Effects of Cross-modal Asynchrony on Informational Masking in Human Cortex

Lars Hausfeld[1], Alexander Gutschalk[2], Elia Formisano[1], and Lars Riecke[1]

Abstract

■ In many everyday listening situations, an otherwise audible sound may go unnoticed amid multiple other sounds. This auditory phenomenon, called informational masking (IM), is sensitive to visual input and involves early (50–250 msec) activity in the auditory cortex (the so-called awareness-related negativity). It is still unclear whether and how the timing of visual input influences the neural correlates of IM in auditory cortex. To address this question, we obtained simultaneous behavioral and neural measures of IM from human listeners in the presence of a visual input stream and varied the asynchrony between the visual stream and the rhythmic auditory target stream (in-phase, antiphase, or random). Results show effects of cross-modal asynchrony on both target detectability (RT and sensitivity) and the awareness-related negativity measured with EEG, which were driven primarily by antiphasic audiovisual stimuli. The neural effect was limited to the interval shortly before listeners' behavioral report of the target. Our results indicate that the relative timing of visual input can influence the IM of a target sound in the human auditory cortex. They further show that this audiovisual influence occurs early during the perceptual buildup of the target sound. In summary, these findings provide novel insights into the interaction of IM and multisensory interaction in the human brain. ■

INTRODUCTION

Our everyday environment confronts us with complex mixtures of sounds originating from various sources (e.g., voices, music, traffic, rain). A proper behavioral reaction to ecologically important sounds requires us to analyze the auditory scene. In relatively quiet scenes, we may hear out individual auditory streams with ease, sometimes even without being aware of it (Sussman, Horvath, Winkler, & Orr, 2007). However, in very noisy scenes where background sounds partially mask the target stream, this streaming task is more complex and demanding.

Informational masking (IM) occurs when an otherwise audible sound goes unnoticed because of the simultaneous presence of spectrally nonoverlapping sounds. This perceptual phenomenon is often studied using a repetitive tone sequence (target) embedded in a cloud of random, spectrally nonoverlapping tones (masker). IM depends on various factors, including the number of masker tones (Sheft & Yost, 2008), the number of possible target frequencies (Kidd, Richards, Mason, Gallun, & Huang, 2008), and the acoustic similarity between masker and target (Kidd, Mason, & Arbogast, 2002). Compared with energetic masking, which is induced in the peripheral auditory system by overlapping neural responses to target and masker, the neural mechanisms and loci of IM are still poorly understood (Shinn-Cunningham, 2008).

Several brain studies have identified neural correlates of IM at early processing stages of the auditory cortex (AC). By comparing cortical magnetoencephalography (MEG) responses elicited by detected target tones versus undetected tones under matched acoustic conditions, Gutschalk, Micheyl, and Oxenham (2008) found a prominent negativity that likely originates from the AC during the 50–250 msec interval following the onsets of target tones (peaking at 150 msec) when the listener detects these tones. They referred to this late latency response as the awareness-related negativity (ARN). Because middle latency, steady-state cortical responses showed no such effect, the authors concluded that IM may emerge between early and late processing stages (50–250 msec) in AC. The early portion of the ARN (75–175 msec) shows a similar latency, polarity (Gutschalk et al., 2008), and sensory response features (Königs & Gutschalk, 2012) as the passive N1 response to suprathreshold tones without masker (Näätänen & Picton, 1987; Picton, Hillyard, Krausz, & Galambos, 1974; Vaughan & Ritter, 1970). Thus, it has been suggested that (the release from) IM arises from a similar processing stage in AC as the N1. Consistently, a related fMRI study using the aforementioned paradigm and statistical comparison (Wiegand & Gutschalk, 2012) demonstrated higher AC activity for detected (vs. undetected) targets under IM. A recent MEG study on IM using effective connectivity analysis (Giani, Belardinelli, Ortiz, Kleiner, & Noppeney, 2015) found that the ARN may arise from recursive processes within AC. The ARN in that study emerged only for the second of two tones, that is, one tone before participants' behavioral report of the target, whereas it was observed two tones before the behavioral report in a previous study using a

[1]Maastricht University, [2]Ruprecht-Karls-Universität Heidelberg

doi:10.1162/jocn_a_01097

series of 12 tones (Gutschalk et al., 2008). This led the authors to conclude that the ARN reflects the perception of an auditory stream rather than the perception of the single tones of the stream. In summary, these studies show that IM is represented at early processing stages in AC, not precluding the possibility that IM originates even earlier at subcortical stages.

An important aspect of IM in everyday listening situations that has not been addressed much so far is its sensitivity to visual input. Visual cues simultaneous to an auditory target can reduce IM at a behavioral level, especially when the target and masker cannot be segregated based on spatial auditory cues (Varghese, Ozmeral, Best, & Shinn-Cunningham, 2012; Helfer & Freyman, 2005). However, it remains unknown whether and how visual input influences neural correlates of IM. MEG studies on auditory streaming found that biasing effects of visual temporal cues on perceptual organization (i.e., auditory stream integration or segregation; Maddox, Atilgan, Bizley, & Lee, 2015; O'Leary & Rhodes, 1984) are reflected in the auditory MMN (Rahne & Bockmann-Barthel, 2009; Rahne, Bockmann, von Specht, & Sussman, 2007), a long latency (150–250 msec) auditory cortical response thought to index preattentive acoustic deviance detection (Näätänen, Tervaniemi, Sussman, Paavilainen, & Winkler, 2001). Under IM, the MMN is only observed when the listeners are aware of the standard stream (Dykstra & Gutschalk, 2015). On the basis of these studies and the notion that IM and auditory streaming involve similar mechanisms (Kidd, Mason, Deliwala, Woods, & Colburn, 1994), it is conceivable that cross-modal (temporal) information influences early auditory cortical correlates of IM.

To address this point and gain more insights into the relative timing of cortical processes for IM and multisensory interaction, we applied a random multitone masker paradigm, which we extended to include a visual stream, and simultaneously measured electroencephalography (EEG) in humans. We varied the temporal relation between the visual stream and a concurrent auditory target stream: Visual events either were presented with short onset asynchrony (in-phase), long onset asynchrony (antiphase), or occurred pseudorandomly relative to the rhythmic auditory stream. We refer to these three conditions as synchronous, alternating, and random, respectively. We hypothesized that synchronous stimuli would cause the visual stream and auditory target to bind together, leading to both a perceptual release from IM (i.e., improved auditory target detectability) and an enhancement of neural correlates of IM (i.e., increased ARN magnitude), whereas alternating stimuli would produce opposite effects. Results indeed show an effect of cross-modal asynchrony on auditory target detection under IM as predicted, but no enhancement of the ARN, despite reliable occurrence of the ARN in all experimental conditions. Instead, we observed an attenuating effect of synchronous cross-modal presentation on the ARN elicited by the earliest-detected target tones during the interval directly preceding the subject's behavioral report. On the basis of these findings, we revisit the cortical processes involved in the release from IM (as indexed by the ARN) and discuss their interaction with multi-sensory integration processes.

METHODS

Participants

Nineteen students of Maastricht University (nine women, age range = 18–26 years, mean age = 21.4 years, *SD* = 1.87 years) took part in the experiment. They received course credit or gift vouchers for their participation. They had normal or corrected-to-normal vision and normal hearing as assessed by pure tone audiometry (<25 dB HL at 0.25, 0.5, 0.75, 1, 2, 3, 4, and 6 kHz in both ears) and the Speech, Spatial and Qualities of Hearing Scale (SSQ12 score > 6.5; Noble, Jensen, Naylor, Bhullar, & Akeroyd, 2013). The local ethical committee of the Faculty of Psychology and Neuroscience (Ethische Commissie Psychologie) at Maastricht University approved the experimental procedures of the study.

Stimuli

The auditory stimuli were similar to those in the study of Gutschalk and colleagues (2008) and consisted of a pulsating tone (target) embedded in a multitone cloud (masker). The stimuli lasted 10.4 sec and comprised tones in 18 frequency bands with logarithmically spaced center frequencies between 239 and 5000 Hz (Figure 1A). All tones (i.e., tones belonging to the target or masker) had the same amplitude and lasted 100 msec including linear on-/off-ramps, which lasted 20 msec each. Targets were a sequence of 12 tones with fixed frequency (target frequency, either 489, 699, 1000, 1430, 2045, or 2924 Hz) and a fixed SOA of 800 msec as before (Gutschalk et al., 2008). The two frequency bands on each side of the target contained no tones. Maskers comprised several tones within the remaining 13 frequency bands. Masker tones had random frequency within one equivalent rectangular bandwidth (Glasberg & Moore, 1990) centered on the respective frequency band and random SOA within a 100–700 msec interval. The resulting average masker SOA was 400 msec, which falls well between previously used masker SOAs (Gutschalk et al., 2008). The onset of the masker preceded the onset of the target by 800 msec. In total, 48 of these auditory stimuli were composed by combining each of the six target frequencies with eight differently randomized maskers. In addition, to enable analysis within the framework of signal detection theory, 24 matching "no-target" stimuli were composed by combining each target frequency with four of the aforementioned maskers while excluding the target.

In the experiment, participants viewed a black central fixation cross on a gray background, shown on a PC screen. During the presentation of the auditory stimuli, the color

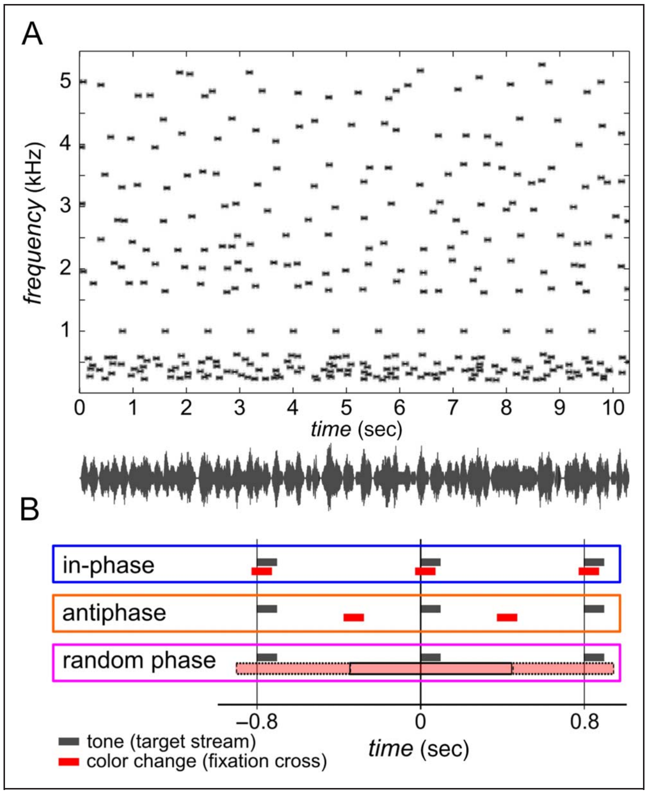

Figure 1. Auditory stimuli and experimental design. A shows the spectrogram (top plot) and the sound waveform (bottom plot) of an example auditory stimulus comprising a 1-kHz target. B illustrates the experimental conditions, which were defined by the delay (onset asynchrony) between the target tones and a color change of a visual fixation cross. In the in-phase condition (IP) and antiphase condition (AP), the color change preceded the target tones by a fixed interval of 25 and 425 msec, respectively. In contrast, in the random phase condition (RP), the color change occurred randomly within −350 to 350 msec relative to the target tones. Gray dashes represent intervals of target tones and red dashes represent intervals of color changes.

of the cross switched occasionally to red. These brief color changes (duration = 100 msec) occurred 12 times with one of three timings relative to the onsets of the individual target tones (or "virtual" tones, for the no-target stimuli). As illustrated in Figure 1B, this relative timing (cross-modal onset asynchrony) defined three experimental conditions: In the in-phase condition (IP) and antiphase condition (AP), the color changes preceded the target tones by a fixed interval of 25 and 425 msec, respectively, whereas in the random phase condition (RP), they occurred randomly within −350 to 350 msec relative to the target tones. The audio-visual lag of 25 msec in condition IP was introduced because audio-visual lags of 20–30 msec have been shown to evoke the highest number of audio-visual simultaneity judgements in a previous multisensory integration study (Zampini, Guest, Shore, & Spence, 2005). Each of four runs of the experiment involved 72 trials (i.e., 24 trials of each experimental condition). Thus, in total data from 216 trials comprising 144 targets and 72 nontargets were collected (corresponding to full presentations of the aforementioned set of 48 targets and 24 nontargets for each of the three experimental conditions). The trial order was fully randomized and balanced for each of the four runs, which were separated by short breaks. Trials began with the presentation of the black fixation cross followed by the presentation of task stimuli and a short rest interval (intertrial interval) of 2 sec.

Auditory stimuli were digitized using a sampling rate of 44.1 kHz and 16 bits. They were presented via a soundcard (Sound Blaster X-Fi Xtreme Audio, Creative Technology Ltd., Singapore), an audio amplifier (AB 200, AB International, Loomis, CA), and two speakers (Control 25, JBL Professional, Northridge, CA) located 1.3 m in front and symmetrically with respect to the participant (60° angle in azimuth) at a comfortable listening level of 60 dB SPL. Example auditory stimuli can be downloaded here: http://dx.doi.org/10.7910/DVN/NIGHLY.

Procedure

After obtaining written informed consent, audiometry, and SSQ12 from the participants, they were familiarized with the auditory task by presenting them with exemplary auditory stimuli comprising a target and no masker. They were instructed to fixate the fixation cross, detect a regularly repeating tone pip, and report the presence of this target as quickly as possible by pressing a button. They were informed that some stimuli contained no target. Finally, they were trained on 18 trials of the task involving 12 target stimuli, 6 no-target stimuli, and no color changes.

EEG Recording

EEG was recorded from 30 scalp electrodes positioned according to a modified 10–20% system (EasyCap; Electro-Cap, Inc., Eaton, OH) and referenced to linked mastoids, using BrainAmp amplifiers (Brain Products, Munich, Germany). Vertical and horizontal EOG was recorded from electrodes placed below and next to the right eye, respectively. Impedances were kept below 5 kΩ. EEG recordings were bandpass-filtered (cutoffs: 0.01 and 124 Hz, analog filter) and digitized with a 250-Hz sampling rate.

Behavioral Data Analysis

Each trial was classified as Hit, Miss, False Alarm, or Correct Rejection depending on the presence of a target and the participant's response. Hits and False Alarms reported before the second target tone (i.e., <1600 msec) were considered as guesses because at least two consecutive tones were necessary to identify the target (Gutschalk et al., 2008; on average 1.5 trials [i.e., 0.7%] were excluded per participant [across participants 6, 9, and 10 trials were excluded for IP, AP, and RP, respectively]). For these early guesses, data to all tones of the respective 10.4-sec stimulus were excluded from further analysis. Detection performance was assessed using the sensitivity index d', computed as the difference between the *z*-transformed hit rates and false alarm rates (Macmillan & Creelman, 1991). To assess changes in performance over time, this measure was also extracted separately for each tone position within the target, that is, for each individual target tone (or "virtual" tone, for the no-target stimuli), except for the first tone because at least two consecutive tones were necessary to identify the target. Finally, RT was assessed for each Hit as the interval between the onset of the second target tone (i.e., 1600 msec after stimulus onset) and button press.

EEG Data Analysis

EEG Data Preprocessing

EEG data were preprocessed and analyzed using the EEGLAB toolbox and custom Matlab scripts. Data preprocessing involved band-pass filtering (cutoffs: 0.5 and 30 Hz, FIR filter), re-referencing to an average reference, and epoching from −0.25 to 10.4 sec relative to auditory stimulus onset. To enable artifact reduction, the channel waveforms from each participant were decomposed into maximally temporally independent components (ICs). This was done using the ErpICASSO algorithm (Artoni et al., 2012; Himber, Hyvärinen, & Esposito, 2004), which determines these ICs based on the ICA results of epoch-wise bootstrapped data and subsequent component clustering. The initial ICA results were obtained by applying the FastICA algorithm (Hyvärinen & Oja, 2000) to 50 bootstraps, using a symmetric decorrelation approach, Gaussian nonlinearity, and a stopping criterion of $\varepsilon = 10^{-6}$. ICs not resembling brain-related activity (e.g., EOG or ECG) were identified based on visual inspection of weight topography and waveform and, in case of EOG, in addition by high correlations between EOG and EEG waveform. Finally, nonartifactual ICs (22.05 ± 2.74 ICs across participants [mean ± *SD*]) were recomposed and back-projected to yield artifact-reduced EEG channel waveforms.

Next, EEG trials were further epoched from −40 msec until 750 msec relative to the individual tone positions, that is, the onset times of individual target tones (or "virtual" tones, for the no-target stimuli). Each of the resulting epochs was then classified as Hit, Miss, Correct Rejection, or False Alarm as before (see Behavioral Data Analysis). Because of a relatively small number of false alarms (5.93 ± 5.82 across conditions [mean ± *SD*]; Figure 2), we restricted the EEG analysis to the three other trial types. Trials comprising residual artifacts not captured by the artifactual ICs were removed via autoadaptive averaging (Talsma, 2008) separately for each condition and response type (on average 16.68 ± 1.24% [mean ± *SD*] of trials were rejected across conditions and response types).

To identify the ARN, ERPs were extracted from a "whole-target" analysis as before (Gutschalk et al., 2008): Trial responses for Hits were derived by averaging the epochs including the two tones preceding the behavioral response and all tones after that response. For Misses and Correct Rejections, trial responses were computed by including all epochs starting with the interval of the second tone for averaging. The final ERPs for Hits, Misses, and Correct Rejections were obtained by averaging these trial responses across trials. Excluding the two tones before the behavioral response from this analysis did not change the results qualitatively.

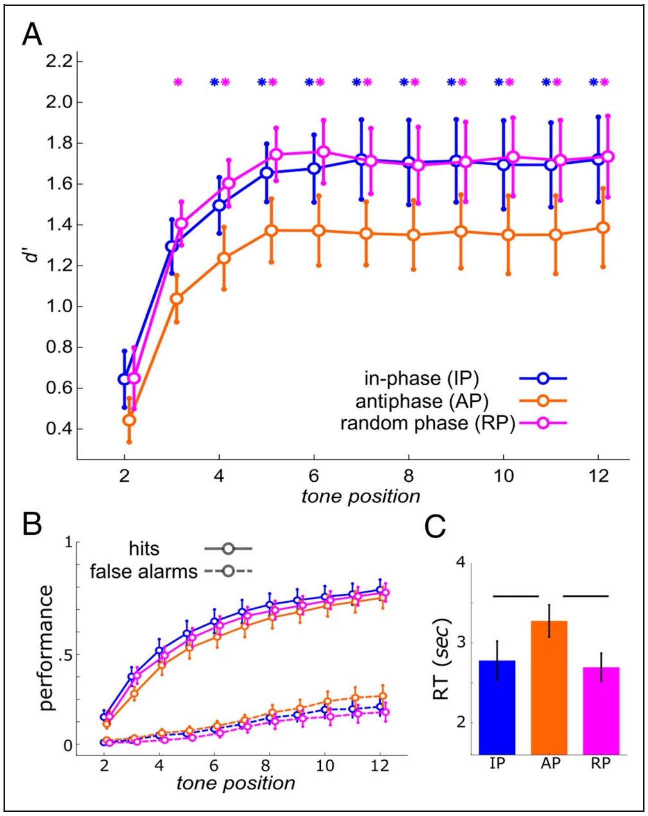

Figure 2. Behavioral results. A shows participants' average target detection performance assessed with d' as a function of the position of the tone within the target sequence, separately for each experimental condition. The dots above the curves indicate tone positions at which performance in IP condition (blue) and RP condition (magenta) was significantly higher than in the AP condition (corrected for multiple comparison with false discovery rate [FDR; Benjamini & Hochberg, 1995]: $q = .05$). Error bars represent *SEM* across participants. B is analogous to A but shows hit rate and false alarm rate instead of d'. RT is depicted in C, lines denote RT differences ($p < .05$).

To assess the emergence of the ARN, that is, its temporal evolution across individual tone positions within the target before the behavioral response, a subsequent "tone-resolved" analysis was conducted. In this single-tone analysis, ERPs were computed by first labeling epochs within each Hit trial according to their position relative to the button press and then averaging epochs with matching (behavioral response-related) positions across all Hit trials. This was done for three positions, that is, the three tones immediately preceding the behavioral response.

Spatial Filtering

As previous findings have indicated an auditory cortical origin of the ARN (Giani et al., 2015; Wiegand & Gutschalk, 2012; Gutschalk et al., 2008), we focused the EEG analysis on neural activity supposedly originating from auditory cortical sources. To emphasize auditory-evoked cortical activity, we estimated a spatial filter based on the N1, an ERP component showing similar properties as the ARN (see Introduction). First, the ERP to the onset of the auditory stimuli was computed and the resulting sound-evoked ERP was averaged across channels FC1, FC2, and Cz. Second, from this channel-averaged ERP, the N1 was identified as the largest negativity within a 75–200 msec interval, and its peak latency was extracted. Third, ERP magnitude at the observed N1 peak latency (on average 114 ± 31.1 msec [mean ± *SD*]) was extracted for each EEG channel to obtain a map of N1 magnitudes. These steps were done separately for each participant. Finally, the resulting individual N1 magnitude maps were averaged and normalized according to Euclidean norm to obtain an average map of spatial filter weights. The resulting filter was applied (i.e., by computing the linear sum of the weighted channel waveforms) in all subsequent analyses. Importantly, data segments for filter estimation did not overlap with data for further analysis and, thus, did not introduce dependency in the data.

Definition of the ARN

Previous studies plotted the ARN for Hits in reference to the baseline (Giani et al., 2015; Gutschalk et al., 2008) and compared Hit and Miss trials in a subsequent statistical analysis. Here, we directly plotted the ARN as difference wave between Hit minus Miss trials, mainly to subtract out the visual evoked response in the two conditions with a fixed phase relationship. Although this reduces the signal-to-noise ratio by about $\sqrt{2}$ compared with comparing only Hit trials, this step is necessary to compare the ARN amplitude between conditions.

Statistical Analysis

Cluster-based Nonparametric Test

To assess the reliability of the ARN and identify its timing after the onset of the detected tones, the statistical significance of the difference Hit versus Miss was first assessed at each time bin of the whole-interval ERP, using a cluster-based permutation test (Maris & Oostenveld, 2007). For each condition, an empirical null distribution of cluster sizes was created, that is, the number of significant consecutive time bins that can be observed by chance given the data (Bullmore et al., 1999). Permutations were created by switching labels of Hits and Misses at the participant level, leading to 2^{15} different sets and their reversed counterparts resulting in total in $2^{16}-2$ permutations (the true labeling and its reverse were not used). For each of the permutation sets, the maximum cluster size of significant time bins was derived (two-sided paired t test with a criterion $p < .025$). Subsequently, the size of observed clusters of significant consecutive bins with true labels are compared with the distribution of maximum cluster sizes under the null hypothesis, which results in a probability estimate (i.e., the number of instances of the permutation distribution with clusters larger than the observed cluster) that is corrected for multiple comparisons. Clusters with a probability of $p < .05$ were labeled as significant and selected for further interval-of-interest analysis. Following this initial ARN identification, ARN magnitude differences between the experimental conditions were assessed using paired t tests. In addition, an exploratory analysis using the same cluster-based nonparametric analysis over the whole trial interval was applied to detect potential differences outside the predefined ARN interval.

After this whole-interval analysis (involving ERPs averaged across all detected tones, see EEG data preprocessing section), a subsequent analysis assessed the emergence of the ARN across individual target tones before the listener's button press and putative cross-modal differences in this emergence. This single-tone analysis involved applying the interval-of-interest analysis to single-tone ERPs.

RESULTS

Behavioral Results

Three participants showed insufficient target detection performance ($d' < 0.12$); consequently their data were excluded from the analysis. The remaining 16 participants performed at an intermediate level ($d' > 0.5$; mean ± *SEM* d': 1.683 ± 0.20). Figure 2A and 2B illustrates the behavioral results from tone-resolved analyses (i.e., for each tone position). Consistent with previous findings (Gutschalk et al., 2008), participants' performance as assessed with d' built up over the duration of the target, reaching ceiling after approximately five tones. Statistical comparisons between the experimental conditions revealed that overall performance was lower in the AP condition than the IP and RP conditions (see Table 1). A one-way repeated-measures ANOVA with factor Condition (IP, AP, RP) showed a main effect of Condition on d' ($F(2, 30) = 6.3728$, $p = .005$). Post hoc tests showed statistically lower performance for AP than for IP or RP (IP vs. AP: $t(15) = 2.861, p = .012$; RP vs. AP: $t(15) = 3.762, p = .002$), and no significant difference between IP and RP ($t(15) = -0.117$, *ns*). In line with these d' results, analysis of RTs (Figure 2C; Table 1) for Hit trials revealed a main effect of Condition ($F(2, 30) = 4.080, p = .027$) and longer RTs for AP than IP or RP (IP vs. AP: $t(15) = -2.503, p = .024$; RP vs. AP: $t(15) = -2.384$, $p = .031$; $q < .05$), but no difference between IP and RP ($t(15) = -0.162$, *ns*). Applying the same analysis to hit rates and false alarm rates yielded no main effect of Condition (hit rate: $F(2, 30) = 1.940, p = .163$; false alarm rate: $F(2, 30) = 2.558, p = .094$). In summary, these results show that visual stimuli can modulate auditory target detection under IM, depending on their cross-modal asynchrony with the auditory target.

Table 1. Behavioral Data: Descriptive Statistics

	Hit Rate		*False Alarm Rate*		*d′*		*RT (sec)*	
	Mean	*SEM*	*Mean*	*SEM*	*Mean*	*SEM*	*Mean*	*SEM*
IP	0.689	0.045	0.167	0.043	1.721	0.208	2.561	0.221
AP	0.653	0.047	0.215	0.047	1.387	0.192	2.934	0.191
RP	0.676	0.040	0.143	0.041	1.735	0.199	2.585	0.182

The table summarizes the behavioral results for each experimental condition (IP, AP and RP). The mean and *SEM* are shown for each behavioral measure (hit rate, false alarm rate, *d′*, and RT).

EEG Results

The topography of the applied spatial filter (see inset Figure 3A) matched well the topography of the auditory-evoked N1 (Näätänen & Picton, 1987; Picton et al., 1974; Vaughan & Ritter, 1970) as expected.

Whole-interval ERPs

Figure 3A–C show results from whole-interval ERP analysis, which served to identify the ARN based on all detected tones. The ARN could be reliably observed in all experimental conditions (IP, AP, and RP). In line with previous findings from purely auditory MEG studies (Giani et al., 2015; Königs & Gutschalk, 2012; Gutschalk et al., 2008), cluster-based permutation tests on ARN magnitude (i.e., Hits–Misses difference curves) for each condition revealed consistently two significant ARN intervals (the exact intervals are listed in Table 2). These intervals were replicable across various alternative analyses (e.g., when restricting the data analysis to channels FC1, FC2, and Cz instead of applying the obtained spatial filter, or when defining the ARN based on hit trials alone rather than the difference hits vs. misses). Because the interval identified in the RP condition between 156 and 312 msec matched well the previously reported “late” ARN interval (175–275 msec; Gutschalk et al., 2008), it was selected for further analysis for cross-modal asynchrony effects (i.e., IP vs. AP). Conditions with fixed audio-visual timing (IP and AP) showed additional visual color change-evoked potentials, reflecting the fact that epochs were time-locked to tone onsets and color changes and therefore had fixed latency with respect to these tone onsets in IP and AP condition. Noteworthy, these visual-evoked responses were more prominent than the auditory target-evoked responses (Figure 3C, compare peaks at ~550 msec vs. ~150 msec), which likely reflects a difference in stimulus saliency induced by the fact that target tones, but not visual stimuli, were embedded in a masker.

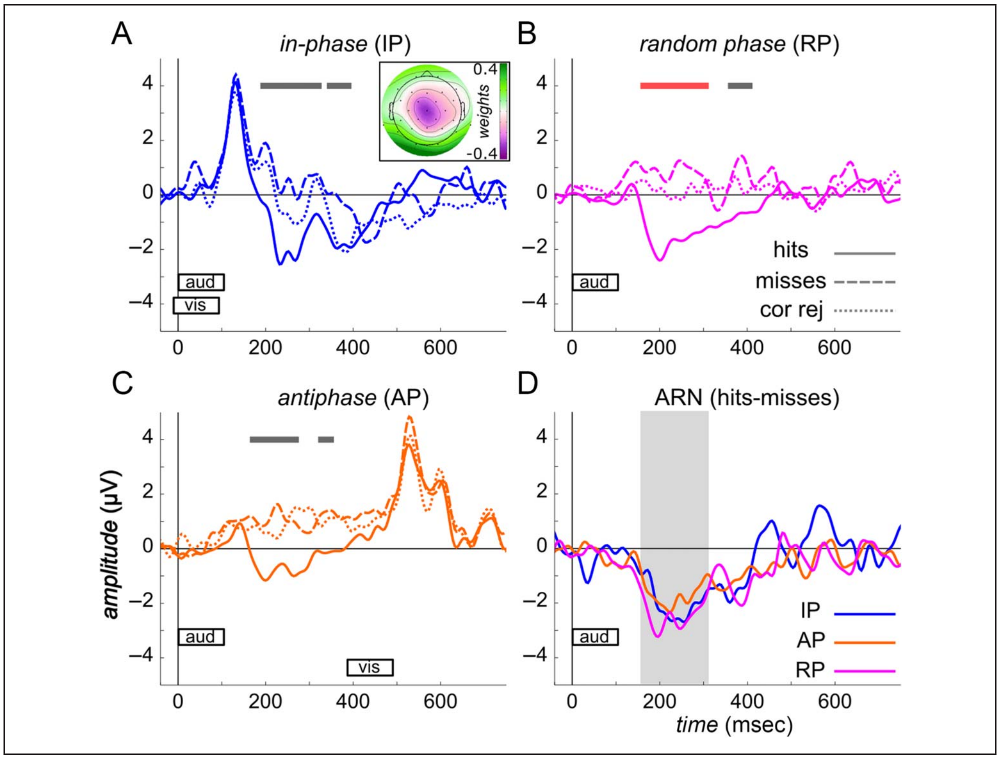

Figure 3. EEG results from analysis of whole-interval ERPs. A shows participants’ average whole-interval ERP associated with Hits (solid line), Misses (dashed line), and Correct Rejections (dotted line) in the IP condition. The inset shows the topography of the spatial filter estimated to extract auditory cortical processes. Gray horizontal bars delimit intervals during which a significant ARN (Hit > Miss, $p < .05$) was observed. The small rectangles represent the intervals of target tones (aud) and visual color changes (vis). Analogously, B and C show data from the RP and AP condition, respectively. D provides a summary of all panels, showing the Hit-minus-Miss difference (ARN) waveform for each condition. The red bar in B and the gray-filled rectangle in D indicate the interval of interest that was selected for further analysis.

Table 2. Identified ARN Intervals

	Interval 1 (msec)	*Interval 2 (msec)*	*Cluster p Interval 1*	*Cluster p Interval 2*
IP	188–328	340–396	.0001	.0056
AP	164–276	320–356	.0001	.0225
RP	156–312	356–412	.0010	.0162

The table shows the two intervals during which a significant ARN was observed and the associated statistical significance values, separately for each condition (IP, AP, and RP). Interval 1 in the RP condition was selected for further interval-of-interest analyses.

Figure 3D illustrates the ARN (Hits–Misses difference) waveform for each experimental condition. Analysis per time bin revealed no significant difference between any pair of conditions (cluster-based permutation test on ARN differences, $p > .10$). Similar outcomes were obtained from interval-of-interest analyses, which revealed no significant difference between IP and AP, neither in terms of ARN magnitude ($t(15) = -0.895, p = .385$; mean ± *SEM* amplitude for IP: −4.37 ± 0.52 μV, AP: −3.68 ± 0.62 μV) nor ARN peak latency ($t(15) = 0.044, p > .5$; mean ± *SEM* peak latency for IP: 241 ± 9 msec, AP: 243 ± 11 msec). In summary, these results suggest that cross-modal asynchrony does not influence the ARN after the release from IM (i.e., after listeners' report of the target tones), in contradiction with our hypothesis.

Single-tone ERPs

A possible explanation for the aforementioned null result is that we applied the ARN analyses to an average measure capturing an entire sequence of detected target tones. Considering that previous studies showed that the ARN builds up before the behavioral response indicating target detection (Giani et al., 2015; Gutschalk et al., 2008), it remains possible that the hypothesized cross-modal effect on ARN is limited to the interval before the target is first detected. Considering further that cross-modal asynchrony can influence both RTs to an auditory target under IM (present study, Figure 2) and behavioral measures of auditory streaming (see Introduction), it is conceivable that cross-modal asynchrony could influence only the initial release from IM, rather than the listeners' ongoing percept of this target. To test this idea, we extracted single-tone ERPs (see EEG Data Preprocessing section) and repeated the ARN analysis at various tone positions preceding the first detected target tone (i.e., before the button press).

Figure 4 illustrates the results from this analysis for three consecutive tone positions (−2 to 0, relative to the listener's behavioral report of the target, where position 0 corresponds to the tone immediately preceding the button press). In line with previous findings (Giani et al., 2015; Gutschalk et al., 2008), the ARN showed slight variations across these positions in each condition (IP, AP, RP). Results from interval-of-interest analysis in Figure 4B show

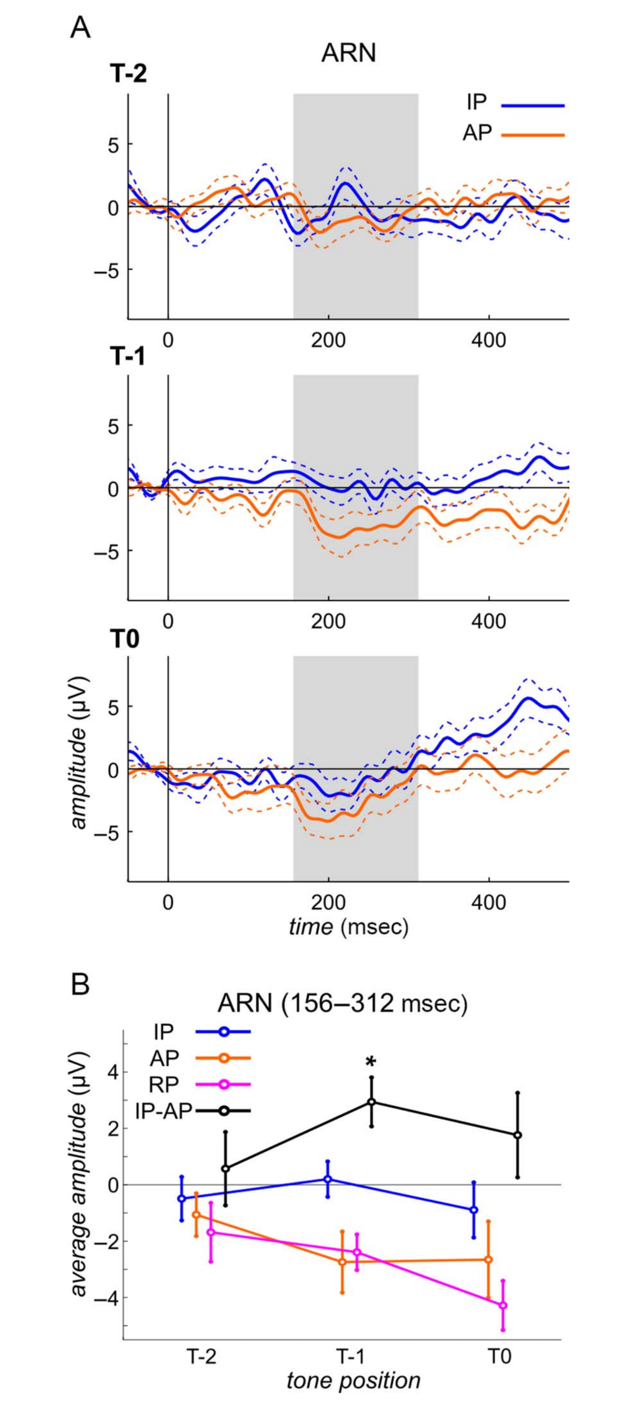

Figure 4. EEG results from analysis of single-tone ERPs. A shows participants' average single-tone ARN (ERP associated with Hits minus Misses) in the IP and AP condition (solid blue and orange waveforms), separately for three different tone positions (−2 to 0) relative to the listener's report of the target, where position 0 corresponds to the last tone preceding the listener's report. The dotted waveforms represent *SEM* across participants and the filled gray rectangle represents the interval of interest. B shows average ARN amplitude within the interval of interest as a function of tone position, separately for the IP condition (blue), AP condition (orange), the difference IP minus AP (cross-model asynchrony effect, black) and—for comparison—the RP condition (pink). Error bars represent *SEM* across participants, and the asterisk indicates the tone position for which a significant cross-model asynchrony effect was observed (FDR-corrected, $q < .05$).

that the ARN emerged shortly before the listeners' report of the target, starting at tone position 0 for condition IP and at position −1 for conditions AP and RP. These observations suggest that the ARN in the AP and RP conditions built up earlier before behaviorally indicated target detection than in the IP condition. A repeated-measures ANOVA with factors Condition (IP, AP) and Tone Position (−2 to 0) partially confirmed this notion, revealing a main effect of Condition ($F(1, 15) = 4.675, p = .047$) but no main effect of Tone Position ($F(2, 30) = 0.561, p = .577$) and no Condition × Tone interaction ($F(2, 30) = 1.039, p = .366$). Post hoc tests showed a significant effect of cross-modal asynchrony exclusively for the tone immediately preceding the listener's report of the target (position T-1: $t(15) = 3.389, p = .004$; two-sided paired t test, FDR-corrected $q < .05$). Data from the RP condition, which had been used for defining the time window of interest, were excluded from these statistical tests to avoid potential bias. Together with the null result from whole-interval analysis, the observed cross-modal effect on the ARN to early-detected tones indicates that visual stimuli modulate the ARN primarily during the initial release from IM, depending on their cross-modal asynchrony with the auditory target.

Combined analysis of the observed cross-modal asynchrony effects on brain response (ARN to early-detected tones) and behavioral response (RT, d', hit rate, false alarm rate) revealed no significant correlation (a trend was found when correlating the difference in ARN between IP and RP with respective RT differences for T-1: $r = .473, p = .065$, uncorrected).

DISCUSSION

We found that the relative timing between a visual stream and a rhythmic auditory target embedded in a multitone masker influences fundamental aspects of IM at both the behavioral and neural level. Visual stimuli that alternate (vs. are synchronous) with the auditory target hamper release from IM (as shown by lower sensitivity and longer RT in the AP vs. IP condition) and modulate IM-related auditory cortical potentials shortly before the listener's report of the target (as shown by larger ARNs during this interval in the AP vs. IP condition). These results are in line with previous behavioral findings showing influences of visual input on IM and reveal that these influences occur in the AC during the initial release of IM. They further demonstrate that cortical correlates of IM (specifically the ARN) can be obtained reliably under various audiovisual conditions and using EEG.

Cross-modal Antiphase Impedes Release from IM

The behavioral results show that visual stimuli that alternate (AP) with an auditory target in an informational masker can reduce or delay (~500 msec) the detection of this target, compared with visual stimuli that are synchronous (IP) or random (RP) relative to it, in line with previous cross-modal streaming studies (Maddox et al., 2015; O'Leary & Rhodes, 1984). Unexpectedly, we found no significant difference between synchronous versus random conditions (IP vs. RP), although hit rates differed qualitatively in the predicted direction. Given that listeners performed no overt task on the visual stimulus (besides fixating their gaze on it) and could readily infer that this stimulus was not a reliable predictor for the occurrence of target tones (cross-modal asynchrony varied unpredictably across trials), it is implausible that listeners attempted to exploit the visual stimulus when performing the auditory task. More plausibly, our null result may reflect that listeners intentionally paid no or only little endogenous (top–down) attention to the visual modality but focused on the auditory input (Crosse, Butler, & Lalor, 2015). In addition, benefits for the synchronous condition (IP) were possibly reduced in some listeners due to the fixed audiovisual delay (auditory lag of 25 msec), which did not take into account possible differences in listeners' optimal delay for audiovisual integration (e.g., Zampini et al., 2005). On the basis of these considerations, our behavioral results indicate that listeners could not benefit much from the temporal information provided by the visual rhythm in the IP condition but were distracted by it in the alternating condition (AP) in a bottom–up manner.

Cross-modal Asynchrony Modulates the Auditory Cortical Representation of IM

Using a whole-target ERP analysis, we found no significant effect of cross-modal asynchrony on the ARN evoked by a sequence of detected target tones. In contradiction with our initial hypothesis, this null result indicates that the ARN during an ongoing auditory target—once this target stream is identified—may be relatively insensitive to temporal changes in concurrent visual input, suggesting that multisensory interaction and IM do not interfere much in AC after the listener extracted the auditory target from the acoustic input.

However, by focusing the analysis on ARNs elicited by the earliest-detected tones (single-tone ERP analysis), we could find a significant effect of cross-modal asynchrony during the interval immediately preceding the listener's behavioral report of the target. This outcome shows that visual temporal information in fact can influence the ARN elicited by the first-detected target tones; specifically, stimuli with alternating audio-visual presentation (AP) show larger amplitudes during these early ARNs. Together with the observation that the ARN built up during the aforementioned preresponse interval (Figure 4B), our results indicate that multisensory interaction can interfere with IM in AC during the initial release from IM, suggesting that multisensory interaction in our paradigm primarily influenced the early detection and attentional selection of the auditory target in AC. Our interpretation does not exclude the possibility that later-occurring highly salient visual stimuli may still interfere with an established target stream percept in AC, although this unlikely occurred in our experiment (informal listening tests suggested that an established percept of the target stream could be easily upheld, and the whole-interval ERP analysis provided no significant result).

Although our study was designed to study visual influences on IM using the ARN as a cortical index of IM, rather than the ARN itself, our results from the whole-target analysis demonstrate that the ARN can be reliably observed across a variety of audio-visual stimulus conditions (AP, IP, and RP) using EEG. These results replicate previous findings, which have been limited to MEG so far (Giani et al., 2015; Königs & Gutschalk, 2012; Gutschalk et al., 2008). Thus, the ARN turns out to be a robust neural phenomenon that can be accessed via various noninvasive neuroelectromagnetic methods and observed under various audio-visual conditions, making it a useful marker for future studies of the neural mechanisms underlying IM.

Possible Explanation for the Observed Behavioral and Neural Effects

A possible interpretation of the observed pattern of behavioral and neural results is that cross-modal asynchrony modulates the amount of the top–down processing resources in AC that are required for the listener to release the auditory target from IM. We explain this idea as follows: Listeners in our study likely performed the auditory task by trying to identify and accumulate sensory evidence for a repetitive auditory pattern (i.e., periodic auditory-evoked responses in AC) and focusing their attention on this auditory-evoked rhythmic evidence (Elhilali, Xiang, Shamma, & Simon, 2009). This notion is supported by the facts that we instructed our listeners to detect a repetitive tone and that our listeners could detect this tone on average only after its fourth repetition (average RT with respect to first presentation: ~3.5 sec, equivalent to 4.3 tone cycles).

It is further conceivable that this listening strategy could have been sensitive to visual temporal input: considering that salient rhythmic visual input can shift the timing of ongoing rhythmic activity in AC and thereby support interaction with synchronous auditory input (Lakatos, Chen, O'Connell, Mills, & Schroeder, 2007), rhythmic visual stimulation (IP and AP) likely biased listeners' temporal attention "bottom–up" toward the visually evoked rhythm in AC. This notion is supported by both our behavioral (Figure 2B) and neural data (Figure 3A–C): In condition IP, visual stimulation provided valid cues regarding the target rhythm on trials where the target was present. Accordingly, we observed concurrent visual- and auditory-evoked periodic responses in AC on such trials and higher false alarm rates and hit rates in this condition compared with condition RP, although the latter behavioral differences did not reach statistical significance. In summary, these observations fit with the idea that the visual-evoked rhythm induced a temporal bias in AC.

Conversely, in condition AP, the visual input provided no valid cue, regardless of the presence of the target. Accordingly, we observed temporally distinct visual- and auditory-evoked periodic responses in AC, higher false alarm rates, and lower hit rates in this condition compared with condition RP, although the behavioral differences again did not reach statistical significance. Moreover, we observed the longest RTs in this condition, suggesting that this condition required listeners to overcome the conflicting visually induced bottom–up temporal bias and redirect attention in a top–down manner toward the alternating auditory-evoked rhythm. This likely involved additional processes in AC including endogenous attention, which has been shown to align the timing of ongoing rhythmic activity in AC to a sensory target rhythm via oscillatory phase reset (Lakatos et al., 2009).

Finally, in condition RP, visual stimulation was irregular and provided no valid cue. Accordingly, we observed no visually evoked periodic auditory cortical responses in this condition as expected, suggesting that listeners identified the auditory-evoked rhythm in the absence of visually induced bottom–up temporal bias. RTs were not significantly longer than in condition IP, possibly because an evoked rhythm in condition RP always constituted a valid cue, whereas in condition IP listeners still had to verify whether such an evoked rhythm actually resembled the auditory target (vs. visual input alone, as on no-target trials).

In summary, it seems that the visual rhythm and listeners' task goal biased temporal attention in bottom–up and top–down fashion, respectively, presumably by phase-shifting ongoing neuronal oscillations in AC. This defined periodic "attentional" time windows (i.e., oscillatory phases of increased neural excitability) during which sensory input was processed more effectively. The visually induced bottom–up attentional bias was either favorable (condition IP) or detrimental (condition AP) to auditory task performance. The task goal-induced top–down attentional bias was inversely related to how well the visually evoked rhythm tagged the auditory-evoked target rhythm, that is, it was strongest when listeners were biased away from the auditory target (AP). Consistent with this notion of a stronger top–down processing for alternating cross-modal stimuli, our neural data show a larger ARN in condition AP versus IP in the interval immediately preceding the listener's report of the auditory target (Figure 4B). In this view, temporal shifts in top–down attention may modulate the buildup of the ARN. Because the current data do not allow fully separating ongoing neuronal oscillations from sensory-evoked rhythm, more research is needed to enable verifying whether oscillatory phase shifts indeed underlie our results.

An alternative but not mutually exclusive view is that the decision regarding the presence of the auditory target is based on matching the visual rhythm to the accumulated evidence for the auditory rhythm. If the two are in phase, then a fast decision can be taken based on their synchrony with less evidence regarding the auditory rhythm. If the two are in antiphase, then the decision is delayed because further auditory evidence is required as basis for the perceptual decision and might be reflected in the larger ARN in condition AP. In this view, the magnitude of the ARN reflects the amount of accumulated auditory evidence that is available to the listener for making a perceptual decision while the listener's criterion for taking this decision is under top–down control and may additionally interact with the salient visual cue.

Potential Limitations

Our interpretation requires a few cautionary remarks. First, it should be noted that our interpretation regarding neural processes in AC presumes that the applied spatial EEG filter successfully extracted AC activity (Figure 3A). Although the spatial specificity of EEG compared with MEG is inherently limited, this assumption seems justified given the fact that the ARN we observed using this filter is highly similar to ARNs observed in previous MEG studies extracting AC activity with other source localization approaches (Giani et al., 2015; Gutschalk et al., 2008). Second, because the ARN is defined based on a neural response to a behaviorally detected target tone, the absolute tone position (within the overall stimulus) from which the ARN is extracted can covary with the latency of the listener's behavioral response. Consequently, ARNs associated with different RTs can be associated with different numbers of preceding tones and differences in neural adaptation. However, these potential differences unlikely confounded our ARN results: The observed RT difference between our experimental conditions was 373 msec or less (see Table 1), which falls well below half of our target tone SOA, implying that we extracted the ARN on average from approximately the same absolute tone position in different conditions.

Conclusions

Our study indicates that visual input can interact with the neural representation of (the release from) IM in early AC. Specifically, visual input that alternates with the temporal pattern of an auditory target can hamper the release from IM by both distracting the listener and inducing need for top–down attention to enhance this representation.

Acknowledgments

We thank Shalina Görg for help with data acquisition. This work was supported by Maastricht University and the Netherlands Organization for Scientific Research (NWO; VICI grant 435-12-002 to E. F. and VENI Grant 451-11-014 to L. R.). L. H. and L. R. conceived the experiment. L. H. implemented the experiment and acquired the data. L. H. and L. R. analyzed the data. L. H., L. R., A. G., and E. F. wrote the manuscript.

Reprint requests should be sent to Lars Hausfeld, Faculty of Psychology and Neuroscience, Department of Cognitive Neuroscience, Maastricht University, Oxfordlaan 55, Maastricht, Netherlands, 6229 EV, or via e-mail: lars.hausfeld@maastrichtuniversity.nl.

REFERENCES

Artoni, F., Gemignani, A., Sebastiani, L., Bedini, R., Landi, A., & Menicucci, D. (2012). ErpICASSO: A tool for reliability estimates of independent components in EEG event-related analysis. In *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE* (pp. 368–371). New York: IEEE.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B, Methodological, 57,* 289–300.

Bullmore, E. T., Suckling, J., Overmeyer, S., Rabe-Hesketh, S., Taylor, E., & Brammer, M. J. (1999). Global, voxel, and cluster tests, by theory and permutation, for a difference between two groups of structural MR images of the brain. *IEEE Transactions on Medical Imaging, 18,* 32–42.

Crosse, M. J., Butler, J. S., & Lalor, E. C. (2015). Congruent visual speech enhances cortical entrainment to continuous auditory speech in noise-free conditions. *Journal of Neuroscience, 35,* 14195–14204.

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

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 Biology, 7,* e1000129.

Giani, A. S., Belardinelli, P., Ortiz, E., Kleiner, M., & Noppeney, U. (2015). Detecting tones in complex auditory scenes. *Neuroimage, 122,* 203–213.

Glasberg, B. R., & Moore, B. C. (1990). Derivation of auditory filter shapes from notched-noise data. *Hearing Research, 47,* 103–138.

Gutschalk, A., Micheyl, C., & Oxenham, A. J. (2008). Neural correlates of auditory perceptual awareness under informational masking. *PLoS Biology, 6,* e138.

Helfer, K. S., & Freyman, R. L. (2005). The role of visual speech cues in reducing energetic and informational masking. *Journal of the Acoustical Society of America, 117,* 842–849.

Himber, J., Hyvärinen, A., & Esposito, F. (2004). Validating the independent components of neuroimaging time series via clustering and visualization. *Neuroimage, 22,* 1214–1222.

Hyvärinen, A., & Oja, E. (2000). Independent component analysis: Algorithms and applications. *Neural Networks, 13,* 411–430.

Kidd, G., Jr., Mason, C. R., & Arbogast, T. L. (2002). Similarity, uncertainty, and masking in the identification of nonspeech auditory patterns. *Journal of the Acoustical Society of America, 111,* 1367–1376.

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.

Kidd, G., Jr., Richards, V. M., Mason, C. R., Gallun, F. J., & Huang, R. (2008). Informational masking increases the costs of monitoring multiple channels. *Journal of the Acoustical Society of America, 124,* EL223–EL229.

Königs, L., & Gutschalk, A. (2012). Functional lateralization in auditory cortex under informational masking and in silence. *The European Journal of Neuroscience, 36,* 3283–3290.

Lakatos, P., Chen, C. M., O'Connell, M. N., Mills, A., & Schroeder, C. E. (2007). Neuronal oscillations and multisensory interaction in primary auditory cortex. *Neuron, 53,* 279–292.

Lakatos, P., O'Connell, M. N., Barczak, A., Mills, A., Javitt, D. C., & Schroeder, C. E. (2009). The leading sense: Supramodal control of neurophysiological context by attention. *Neuron, 64,* 419–430.

Macmillan, N. A., & Creelman, C. D. (1991). *Detection theory: A user's guide*. Cambridge: Cambridge UP.

Maddox, R. K., Atilgan, H., Bizley, J. K., & Lee, A. K. (2015). Auditory selective attention is enhanced by a task-irrelevant temporally coherent visual stimulus in human listeners. *Elife, 4,* e04995.

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods, 164,* 177–190.

Näätänen, R., & Picton, T. (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., Tervaniemi, M., Sussman, E., Paavilainen, P., & Winkler, I. (2001). "Primitive intelligence" in the auditory cortex. *Trends in Neuroscience, 24,* 283–288.

Noble, W., Jensen, N. S., Naylor, G., Bhullar, N., & Akeroyd, M. A. (2013). A short form of the Speech, Spatial and Qualities of Hearing scale suitable for clinical use: The SSQ12. *International Journal of Audiology, 52,* 409–412.

O'Leary, A., & Rhodes, G. (1984). Cross-modal effects on visual and auditory object perception. *Perception & Psychophysics, 35,* 565–569.

Picton, T. W., Hillyard, S. A., Krausz, H. I., & Galambos, R. (1974). Human auditory evoked potentials. I. Evaluation of components. *Electroencephalography and Clinical Neurophysiology, 36,* 179–190.

Rahne, T., Bockmann, M., von Specht, H., & Sussman, E. S. (2007). Visual cues can modulate integration and segregation of objects in auditory scene analysis. *Brain Research, 1144,* 127–135.

Rahne, T., & Bockmann-Barthel, M. (2009). Visual cues release the temporal coherence of auditory objects in auditory scene analysis. *Brain Research, 1300,* 125–134.

Sheft, S., & Yost, W. A. (2008). Method-of-adjustment measures of informational masking between auditory streams. *Journal of the Acoustical Society of America, 124,* EL1–EL7.

Shinn-Cunningham, B. G. (2008). Object-based auditory and visual attention. *Trends in Cognitive Sciences, 12,* 182–186.

Sussman, E. S., Horvath, J., Winkler, I., & Orr, M. (2007). The role of attention in the formation of auditory streams. *Perception & Psychophysics, 69,* 136–152.

Talsma, D. (2008). Auto-adaptive averaging: Detecting in event-related potential data using a fully automated procedure. *Psychophysiology, 45,* 216–228.

Varghese, L. A., Ozmeral, E. J., Best, V., & Shinn-Cunningham, B. G. (2012). How visual cues for when to listen aid selective auditory attention. *Journal of the Association for Research in Otolaryngology, 13,* 359–368.

Vaughan, H. G., Jr., & Ritter, W. (1970). The sources of auditory evoked responses recorded from the human scalp. *Electroencephalography and Clinical Neurophysiology, 28,* 360–367.

Wiegand, K., & Gutschalk, A. (2012). Correlates of perceptual awareness in human primary auditory cortex revealed by an informational masking experiment. *Neuroimage, 61,* 62–69.

Zampini, M., Guest, S., Shore, D. I., & Spence, C. (2005). Audio-visual simultaneity judgments. *Perception & Psychophysics, 67,* 531–544.