra & Gutschalk, 2015; Sussman, Chen, Sussman-Fort, & Dinces, 2014), but change detection may still operate differently in the fundamentally different setups used to probe change deafness.

Here, we evaluated the neural representation of auditory streams before and after a change, to explore if selective attention to the relevant auditory stream is required for change detection, and if a lack thereof is a source of change deafness. The scenes comprised four different speakers, which were presented twice (Figure 1). In the second presentation of the scene, one of the speakers was omitted (80%) or all four speakers were presented again (20%).

The ongoing activity evoked by each speaker in the auditory cortex was separately reconstructed by cross-correlation analysis with the speech envelope (Abrams, Nicol, Zecker, & Kraus, 2008; Aiken & Picton, 2008; Hertrich, Dietrich, Trouvain, Moos, & Ackermann, 2012). As it has been previously shown that speech-locked activity is strongly enhanced when one of two speakers is selectively attended (Ding & Simon, 2012a; Ding & Simon, 2012b; Kerlin, Shahin, & Miller, 2010; Mesgarani & Chang, 2012; Zion Golumbic et al., 2013), we first establish that this finding can be reliably reproduced with four simultaneous speakers. This was done in the first experiment, where we cued participants to the relevant speaker by presenting this speaker in isolation before the two multi-speaker scenes, and instructed them to attend to this speaker in order to report if it was omitted in Scene 2. In the second experiment, participants were not cued to the relevant speaker. Based on the assumption that change detection depends on attention to the relevant stream (Eramudugolla et al., 2005), we tested the hypothesis that activity evoked by the omitted speaker in Experiment 2 is enhanced in the first scene in trials where the change is subsequently detected.

## 2 | MATERIALS AND METHODS

### 2.1 | Participants

In Experiment 1, 13 listeners (four males, 20–32 years old, mean age 25.4 years) participated. Two participants were removed from further analysis because of a high false alarm rate (>80%). In Experiment 2, 16 listeners (eight females, 19–32 years old, mean age 25.3 years) participated. One participant was removed from analysis because he/she did not comply with the task instruction. All participants were native speakers of German.

### 2.2 | Stimuli

Multi-speaker scenes were based on German audiobooks with 12 different male and 12 female speakers (source: vorleser.net), in which 8 of the 24 speakers (four males, four females) were potential target speakers, who were sometimes omitted in Scene 2. The files were cut into segments of 5 s. All segments with silent periods longer than 0.5 s were discarded. The remaining segments were normalized to the same maximal root-mean-square value. A 10-ms linear ramp at the beginning and the end of each sound clip was applied. All sounds were presented diotically using ER-3 earphones (Etymotic Research, Elk Grove Village, IL) with a mean sound pressure level of the target streams of 75 dB.

Sequences of sine tones were presented for the purpose of modeling source activity. A sequence consisted of 20 different sine tones with 5-ms raised cosine onset and offset ramps with frequencies ranging from 700 to 1,300 Hz (logarithmically scaled) and a duration of 100 ms. The tones were played in random order with an interstimulus interval of 0.8 s. A sine-tone sequence was played after every 20th change-detection trial, that is, six times during the whole session. Subjects were instructed to listen passively to the tones.

Change-detection trials consisted of two 5-s scenes separated by a 500-ms noise burst to indicate the end of the first and the beginning of the second scene. The purpose of the noise burst was solely to structure the stimulus, because the transition between scenes was otherwise hard to identify when the scenes were separated by silence, only. Scene 1 comprised two male and two female speakers. In 20% of all cases (32 trials), Scene 2 comprised the same four speakers reading another, randomly chosen passage. These trials were used as catch trials to estimate the listeners' false alarm rate. The other 80% (128 trials) were change trials, where one of the four speakers was omitted in Scene 2. Each set with a disappearing speaker was used two times in the experiment, where in one occasion, the target was a male speaker, and in the other the target was a female speaker. The set of trials was identical for all experiments. Participants indicated with a button press after the second scene if they had detected a change or not. The same button press started a new trial.

In the first experiment, the target speaker was played alone as a 5-s-long cue before each trial, speaking the same text as in the directly following multi-speaker Scene 1. Listeners were instructed to attend to this speaker in Scene 1 and indicate if the speaker disappeared in Scene 2. No feedback was provided whether the response was correct or not. The text in Scene 2 was different from Scene 1. While the text for each speaker in Scene 2 was taken from the same story as in

Scene 1, it was not the direct continuation of Scene 1, to avoid strong contextual cues. This setup was chosen to base the change detection on the speaker identity rather than on the context.

In the second experiment, a different group of participants was instructed to listen to Scene 1 and to judge whether the same speakers were also present in Scene 2 without prior cue to the target speaker. Except for the omitted cue sequence, the experimental setup was identical to Experiment 1.

## 2.3 | Speech-data transformation

For the cross-correlation analysis, the speech data were processed separately for each speaker comprised in the multi-speaker scenes (Figure 2). The speech data were band-pass filtered (Butterworth second order, cutoff frequencies: 200 and 5,000 Hz), because previous studies showed that the sensitivity of the auditory cortex is maximal for frequency content in the range of 400–2000 Hz (Ding & Simon, 2012a). In order to derive the speakers' temporal envelope, a cochlear model based on a gammatone filter bank (200–5,000 Hz, number of filters $N = 100$) and half-wave rectification with compression and low-pass filtering (cutoff frequency: 10 Hz) was applied to the speech data (Fontaine, Goodman, Benichoux, & Brette, 2011). The resulting amplitude values of the different filters were summed up, sampled down to the magnetoencephalography (MEG) sample rate (1,000 Hz), and normalized by the number of filters ($N = 100$). This type of signal processing was chosen to be consistent with the physiological processing up to the cochlear nucleus (Dau, Kollmeier, & Kohlrausch, 1997), but we expect only minor difference in comparison to envelope extraction based on a Hilbert transform, which has alternatively been used in previous studies (Abrams et al., 2008; Aiken & Picton, 2008; Hertrich et al., 2012).

Finally, the first derivative of the envelope was calculated and negative values of the first derivative were set to zero (half-wave rectification). The first derivate rather than the unprocessed envelope was used based on the assumption that the speech-synchronized MEG signal is mostly driven by the rising part of the amplitude fluctuations (Biermann & Heil, 2000). It has been pointed out that these gains in intensity are loosely correlated with syllable onsets (Hertrich et al., 2012). If the original envelope was used, instead, falling slopes would theoretically predict transient response of opposite polarity as rising slopes. However, this is in contrast to the observations that offset responses in the auditory cortex have a similar structure and polarity as onset responses (Pantev, Eulitz, Hampson, Ross, & Roberts, 1996). Using the first derivative and subsequent half-wave rectification, the estimated waveform is restricted to onset responses, instead. This is reasonable for our case, because offset responses are smaller than onset responses, unless they follow multiple-second-long sounds. An analysis of offset responses with identical polarity as onset responses would be possible by including the negative values after polarity inversion, but this was omitted in the present analysis.

## 2.4 | MEG acquisition and analysis

The MEG was recorded continuously with a Neuromag-122 (Elekta Neuromag, Helsinki, Finland) whole-head MEG system in a four-layer magnetically shielded room (IMEDCO, Hägendorf, Switzerland) using a sampling rate of 1 kHz. The head position inside the device was measured with four position-indicator coils fixed to the listeners' head. The position of these coils relative to the head surface was registered with an Isotrak II digitizer (Polhemus, Colchester, CT).

The averaged response evoked by sine tones was used to fit a pair of dipoles in auditory cortex to the $N_1$m peak using BESA 5.1 (BESA GmbH, Gräfelfing, Germany) individually for each participant (one dipole in the left and one in the right auditory cortex). The $N_1$m was chosen because this component showed the strongest attention-related modulation in previous studies (Ding & Simon, 2012a; Hillyard, Hink, Schwent, & Picton, 1973). Furthermore, a pair of regional sources was set at the position of the eyes in order to model artifacts caused by eye movements. A principal component analysis-based spatial filtering of streetcar artifacts was additionally applied. All topographies were combined into one individual spatial filter for each subject, which was used to calculate dipole-source waveforms of the raw MEG data for further processing with Python (Scipy and Brian libraries).

The speech-data trials were classified by their type (first and second scene; hit, miss, false alarm, and correct rejection). The analysis was limited to change trials, because the number of trials where no change was present was too small to evaluate the MEG response to false alarm and correct rejections with sufficient signal-to-noise ratio. After linear de-trending, the speaker-specific, speech-locked MEG signal was obtained by calculating the cross correlation between the processed speech data (cf. previous section) and the corresponding continuous source-waveform segment with a maximal lag of 350 ms. The resulting signals were band-pass filtered (second order, zero-phase-shift Butterworth filter, cutoff frequencies: 0.5 and 70 Hz), and averaged according to the trial and speaker type.

The resulting waveforms were then analyzed like standard evoked-response source waveforms. Based on the average latency across conditions and subjects, the $P_1$m and $N_1$m source strength were measured as mean activity in the time windows 50–80 ms for the $P_1$m and 120–170 ms for the $N_1$m. Preliminary analysis of the data showed no differences between hemispheres. Therefore, the data were collapsed across hemispheres for the final analysis reported here. Amplitudes were measured for: (a) the target speaker, that is, the speaker who was omitted in Scene 2 and was additionally cued in Experiment 1; (b) the distractor of the same sex as the target speaker; and (c) the two distractor speakers of the opposite sex (the response to the two opposite-sex distractor speakers were subsequently averaged). Differences between same-sex and different-sex distractors were evaluated with an ANOVA with the factors trial type (hit, miss) and speaker type (same-sex distractor, different-sex distractor), which was separately computed for the $P_1$m and $N_1$m, and for Scenes 1 and 2. As no difference was observed between same- and different-sex distractors in Scene 1, the distractor data were averaged for the comparison with the target-evoked data. The statistical analysis of target-speaker effects was performed with an ANOVA with the factors trial type (hit, miss) and speaker type (target, distractor), separately computed for the $P_1$m and $N_1$m components. Significant effects were further evaluated with Tukey's test for multiple comparisons of means, with a 95% family wise confidence level.

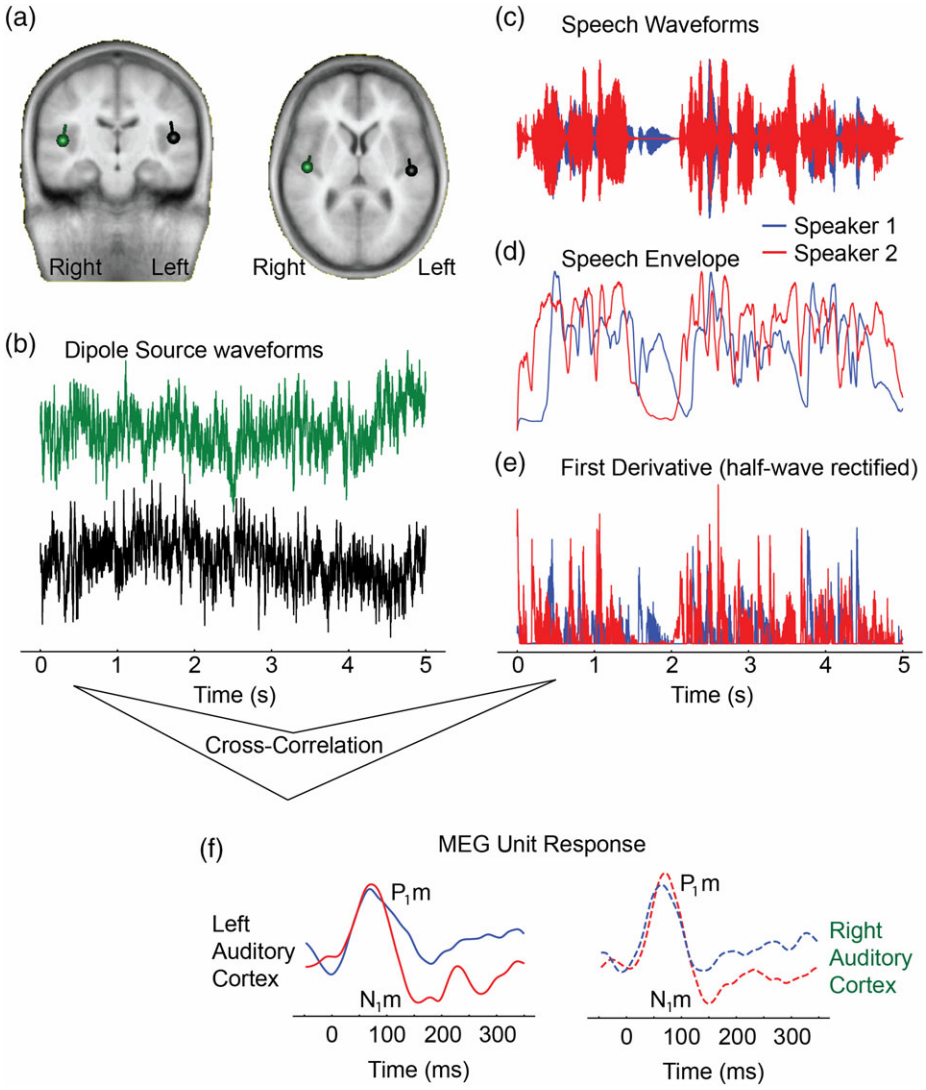


**FIGURE 2** Procedures for the cross-correlation analysis: (a) as localizer, the $N_1$m evoked by pure tones was modeled with two dipoles, one in the right and one in the left auditory cortex. (b) The individual dipole model was used as a spatial filter in combination with artifact topographies to extract continuous source waveforms for 5-s long time intervals, which correspond to the presentation of the speech stimuli. (c) The corresponding speech stimuli were processed separately (here we show only two of four speakers for convenience). (d) First, the envelope of each speaker was computed with an auditory model, including a gamma-tone filter bank, half-wave rectification, and compression. The results were then summed across frequency channels again, the result of which is plotted in panel d. (e) Finally, the first derivative of the signal was calculated and half-wave rectified. A cross correlation with a maximum time lack 350 ms was then calculated between the signals shown in (b) and (e), resulting in the unit responses shown in (f ), which models the auditory cortex response to the (syllable) onsets represented by the signal shown in (e). There is one waveform for each cross correlation, that is, two for each speaker (speaker 1: red, speaker 2: blue) and two for each hemisphere (left: solid, right: dashed), resulting in the four waveforms shown in this example. The unit response is highly similar to the tone-evoked response in auditory cortex, with prominent peaks $P_1$m and $N_1$m. Here, the listener attended to speaker 1, and accordingly the $N_1$m is more prominent for speaker 1. The activity was overall similar in the left and right auditory cortex, and the two hemispheres were therefore averaged for the main analysis [Color figure can be viewed at wileyonlinelibrary.com]

## 3 | RESULTS

### 3.1 | Experiment 1

When listeners were cued to the potentially omitted speaker beforehand, they correctly detected 77% of speaker omissions in Scene 2. Behavioral results are shown in Figure 3, the MEG source activity related to the speakers' speech envelope are shown in Figure 4. Based on the listeners' responses, change trials were sorted into correct change detection (Figure 4a,c) and miss trials (Figure 4b,d), where change deafness occurred despite cuing. The cued speaker (red) evoked stronger responses than concurrent distractor speakers in the $N_1$m interval (Figure 4a; Table 1), as was expected when listeners focused their attention on the cued speaker for the whole interval. Moreover, there was a significant effect of the behavioral classification (Table 1), indicating stronger $N_1$m responses in trials with correct change detection. While this effect was driven by the neural response evoked by the target speaker (see Table 1), the interaction of speaker × trial missed significance. It may therefore be that the attentional modulation is not limited to the cued target speaker, but that there is also a general response enhancement in hit trials, even though

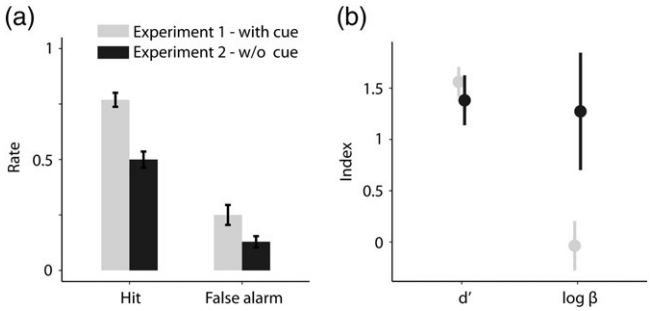


**FIGURE 3** Behavioral results. (a) The hit and false alarm rates (mean ± *SEM*) for Experiment 1 (with cue, red) were greater than for Experiment 2 (no cue, blue). (b) The sensitivity index d' was not statistically different between experiments. However, the positive bias log β for the uncued experiment indicated a more conservative response behavior in the uncued Experiment 2

this effect is numerically smaller than the effect observed for the target speaker (Figures 4a,b and 7c). The lack of an $N_1$m enhancement in miss trials suggests that listeners deployed their attention less effectively during Scene 1 of miss trials, thereby impeding correct change detection at the onset of Scene 2. There was no speaker-specific effect, but a significant trial effect was observed in the $P_1$m interval (Table 1). The effect was such that $P_1$m responses were smaller for correct detection trials (Figure 7a). Most likely, this difference also reflects the $N_1$m enhancement, given that the two response components overlap in time.

In change trials, the focus of attention in Scene 2 was apparently transferred to the remaining speaker of the same sex as the cued speaker. This is revealed by the comparison of the response to same- and opposite-sex speakers (Figure 4c,d), which reveals a significantly larger $N_1$m for the same-sex speaker compared to the other-sex speakers in Scene 2 (Table 1), but not in Scene 1.

For comparison, Figure 5 shows the response evoked by the cue speaker, presented alone, and by the target-speaker in Scene 1, presented together with three distractor speakers, averaged across hit and miss trials. The response to the cue is earlier in comparison to the target for the $P_1$m (cue: 53.2 ± 4.6 ms, target: 64.8 ± 9.3 ms; $p$ = 0.00368 [mean ± *SD*; two-tailed paired-sample $t$ test]) and the $N_1$m (cue: 115.6 ± 20.2 ms, target: 133.8 ± 13.4 ms; $p$ = 0.00612). The $P_1$m also showed a significantly larger amplitude (cue: 6.4 ± 3.1 ms, target: 1.5 ± 13.4 ms; $p$ = 0.00062) for cues, whereas the $N_1$m amplitude was not significantly different between cues and targets (cue: 3.0 ± 1.4 ms; target: 1.5 ± 0.6 ms, $p$ = 0.22572).

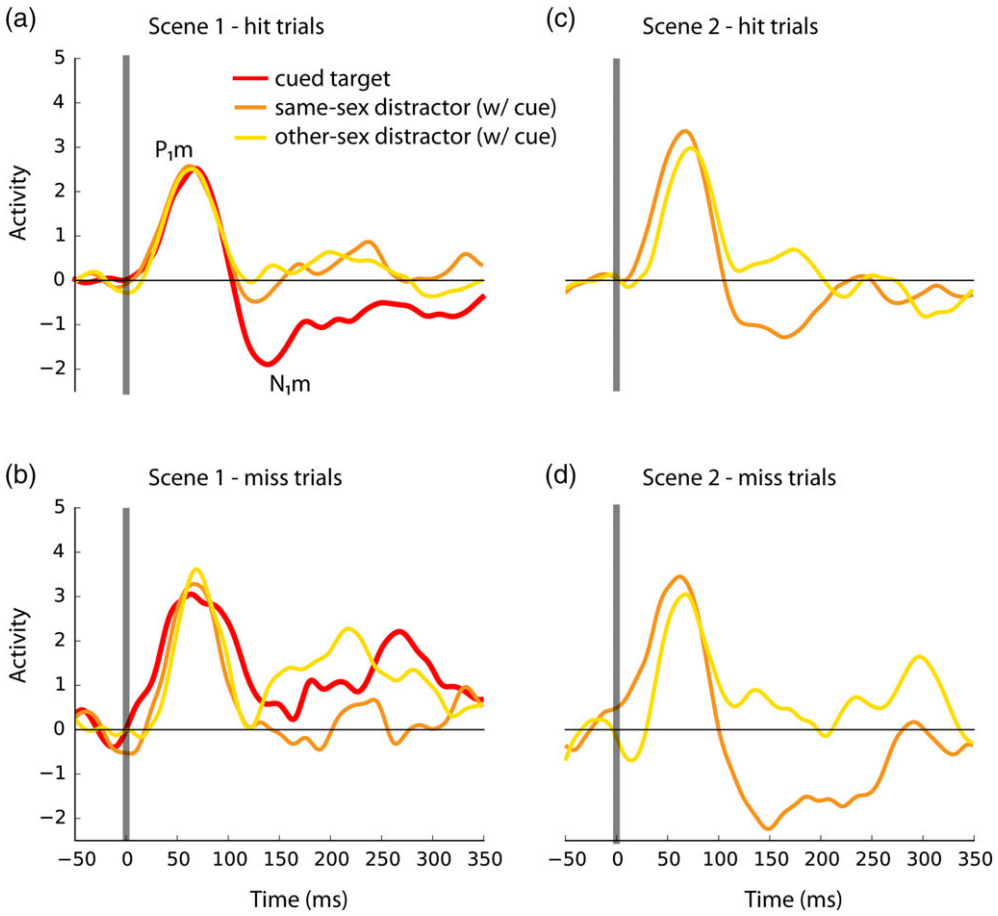


**FIGURE 4** Source activity in auditory cortex for the cued Experiment 1. Only change trials were evaluated; the waveforms represent a grand average across participants ($N$ = 11). Left panels Scene 1 (a andb), right panels Scene 2 (c and d). Hit trials in the upper panels (a and c), miss trials in the lower panels (b and d). There is clearly an effect of attentional modulation visible in the $N_1$m time interval (around 150 ms) indicating successful focusing of attention on the cued target speaker (red) in hit trials (a) but not in miss trials (b) of scene 1. Without the target speaker in Scene 2, subjects apparently switched their attention to the same-sex speaker (orange in c and d) in both trial types [Color figure can be viewed at wileyonlinelibrary.com]

**TABLE 1** Statistical analysis of MEG response amplitudes in Experiment 1 with cued target speaker

| Analysis | Factor | *F* value | *p* value |
|---|---|---|---|
| Cued target-speaker analysis | | | |
| Scene 1—$N_1$m | Speaker | $F_{1,116}$ = 5.0871 | 0.02598* |
| | Trial | $F_{1,116}$ = 10.7649 | 0.00137** |
| | Speaker × trial | $F_{1,116}$ = 2.3587 | 0.12731 |
| | Hit-target vs. hit-distractor | | 0.021* |
| | Miss-target vs. miss-distractor | | 0.944 |
| | Miss-target vs. hit-distractor | | 0.721 |
| | Hit-target vs. miss-distractor | | $p$ < 0.0001*** |
| | Hit-target vs. miss-target | | 0.004** |
| | Hit-distractor vs. miss-distractor | | 0.206 |
| Scene 1—$P_1$m | Speaker | $F_{1,116}$ = 1.3 | 0.697 |
| | Trial | $F_{1,116}$ = 6.18 | 0.014* |
| | Speaker × trial | $F_{1,116}$ = 0.39 | 0.535 |
| Distractor-speaker analysis | | | |
| Scene 1—$N_1$m | Speaker | $F_{1,76}$ = 2.35 | 0.13 |
| | Trial | $F_{1,76}$ = 4.11 | 0.046* |
| | Speaker × trial | $F_{1,76}$ = 0.55 | 0.459 |
| Scene 1—$P_1$m | Speaker | $F_{1,76}$ = 0.75 | 0.389 |
| | Trial | $F_{1,76}$ = 5.23 | 0.025* |
| | Speaker × trial | $F_{1,76}$ = 0.76 | 0.385 |
| Scene 2—$N_1$m | Speaker | $F_{1,76}$ = 7.17 | 0.00910** |
| | Trial | $F_{1,76}$ = 0.04 | 0.84323 |
| | Speaker × trial | $F_{1,76}$ = 0.37 | 0.54216 |
| Scene 2—$P_1$m | Speaker | $F_{1,76}$ = 0.15 | 0.6977 |
| | Trial | $F_{1,76}$ = 0.04 | 0.8451 |
| | Speaker × trial | $F_{1,76}$ = 0.09 | 0.7639 |

The effect on the cued target speaker is evaluated with an ANOVA with the factors trial (hit, miss) and speaker (target, distractor). Differences between the activity evoked by distractors are additionally evaluated with an ANOVA with the factors sex (same or other as target speaker) and trial (hit, miss). The latter analysis was performed separately for Scenes 1 and 2. Tukey's multiple comparisons of means are used for paired post hoc tests in both cases. Only significant paired tests are reported with the exception of the main hypothesis (target $N_1$m in Scene 1). * < 0.05; ** < 0.01; *** < 0.001

### 3.2 | Experiment 2

When listeners were not cued to the potentially omitted speaker, their average detection rate was only 50% and thus lower than in the cued Experiment 1 (Welch two-sample *t* test $t$ = 5.6, $df$ = 23.96, $p = 8.8 \cdot 10^{-6}$). While a lower detection rate was expected without a cue based on previous studies (Eramudugolla et al., 2005), the detection rate does not reveal the full story, because the false alarm rate was also significantly lower than in Experiment 1 (13 vs. 25%; Welch two-sample *t* test $t$ = 2.3, $df$ = 16.2, $p$ = 0.03). Therefore, the detectability index d' was not significantly different between Experiments 1 and 2 (Welch two-sample *t* test $t$ = 0.63, $df$ = 21.9, $p$ = 0.53), but only the bias log β (Welch two-sample *t* test $t$ = −2.11, $df$ = 18.6, $p$ = 0.049), that is, listeners made more conservative decision when no cue to the target was available.

Figure 6 shows the source waveforms obtained in Experiment 2 plotted in the same schema used for Experiment 1. The darkest blue, bold line is the response evoked by the speaker that is omitted in Scene 2. As can be observed in this figure, the evoked response in the $N_1$m interval is overall similar across all conditions plotted. In the statistical analysis (Table 2), there was no significant difference between target and distractor speakers or between hit and miss trials in the $P_1$m and $N_1$m intervals, and there was no numerical trend of the same direction as the effect observed in Experiment 1 (Figure 7). We further tested if significant differences were observed in other time windows or between speakers of the target sex vs. the opposite sex, but could not find any relevant statistical trends in these exploratory analyses. Based on the results of Experiment 1, the minimal effect size that could have been detected with $N$ = 15 participants in Experiment 2 was estimated to be approximately 75% of the effect for stronger target than distractor responses in hit trials, and 50% of the effect for stronger target responses in hit compared to miss trials (power = 0.8, alpha = 0.05, paired *t* test). Because of the lower false alarm rate in Experiment 2, the signal-to-noise ratio for the comparison of hit and miss trials should rather be somewhat better than in Experiment 1.

In addition to the lack of a significant difference in the $N_1$m interval, the $N_1$m appears to be generally smaller in Experiment 2. The comparison of response amplitude in the $N_1$m time interval between Experiments 1 and 2 (*t* test, corrected for multiple comparison (Benjamini & Yekutieli, 2001)) reveals a significantly more positive response in Experiment 2 across all speakers, including the supposedly unattended opposite-sex distractors (see Figure 7). No significant difference was found in the $P_1$m interval.

Finally, to probe whether attention in Experiment 2 was instead directed toward the whole scene, we evaluated if an (enhanced) $N_1$m response was observed when the analysis was based on the envelope of the summed scene instead of the single speakers, but no such difference was observed (data not shown).

## 4 | DISCUSSION

These results indicate that different modes might be used for detecting a speaker's disappearance from a crowd. In Experiment 1, listeners focused on the cued speaker in Scene 1 and supposedly searched for the voice of this speaker in Scene 2. The successful focusing of the cued speaker in Scene 1 is well documented by the selective

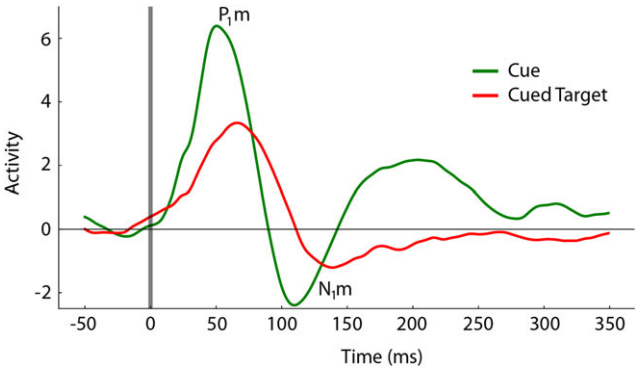


**FIGURE 5** Comparison of the response evoked by the cue speaker (green) and the subsequent, cued target speaker (red) in Scene 1 of Experiment 1. The same speech stimuli are used in both cases, but the target was presented together with three distractor speakers [Color figure can be viewed at wileyonlinelibrary.com]

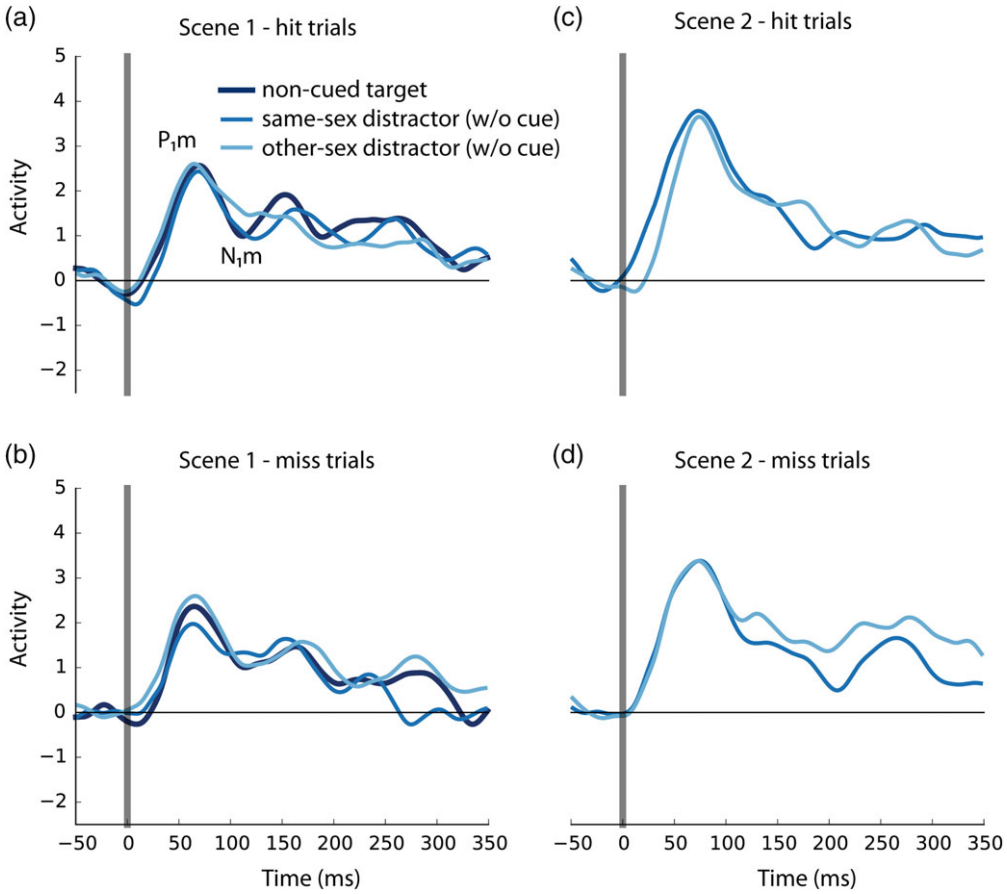


**FIGURE 6** Source activity in auditory cortex for the uncued Experiment 2. Only change trials were evaluated; the waveforms represent a grand average across participants ($N$ = 15). Left panels Scene 1 (a and b), right panels Scene 2 (c and d). Hit trials in the upper panels (a and c), miss trials in the lower panels (b and d). In contrast to Experiment 1, no $N_1$m enhancement for the target speaker (darkest blue) was observed in hit trials (a) [Color figure can be viewed at wileyonlinelibrary.com]

enhancement of the $N_1$m evoked by the cued speaker. Moreover, the relevance of attentional enhancement for the subsequent comparison with Scene 2 is demonstrated by the lower $N_1$m amplitude in miss trials. While the nature of the comparison process cannot be directly reconstructed from the data, the enhancement of the same-sex speaker in Scene 2 is suggestive of a search for the cued voice. Because the target speaker is not present in the scene, listeners focus their attention on the most similar, same-sex-distracter speaker instead, and supposedly decide whether this speaker is the same as the cued speaker followed along Scene 1 or not. Other mechanisms, as for example, an automatic, mismatch-negativity-like (Näätänen et al., 1978) comparison process, cannot be excluded, but would probably not invoke attentional orienting to the previously competing, most similar speech stream.

This mode of auditory change detection would generally be in line with previous suggestions that change deafness was due to limited attentional resources (Eramudugolla et al., 2005), or limited resources of working memory capacity (McAnally et al., 2010; Pavani & Turatto, 2008). While $N_1$m enhancement is an indicator of current attentional focus, most recent models of working memory consider the same resources active for selective attention and working memory processes alike (D'Esposito & Postle, 2015). In "one-shot" change deafness paradigms with distinctively different sound objects (Eramudugolla et al., 2005; Gregg & Samuel, 2008; McAnally et al., 2010; Pavani & Turatto, 2008), the task could potentially be solved by listening only to the cue and to Scene 2, to decide if the object was part of the second scene or not. In contrast, the detection of a disappearing speaker is a more difficult task, unless the speakers are very unusual or have been previously known. While four voices may appear few at first, the listening impression is complex and one cannot easily capture all voices at once. Therefore, listeners may easily confuse one of the same-sex speakers with the cued speaker, in particular when attention was not sufficiently focused on the cued speaker until the end of Scene 1. The results strongly suggest that the intervening time period is too long and the risk of speaker confusion is high when the cued speaker is not followed along both scenes.

Based on the findings of Experiment 1, we tested in Experiment 2 if the randomly chosen focus of selective attention was required for change detection in an uncued setup. In this case, we would have expected that $N_1$m amplitude, as indicator of the variable attentional resource, would have been on average equally distributed across all four speakers in Scene 1. The subset of trials, where attention was incidentally focused on the target speaker, would be expected to more likely enable successful change detection (Eramudugolla et al., 2005). Accordingly, sorting trials based on behavioral performance should then reveal stronger $N_1$m for trials where the change was detected. The detection rate of about 50% would suggest that roughly two listeners can be attended sufficiently to enable change detection

**TABLE 2** Statistical analysis of MEG response amplitudes in Experiment 2 without cue

| Analysis | Factor | *F* value | *p* value |
|---|---|---|---|
| Uncued target-speaker analysis | | | |
| Scene 1—$N_1$m | Speaker | $F_{2,176} = 0.06$ | 0.81 |
| | Trial | $F_{1,176} = 0.86$ | 0.36 |
| | Speaker × trial | $F_{2,176} = 0.13$ | 0.71 |
| Scene 1—$P_1$m | Speaker | $F_{2,176} = 0.05$ | 0.82 |
| | Trial | $F_{1,176} = 0.42$ | 0.52 |
| | Speaker × trial | $F_{2,176} = 0.15$ | 0.70 |
| Distractor-speaker analysis | | | |
| Scene 1—$N_1$m | Speaker | $F_{2,116} = 0.02$ | 0.88 |
| | Trial | $F_{1,116} = 0.36$ | 0.55 |
| | Speaker × trial | $F_{2,116} = 0.19$ | 0.66 |
| Scene 2—$N_1$m | Speaker | $F_{1,116} = 0.28$ | 0.60 |
| | Trial | $F_{1,116} = 0.00$ | 0.94 |
| | Speaker × trial | $F_{1,116} = 0.18$ | 0.67 |
| Scene 1—$P_1$m | Speaker | $F_{2,116} = 1.25$ | 0.27 |
| | Trial | $F_{1,116} = 0.00$ | 0.99 |
| | Speaker × trial | $F_{2,116} = 0.03$ | 0.86 |
| Scene 2—$P_1$m | Speaker | $F_{1,116} = 0.01$ | 0.94 |
| | Trial | $F_{1,116} = 0.19$ | 0.66 |
| | Speaker × trial | $F_{1,116} = 0.24$ | 0.62 |

The effect on the cued target speaker is evaluated with an ANOVA with the factors trial (hit, miss) and speaker (target, distractor). Differences between the activity evoked by distractors are additionally evaluated with an ANOVA with the factors sex (same or other as target speaker) and trial (hit, miss). The latter analysis was performed separately for Scenes 1 and 2.

in the present setup. The results of Experiment 2 rule it unlikely that selective attention to one or two speakers was required for change detection, however. There was also no significant overall difference between hit and miss trials, as was observed in Experiment 1, making it unlikely that reduced nonselective attentional effort was the source of change deafness. Moreover, the data indicated overall less distinct $N_1$m responses in comparison to Experiment 1, suggesting that listeners were engaged in a different listening mode to solve the change-detection task in Experiment 2.

There are several possible strategies for how the changes could have been detected in Experiment 2, which cannot be dissociated based on the present data alone. Based on the finding that retro cues, that is, cues presented subsequent to Scene 1, can still enhance change detection (Backer & Alain, 2012), it could be argued that selective attention was not oriented during Scene 1, but that the same resources were used to maintain a subset of speakers in working memory at (after) the end of Scene 1. However, we find it unlikely that listeners would focus their attention on a subset of speakers at the end of the scene, but not during the scene. Moreover, if attention was focused on a single speaker, we might still expect that listeners perform a search for this speaker in Scene 2, and once they do not find a particular speaker are likely to focus their attention on the same-sex speaker, given that the other-sex speakers are easier to segregate. As no bias was observed in Scene 2, as well, we consider explanations involving the search for a single speaker unlikely.

Another possibility is that the listeners scanned all speakers serially and that it was not the amount of attention per speaker, but the scan order that determined the likelihood of detecting an omission. While this alternative hypothesis cannot be excluded, we would expect some $N_1$m enhancement for this case, equally distributed across speakers. The finding of a smaller $N_1$m in Experiment 2 compared to Experiment 1 rather suggests that selective attention toward single speakers was not used at all. Note, however, that this between-group comparison should be interpreted with caution, and that we have not yet confirmed that serial search produces similar $N_1$m enhancement like focusing a single speaker. Moreover, the power to detect attention effects in Experiment 2 are limited, and we can certainly not exclude small differences in selective attention between the four speakers with this setup.

It can also not be excluded that listeners tried to count the number of speakers in Scenes 1 and 2 and based their decision on the estimated number of speakers. In general, counting the number of speakers is difficult, in particular given the short time. If one discovers the scene structure of the experiment, it may be possible to separately count the number of male and female speakers, or listen for whether there are one or two of each kind. However, this is difficult and affords a dedicated listening strategy, and participants in Experiment 2 did not generally report using such a strategy, with one exception whose data were excluded from the analysis.

In general, the cue in Experiment 1 may not only have indicated which speaker is important, but additionally facilitated the segregation of this speaker from the multi-speaker babble. It is therefore conceivable that the cue produced a bias toward a speaker-based strategy beyond the task instruction, by a promotion of speaker segregation that may otherwise not emerge so easily. Without the cue, the paradigm might therefore be biased toward alternative listening strategies.

It has been previously suggested that change deafness may not be based on object-level processing but rather on the detection of transients within the scene (Cervantes Constantino et al., 2012). While the latter may work well for the appearance of an additional sound source, we consider it unlikely that transients are of major importance for the speaker-omission paradigm used here. Similarly, automatic (frequency shift) change detectors, which have been shown to operate for subliminal auditory stimuli (Demany, Semal, Cazalets, & Pressnitzer, 2010; Demany & Ramos, 2005), are unlikely to play a role for our highly variable stimuli in which one speaker was simply omitted.

Another possibility for a global change-detection strategy could be by representation of the multi-speaker babble as a sound texture (McDermott & Simoncelli, 2011), and comparison of the subsequent scenes or textures by summary statistics (McDermott, Schemitsch, & Simoncelli, 2013). This model would predict that the change detection performance would deteriorate when the speaker was exchanged by another speaker in Scene 2, instead of being omitted, because the summary statistics of the two scenes would soon converge (McDermott et al., 2013). In contrast, a selective attention strategy may be more effective in this case, because it allows for a direct comparison between two streams or speakers at an object level (Gutschalk & Dykstra, 2014; Shamma et al., 2013) and should therefore allow for similar performance as observed here for speaker omission, provided that the target speaker is indicated to the listener.

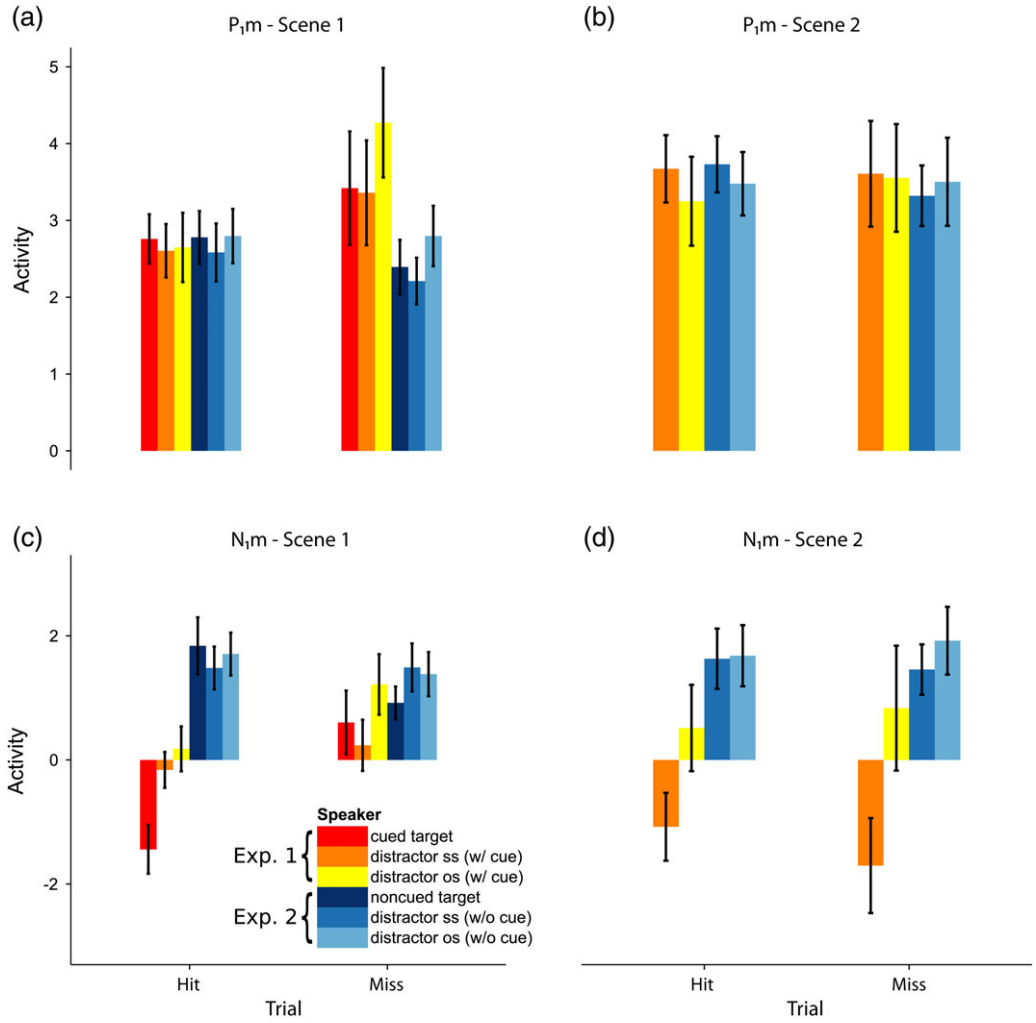


**FIGURE 7** Amplitudes for $P_1$m and $N_1$m components in Experiments 1 and 2. Average source amplitudes (mean ± *SEM*) were measured in the time interval 50–80 ms for the $P_1$m (upper panel, a and b) and 120–170 ms for the $N_1$m (lower panel, c and d). Colors code for speakers and experiments (hot colors for Experiment 1, cold colors for Experiment 2) are similar to Figures 3 and 4 [Color figure can be viewed at wileyonlinelibrary.com]

In general, future research on change deafness needs to acknowledge the context dependence of change detection strategies. It seems quite likely that change detection works adaptively and that different experimental setups therefore highlight different mechanisms, and that these mechanisms are not mutually exclusive.

## ACKNOWLEDGMENT

This work was supported by Deutsche Forschungsgemeinschaft grant DFG GU593/3-2 to A.G.

## ORCID

*Christian Starzynski* https://orcid.org/0000-0002-2905-4059
*Alexander Gutschalk* https://orcid.org/0000-0002-2523-8846

## REFERENCES

Abrams, D. A., Nicol, T., Zecker, S., & Kraus, N. (2008). Right-hemisphere auditory cortex is dominant for coding syllable patterns in speech. *The Journal of Neuroscience*, *28*, 3958–3965.

Aiken, S. J., & Picton, T. W. (2008). Human cortical responses to the speech envelope. *Ear and Hearing*, *29*, 139–157.

Backer, K. C., & Alain, C. (2012). Orienting attention to sound object representations attenuates change deafness. *Journal of Experimental Psychology: Human Perception and Performance*, *38*, 1554–1566.

Benjamini, Y., & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, *29*, 1165–1188.

Biermann, S., & Heil, P. (2000). Parallels between timing of onset responses of single neurons in cat and of evoked magnetic fields in human auditory cortex. *Journal of Neurophysiology*, *84*, 2426–2439.

Cervantes Constantino, F., Pinggera, L., Paranamana, S., Kashino, M., & Chait, M. (2012). Detection of appearing and disappearing objects in complex acoustic scenes. *PLoS One*, *7*, e46167.

Dau, T., Kollmeier, B., & Kohlrausch, A. (1997). Modeling auditory processing of amplitude modulation. I. Detection and masking with narrow-band carriers. *The Journal of the Acoustical Society of America*, *102*, 2892–2905.

Demany, L., & Ramos, C. (2005). On the binding of successive sounds: Perceiving shifts in nonperceived pitches. *The Journal of the Acoustical Society of America*, *117*, 833–841.

Demany, L., Semal, C., Cazalets, J. R., & Pressnitzer, D. (2010). Fundamental differences in change detection between vision and audition. *Experimental Brain Research*, *203*, 261–270.

D'Esposito, M., & Postle, B. R. (2015). The cognitive neuroscience of working memory. *Annual Review of Psychology*, *66*, 115–142.

Ding, N., & Simon, J. Z. (2012a). Neural coding of continuous speech in auditory cortex during monaural and dichotic listening. *Journal of Neurophysiology*, *107*, 78–89.

Ding, N., & Simon, J. Z. (2012b). Emergence of neural encoding of auditory objects while listening to competing speakers. *Proceedings of the National Academy of Sciences of the United States of America*, *109*, 11854–11859.

Dykstra, A. R., & Gutschalk, A. (2015). Does the mismatch negativity operate on a consciously accessible memory trace? *Science Advances*, *1*, e1500677.

Eramudugolla, R., Irvine, D. R., McAnally, K. I., Martin, R. L., & Mattingley, J. B. (2005). Directed attention eliminates "change deafness" in complex auditory scenes. *Current Biology*, *15*, 1108–1113.

Fontaine, B., Goodman, D. F. M., Benichoux, V., & Brette, R. (2011). Brian hears: Online auditory processing using vectorization over channels. *Frontiers in Neuroinformatics*, *5*, 1–9.

Gregg, M. K., & Samuel, A. G. (2008). Change deafness and the organizational properties of sounds. *Journal of Experimental Psychology: Human Perception and Performance*, *34*, 974–991.

Gregg, M. K., & Snyder, J. S. (2012). Enhanced sensory processing accompanies successful detection of change for real-world sounds. *NeuroImage*, *62*, 113–119.

Gutschalk, A., & Dykstra, A. R. (2014). Functional imaging of auditory scene analysis. *Hearing Research*, *307*, 98–110. https://doi.org/10.1016/j.heares.2013.08.003

Hertrich, I., Dietrich, S., Trouvain, J., Moos, A., & Ackermann, H. (2012). Magnetic brain activity phase-locked to the envelope, the syllable onsets, and the fundamental frequency of a perceived speech signal. *Psychophysiology*, *49*, 322–334.

Hillyard, S. A., Hink, R. F., Schwent, V. L., & Picton, T. W. (1973). Electrical signs of selective attention in the human brain. *Science*, *182*, 177–180.

Irsik, V. C., Vanden Bosch der Nederlanden, C. M., & Snyder, J. S. (2016). Broad attention to multiple individual objects may facilitate change detection with complex auditory scenes. *Journal of Experimental Psychology: Human Perception and Performance*, *42*, 1806–1817.

Kerlin, J. R., Shahin, A. J., & Miller, L. M. (2010). Attentional gain control of ongoing cortical speech representations in a "cocktail party". *The Journal of Neuroscience*, *30*, 620–628.

McAnally, K. I., Martin, R. L., Eramudugolla, R., Stuart, G. W., Irvine, D. R. F., & Mattingley, J. B. (2010). A dual-process account of auditory change detection. *Journal of Experimental Psychology: Human Perception and Performance*, *36*, 994–1004.

McDermott, J. H., Schemitsch, M., & Simoncelli, E. P. (2013). Summary statistics in auditory perception. *Nature Neuroscience*, *16*, 493–498.

McDermott, J. H., & Simoncelli, E. P. (2011). Sound texture perception via statistics of the auditory periphery: Evidence from sound synthesis. *Neuron*, *71*, 926–940.

Mesgarani, N., & Chang, E. F. (2012). Selective cortical representation of attended speaker in multi-talker speech perception. *Nature*, *485*, 233–236.

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

Pantev, C., Eulitz, C., Hampson, S., Ross, B., & Roberts, L. E. (1996). The auditory evoked "off" response: Sources and comparison with the "on" and the "sustained" responses. *Ear and Hearing*, *17*, 255–265.

Pavani, F., & Turatto, M. (2008). Change perception in complex auditory scenes. *Perception & Psychophysics*, *70*, 619–629.

Puschmann, S., Sandmann, P., Ahrens, J., Thorne, J., Weerda, R., Klump, G., … Thiel, C. M. (2013). Electrophysiological correlates of auditory change detection and change deafness in complex auditory scenes. *Neuroimage*, *75*, 155–164.

Shamma, S. A., Elhilali, M., Ma, L., Micheyl, C., Oxenham, A. J., Pressnitzer, D., … Yanbo, X. (2013). Temporal coherence and the streaming of complex sounds. In B. C. J. Moore, R. P. Carlyon, R. D. Patterson, & H. Gockel (Eds.), *Basic aspects of hearing: Physiology and perception* (pp. 535–544). New York, NY: Springer.

Sohoglu, E., & Chait, M. (2016). Neural dynamics of change detection in crowded acoustic scenes. *Neuroimage*, *126*, 164–172.

Sussman, E. S., Chen, S., Sussman-Fort, J., & Dinces, E. (2014). The five myths of MMN: Redefining how to use MMN in basic and clinical research. *Brain Topography*, *27*, 553–564.

Zion Golumbic, E. M., Ding, N., Bickel, S., Lakatos, P., Schevon, C. A., McKhann, G. M., … Schroeder, C. E. (2013). Mechanisms underlying selective neuronal tracking of attended speech at a "cocktail party". *Neuron*, *77*, 980–991.

## SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

**How to cite this article:** Starzynski C, Gutschalk A. Context-dependent role of selective attention for change detection in multi-speaker scenes. *Hum Brain Mapp*. 2018;1–10. https://doi.org/10.1002/hbm.24310