JNeurosci
THE JOURNAL OF NEUROSCIENCE

Research Articles | Behavioral/Cognitive

Attention drives visual processing and audiovisual integration during multimodal communication

https://doi.org/10.1523/JNEUROSCI.0870-23.2023

Received: 9 May 2023
Revised: 20 December 2023
Accepted: 21 December 2023

Attention drives visual processing and audiovisual integration during multimodal communication

Authors and affiliations:

Noor Seijdel[1], Jan-Mathijs Schoffelen[2], Peter Hagoort[1,2] & Linda Drijvers[1,2]

[1] Max Planck Institute for Psycholinguistics, Nijmegen, the Netherlands

[2] Donders Institute for Brain, Cognition and Behaviour, Radboud University, Nijmegen, the Netherlands

Corresponding author:

Noor Seijdel

Neurobiology of Language Department - The Communicative Brain

Max Planck Institute for Psycholinguistics

Wundtlaan 1, 6525 XD Nijmegen, The Netherlands

E-mail: noor.seijdel@mpi.nl

Number of pages: 23

Number of figures: 6

Number of words: Abstract: 239; Introduction: 661; Discussion: 1712

Conflict of interest:

The authors declare no competing financial interests.

Acknowledgements:

This work was supported by a Minerva Fast Track Fellowship from the Max Planck Society awarded to LD.

Data and code availability:

Data and code to reproduce the analyses in this article are available at https://osf.io/p36xf/

Abstract

During communication in real-life settings, our brain often needs to integrate auditory and visual information, and at the same time actively focus on the relevant sources of information, while ignoring interference from irrelevant events. The interaction between integration and attention processes remains poorly understood. Here, we use rapid invisible frequency tagging (RIFT) and magnetoencephalography (MEG) to investigate how attention affects auditory and visual information processing and integration, during multimodal communication. We presented human participants (male and female) with videos of an actress uttering action verbs (auditory; tagged at 58 Hz) accompanied by two movie clips of hand gestures on both sides of fixation (attended stimulus tagged at 65 Hz; unattended stimulus tagged at 63 Hz). Integration difficulty was manipulated by a lower-order auditory factor (clear/degraded speech) and a higher-order visual semantic factor (matching/mismatching gesture). We observed an enhanced neural response to the attended visual information during degraded speech compared to clear speech. For the unattended information, the neural response to mismatching gestures was enhanced compared to matching gestures. Furthermore, signal power at the intermodulation frequencies of the frequency tags, indexing non-linear signal interactions, was enhanced in left frontotemporal and frontal regions. Focusing on LIFG (Left Inferior Frontal Gyrus), this enhancement was specific for the attended information, for those trials that benefitted from integration with a matching gesture. Together, our results suggest that attention modulates audiovisual processing and interaction, depending on the congruence and quality of the sensory input.

Significance statement

This research advances our understanding of how attention influences the processing and integration of auditory and visual information during multimodal stimulus presentation. By utilizing Rapid Invisible Frequency Tagging (RIFT) and magnetoencephalography (MEG), the study offers novel insights into the neural activity and interactions between attended and unattended stimuli within a controlled experimental setting. Our findings reveal that attention modulates audiovisual processing and interaction, contingent on the congruence and quality of the sensory input. Gaining a deeper understanding of how our brains process and integrate complex sensory information is essential for optimizing communication and interaction in everyday life, with potential implications for fields such as education, technology, and the treatment of communication disorders.

Introduction

In daily conversations, our brains are bombarded with sensory input from various modalities, making it impossible to comprehensively process everything and everyone in our environment. To effectively communicate in real-life settings, we must not only process auditory information, such as speech, and visual information, like mouth movements and co-speech gestures, but also selectively attend to relevant sources of information while ignoring irrelevant ones. The extent to which the integration of audiovisual speech information is automatic, or influenced by diverted attention conditions, is still a topic of debate (for reviews see: Navarra et al., 2010; Koelewijn et al., 2010; Talsma et al., 2010; Macaluso et al., 2016). While some studies have demonstrated that audiovisual integration is a rather unavoidable process, even when the relevant stimuli are outside the focus of attention (Foxe et al., 2000; Driver 1996; Bertelson et al., 2000; Vroomen et al., 2001a, 2001b), others have shown that audiovisual integration is vulnerable to diverted attention conditions or to visually crowded scenarios (Ahmed et al., 2021; Alsius et al., 2005; 2007; 2014; Alsius and Soto-Faraco, 2011, Andersen Tobias et

al., 2009; Buchan and Munhall, 2011; 2012; Fairhall and Macaluso, 2009; Fujisaki et al., 2006; Senkowski et al., 2005; Tiippana et al., 2011). Thus, how audiovisual integration and attention interact remains poorly understood.

Recent developments put forward a new technique, Rapid Invisible Frequency Tagging (RIFT), as an important tool to investigate exactly this question. RIFT enables researchers to track both attention to multiple stimuli, and investigate the integration of audiovisual signals (Drijvers et al., 2021; Seijdel et al., 2023; Brickwedde et al., 2022; Minarik et al., 2022; Pan et al., 2021; Duecker et al., 2021; Marshall et al., 2021; Zhigalov et al., 2019;2021; Zhigalov & Jensen, 2020; Ferrante et al., 2023). This technique, in which visual stimuli are periodically modulated at high (>50 Hz), stimulus specific 'tagging frequencies', generates steady-state evoked potentials with strong power at the tagged frequencies (Norcia et al., 2015; Vialatte et al., 2010). Frequency tagging has been shown to be a flexible technique to investigate the tracking of attention to multiple different stimuli, with a functional relationship between the amplitude of the SSVEP and the deployment of attention (Toffanin et al. 2009), reflecting the benefit of spatial attention on perceptual processing (Zhigalov et al., 2019). Frequency tagging is interesting in the context of studying audiovisual integration, to investigate whether and how auditory and visual input interact in the brain. Tagging simultaneously presented auditory (using e.g. amplitude modulation) and visual stimuli at different frequencies may lead to non-linear signal interactions indexed by a change in signal power at so-called intermodulation frequencies. For example, using RIFT and magnetoencephalography (MEG), Drijvers et al. (2021) identified an intermodulation frequency at 7 Hz ($f_{visual} - f_{auditory}$) as a result of the interaction between a visual frequency-tagged signal (gesture; 68 Hz) and an auditory frequency-tagged signal (speech; 61 Hz).

In the present study, we investigated how attention affects the processing of auditory and visual information, as well as their integration, during multimodal stimulus presentation. Specifically, we used RIFT and MEG to measure neural activity in response to videos of an actress uttering action verbs (auditory) accompanied with visual gestures on both sides of fixation. We manipulated integration difficulty by varying a lower-order auditory factor (clear/degraded speech) and a higher-order visual factor (congruent/incongruent gesture) and tagged the stimuli at different frequencies for the attended and unattended stimuli. We expected power in visual regions to reflect attention towards the visually tagged input. For the auditory input, we expected power in auditory regions reflecting attention to the auditory tagged input. We expected the interaction between the (attended and ignored) visually tagged signals and the auditory tagged signal to result in spectral peaks at the intermodulation frequencies (65-58 and 63-58; 7 Hz and 5 Hz) respectively. Specifically, we expected this peak to be higher for the attended information (7 Hz) and we expected this activity to occur in the left inferior frontal gyrus (LIFG), a region known to be involved in speech-gesture integration.

Methods

Participants

Forty participants (20 females, 18-40 years old) took part in the experiment. Data from two participants were excluded after data collection, due to missed exclusion criteria (one participant was too old) and problems with comprehension of the task instructions (one participant always answered using the visual information as leading information). For the MEG analyses, participants with inconsistent fixations (gaze outside the fixation for more than 50% of trials during parts of the video)

were excluded. All remaining participants were right-handed and reported corrected-to-normal or normal vision. None of the participants had language, motor or neurological impairment and all reported normal hearing. All participants gave written consent before they participated in the experiment. Participants received monetary compensation or research credits for their participation. The study was approved by the local ethical committee (CMO:2014/288).

Stimuli

The same stimuli as in Drijvers et al. (2021) were used. Participants were presented with 160 video clips showing an actress uttering a highly-frequent action verb accompanied by a matching or a mismatching iconic gesture. Auditory information could be clear or degraded and visual information (gestures) could be congruent or incongruent. In total, there were four conditions, each consisting of 40 trials: clear speech + matching gesture (CM), clear speech, mismatching gesture (CMM), degraded speech + matching gesture (DM) and degrading speech + mismatching gesture (DMM). In all videos, the actress was standing in front of a neutrally colored curtain, in neutrally colored clothes.

During recording of the videos, all gestures were performed by the actress on the fly. The gestures were not predetermined to avoid choreographed or unnatural gestures, as explicit instructions risk drawing undue attention from participants to the gesture's specific form. Verbs for the mismatching gestures were predefined to allow the actress to utter the action verb and depict the mismatching gesture while the face and lips still matched the speech. Videos were on average 2000 ms long. After 120 ms, the preparation (i.e., the first frame in which the hands of the actress moved) of the gesture started. On average, at 550 ms the meaningful part of the gesture (i.e., the stroke) started, followed by speech onset at 680 ms, and average speech offset at 1435 ms. None of these timings differed between conditions. All audio files were intensity-scaled to 70 dB and denoised using Praat (Boersma & Weenink, 2015), before they were recombined with their corresponding video files using Adobe Premiere Pro. To degrade the audio, files were noise-vocoded using Praat. Noise-vocoding preserves the temporal envelope of the audio signal, but degrades the spectral content (Shannon et al., 1995). Based on previous work (Drijvers & Ozyürek, 2017), we used 6-band noise-vocoding, to ensure participants still were able to understand enough of the auditory features of the speech signal to integrate the visual semantic information from the gesture. Our stimulus set comprised frequently-used Dutch action verbs previously employed and validated (Drijvers & Ozyurek, 2017; 2018). All gestures were pretested for iconicity, scoring a mean of 6.1 (SD = 0.64) out of 7, indicating a robust match between gesture and verb. Each video began with the actress in a consistent starting position. Participants were asked to identify the spoken verb and the response choices always included a phonological distractor, semantic distractor, unrelated answer, and the correct answer. While the selected stimuli underwent rigorous validation and vetting to minimize potential stimulus-specific effects, it's noteworthy that they were not counterbalanced among conditions or subjects, which may introduce potential confounds. For further details and descriptions see Drijvers et al., 2017 and Drijvers et al., 2021.

Experimental design and statistical analyses

Participants were tested in a dimly-lit magnetically shielded room and seated 70 cm from the projection screen. All stimuli were presented using MATLAB 2016b (Mathworks Inc, Natrick, USA) and the Psychophysics Toolbox, version 3.0.11 (Brainard, 1997; Kleiner, Brainard, & Pelli, 2007). To achieve

RIFT, we used a GeForce GTX960 2GB graphics card with a refresh rate of 120 Hz, in combination with a PROPixx DLP LED projector (VPixx Technologies Inc., Saint-Bruno-de-Montarville, Canada), which can achieve a presentation rate up to 1,440 Hz. This high presentation rate is achieved by the projector interpreting the four quadrants and three color channels of the GPU screen buffer as individual smaller, grayscale frames, which it then projects in rapid succession, leading to an increase of a factor 12 (4 quadrants * 3 color channels * 120 Hz = 1,440 Hz). The area of the video that would be frequency-tagged was defined by the rectangle in which all gestures occurred. This was achieved by multiplying the luminance of the pixels within that square with a 65/63 Hz sinusoid (modulation depth = 100%; modulation signal equal to 0.5 at sine wave zero-crossing, in order to preserve the mean luminance of the video), phase-locked across trials. For the auditory stimuli, frequency tagging was achieved by multiplying the amplitude of the signal with a 58 Hz sinusoid, with a modulation depth of 100% (following Drijvers et al 2021; Lamminmäki, Parkkonen, & Hari, 2014).

To manipulate spatial attention, we added an attentional cue (arrow pointing to the left or right presented before video onset) and presented the same visual stimulus twice, with different tagging frequencies left and right of fixation. We presented the same video side-by-side on a single trial to avoid unwanted effects from different properties of the videos (e.g. differences in salience, movement kinematics). On half of the trials, participants were asked to attend to the video on the left side of fixation, on the other half of the trials participants were asked to attend to the video on the right side of fixation. The attended video was frequency-tagged at 65 Hz, the unattended video at 63 Hz. We fixed the tagging frequencies at 65 Hz (attended) and 63 Hz (unattended) based on the 1/f power distribution, where lower frequencies typically show higher power (Hermann, 2001). This choice aimed to control for inherent power discrepancies and potential artifacts. The area of the videos that would be frequency-tagged was defined by the rectangle in which all gestures occurred (see Drijvers et al. 2021 for full procedure). Participants were asked to attentively watch and listen to the videos. Auditory information was presented to both ears using MEG-compatible air tubes. Every trial started with a fixation cross (1000 ms), followed by the attentional cue (1000 ms), the videos (2000 ms), a short delay period (1500 ms), and a 4-alternative forced choice identification task (max 3000 ms, followed by the fixation cross of the next trial as soon as a participant pressed one of the 4 buttons). In the 4-alternative forced choice identification task, participants were presented with four written options, and had to identify which verb they heard in the video by pressing one of 4 buttons on an MEG-compatible button box (Figure 1). These answering options always contained a phonological distractor, a semantic distractor, an unrelated answer, and the correct answer. For example, the correct answer could be “strikken” (to tie), the phonological distractor could be “tikken” (to tick), the semantic distractor, which would fit with the gesture, could be “knopen” (to button), and the unrelated answer could be “zouten” (to salt). This task ensured that participants were attentively watching the videos, and enabled us to check whether the verbs were understood. Participants were instructed not to blink during video presentation. The stimuli were presented in four blocks of 40 trials each. In addition to the normal trials, 20 “attention trials” were included to stimulate and monitor attention (see Figure 1). During these trials, participants performed an orthogonal task using already presented stimuli. In these trials, a change in brightness could occur in the attended video, at different latencies and participants were asked to detect this change in brightness. All participants were attentively engaging with the videos throughout the experiment. The whole experiment lasted ~30 minutes and participants were allowed to take a self-paced break after every block. All stimuli were presented in a randomized order per participant.

Data acquisition

Brain activity was measured using MEG, and was recorded throughout the experiment. MEG was acquired using a whole-brain CTF-275 system with axial gradiometers (CTF MEG systems, Coquitlam, Canada). Data were sampled at 1200 Hz after a 300 Hz low-pass filter was applied. Six sensors (MRF66, MLC11, MLC32, and MLO33, MRO33 and MLC61) were permanently disabled due to high noise. Head location was measured using localization coils in both ear canals and on the nasion and was monitored continuously using online head localization software (Stolk et al., 2013). In case of large deviations from the initial head position, we paused the experiment and instructed the subject to move back to the original position. Participants' eye gaze was recorded by an SR Research Eyelink 1,000 eye tracker for artifact rejection purposes. During the task, participants responded using a Fiber Optic Response Pad placed at their right hand.

After the experiment, T1-weighted anatomical magnetic resonance images (MRI) were acquired in the sagittal orientation (or obtained in case of previous participation in MRI/MEG research) using a 3D MPRAGE sequence with the following parameters: TR/TI/TE = 2300/1100/3 ms, FA = 8º, FOV = 256x225x192 mm and a 1 mm isotropic resolution. Parallel imaging (iPAT = 2) was used to accelerate the acquisition resulting in an acquisition time of 5 min and 21 sec. To align structural MRI to MEG, we placed vitamin E capsules in the external meatus of the ear canals, at the same locations as the localizer coils in the MEG system. These anatomical scans were used for source reconstruction of the MEG signals.

Behavioral analysis

Choice accuracy and reaction times (RT) were computed for each condition and each participant. RT analysis was performed on correct responses only. RTs < 100 ms were considered “fast guesses” and removed. Behavioral data were analyzed in Python using the following packages: Statsmodels, Pingouin, SciPy, NumPy, Pandas, (Jones et al., 2001; Vallat, 2018; Oliphant, 2006; Seabold and Perktold, 2010; McKinney, 2011).

MEG preprocessing

MEG data were preprocessed and analyzed using the FieldTrip toolbox (Oostenveld et al., 2011) and custom-built MATLAB scripts (2021b). The MEG signal was epoched based on the onset of the video (t = -1 to 3 s). The data were downsampled to a sampling frequency of 400 Hz after applying a notch filter to remove line noise and harmonics (50, 100, 150, 200, 250, 300 and 350 Hz). Bad channels and trials were rejected via a semi-automatic routine before independent-component analysis (Bell et al., 1995; Jung et al., 2001) was applied. Subsequently, components representing eye-related and heart-related artifacts were projected out of the data (on average, 3.7 components were removed per participant). These procedures resulted in rejection of 9.3% of the trials. The number of rejected trials did not differ significantly between conditions. Participants were instructed to maintain central fixation. Participants with inconsistent fixations (gaze outside the fixation for more than 50% of trials during parts of the video) were excluded, leaving us with 34 participants who consistently fixated throughout the videos.

Frequency Tagging - sensor and source

We first evaluated power at the tagging frequencies in visual and auditory sensory areas by calculating power spectra in the stimulus time window (0.5-1.5 s) and the post-stimulus time window (2.0-3.0 s). With 1-second video segments, we achieved a 1 Hz spectral resolution, aligning with our research objectives. We selected distinct frequencies (65 and 63 Hz; 5 and 7 Hz) for clear differentiation, confirmed by the observed peaks at 63 Hz and 65 Hz. In prior work, we discerned that intermodulation frequency effects were predominantly manifested in power rather than coherence (Drijvers et al., 2021). Because of this, in combination with technical challenges encountered in previous work (i.e., occasional brief delays in video presentation experienced by several participants), we evaluated power changes in visual and auditory sensory areas. We chose a post-stimulus time window as a baseline because, contrary to a prestimulus time window, it is not affected by the button press of the 4-alternative forced choice identification task (following the procedure in Drijvers et al., 2021). To facilitate interpretation of the MEG data, we calculated synthetic planar gradients, as planar gradient maxima are known to be located above neural sources that may underlie them (Bastiaansen & Knösche, 2000). For each individual and each condition, we conducted a spectral analysis for all frequencies between 1 and 130 Hz with a step size of 1 Hz. We applied the fast Fourier transform to the planar-transformed time domain data, after tapering with a boxcar window. Afterward, the horizontal and vertical components of the planar gradient were combined by summing. Using the power spectrum during the baseline condition, the percentage increase in power during stimulus presentation was computed. The resulting power per frequency was averaged over participants and visualized. For the auditory tagging we evaluated all available temporal sensors (MLT11, MLT12, MLT13, MLT14, MLT15, MLT16, MLT21, MLT22, MLT23, MLT24, MLT25, MLT26, MLT27, MLT31, MLT32, MLT33, MLT34, MLT35, MLT36, MLT37, MLT41, MLT42, MLT43, MLT44, MLT45, MLT46, MLT47, MLT51, MLT52, MLT53, MLT54, MLT55, MLT56, MLT57, MRT11, MRT12, MRT13, MRT14, MRT15, MRT16, MRT21, MRT22, MRT23, MRT24, MRT25, MRT26, MRT27, MRT31, MRT32, MRT33, MRT34, MRT35, MRT36, MRT37, MRT41, MRT42, MRT43, MRT44, MRT45, MRT46, MRT47, MRT51, MRT52, MRT53, MRT54, MRT55, MRT56, MRT57), for the visual tagging we evaluated all occipital sensors (MLO11, MLO12, MLO13, MLO14, MLO21, MLO22, MLO23, MLO24, MLO31, MLO32, MLO33, MLO34, MLO41, MLO42, MLO43, MLO44, MLO51, MLO52, MLO53, MRO11, MRO12, MRO13, MRO14, MRO21, MRO22, MRO23, MRO24, MRO31, MRO32, MRO33, MRO34, MRO41, MRO42, MRO43, MRO44, MRO51, MRO52, MRO53, MZO01, MZO02, MZO03). Then, to investigate whether RIFT can be used to identify intermodulation frequencies as a result of the interaction between visual and auditory tagged signals, we repeated the procedure and evaluated power at the intermodulation frequencies (5 Hz and 7 Hz). Here, we focused on left frontal sensors, as the left frontal cortex is known to be involved in the integration of speech and gesture.

Source analysis was performed using dynamic imaging of coherent sources (DICS; Gross et al., 2001). DICS computes source level power at specified frequencies for a set of predefined locations. For each of these locations a beamformer spatial filter is constructed from the sensor-level cross-spectral density matrix (CSD) and the location's lead field matrix. We obtained individual lead fields for every participant using the anatomical information from their MRI. First, we spatially co-registered the individual anatomical MRI to sensor space MEG data by identifying the anatomical markers at the nasion and the two ear canals. We then constructed a realistically shaped single-shell volume conduction model on the basis of the segmented MRI for each participant, and divided the brain volume into a 10 mm spaced grid and warped it to a template brain (MNI). To evaluate power spectra in our sensory regions-of-interest (ROIs) we evaluated visual tagging in all occipital channels, and

auditory tagging in all temporal channels. At source level, we evaluated visual tagging in occipital cortex, including all occipital regions involved with visual processing based on the The Human Brainnetome Atlas (regions 189-196; 199-201; Fan et al. 2016). Auditory tagging was evaluated in temporal regions A41/42 and A22 (regions 71, 72, 75, 76, 79, 80).

Next, we zoomed in on the tagging frequencies and identified the sources of the oscillatory activity. After establishing regions that showed enhanced power at the tagging- and intermodulation frequencies, we proceeded to test the effect of the experimental conditions (clear vs. degraded speech; matching vs. mismatching gesture) within these regions-of-interest (ROIs). The ROIs for the auditory and visual tagged signals were defined by taking the grid points that exceeded 80% of the peak power difference value between stimulus and baseline, across all conditions. For these ROIs, power difference values were extracted per condition. Based on previous studies, the ROI for the intermodulation frequencies at 5 and 7 Hz was anatomically defined by taking those grid points that were part of the Left Inferior Frontal Gyrus (LIFG), using the The Human Brainnetome Atlas; (Fan et al. 2016). To evaluate whether power at the intermodulation frequencies in LIFG was increased during the stimulus window compared to the post-stimulus baseline window, 1 sample permutation tests against zero were performed, using 5000 permutations. For each permutation the signs of a random number of entries in the sample were flipped and the difference in means from the null population mean was recomputed. We repeated this until all permutations were evaluated and stored the differences. The p-value was computed by taking the number of times the stored differences were at least as extreme as the original difference, divided by the total number of permutations. In each iteration, all samples were taken into account (resampling was dependent only on the assignment of values to condition groups)

Results

In the behavioral task we replicated previous results (see Drijvers, Ozyürek & Jensen 2018; Drijvers & Özyürek, 2018; Drijvers, Jensen, Spaak, 2021) and observed that when the speech signal was clear, response accuracy was higher than when speech was degraded (F[1, 37]= 649.82, p<.001, partial η2 = 0.946). Participants performed better when the gesture matched the speech signal compared to when the gesture mismatched the speech signal (F[1, 37]=39.95, p<.001, partial η2 = 0.519). There was a significant interaction between Speech (clear/degraded) and Gesture (matching/mismatching) (F[1, 37]=46.30, p<.001, partial η2 = 0.556). Gestures hindered comprehension when the actress performed a mismatching gesture and speech was degraded (Figure 2).

We observed similar results in the RTs. Participants were faster to identify the verbs when speech was clear, compared to when speech was degraded (F[1, 37] = 568.76, p <.001, partial η^2 = 0.939). Participants were also faster to identify the verbs when the gesture matched the speech signal, compared to when the gesture mismatched the speech signal (F[1, 37] = 31.04, p < .001, partial η^2 = 0.456). There was a significant interaction between Speech (clear/degraded) and Gesture (matching/mismatching) (F[1, 37] = 47.41, p <.001, partial η2 = 0.562). Gestures slowed responses when the actress performed a mismatching gesture and speech was degraded.

In sum, these results demonstrate that the presence of a matching or a mismatching gesture modulates speech comprehension. This effect was larger in degraded speech than in clear speech.

Both visual and auditory frequency tagging produced a clear response that is larger than baseline

As a first step, we calculated the time-locked averages of the event-related fields pooled over conditions. Auditory frequency tagging at 58 Hz produced an auditory steady-state response over left and right-temporal regions (see Figure 3A), and visual frequency tagging at 63 and 65 Hz produced clear visual steady-state responses at occipital regions (see Figure 3B). Both visual and auditory frequency tagging produced a clear steady-state response that was larger than baseline. A one-sample permutation test against zero with 5000 permutations indicated that for the temporal sensors, spectral power was increased at the auditory tagging frequency, 58 Hz (Figure 3A), $p < .001$. For occipital sensors power was increased at the visual tagging frequencies, 63 and 65 Hz (Figure 3B), $p < .001$ and $p < .001$, respectively. We confirmed these results at the source level, by computing the source spectra to evaluate power at the different frequencies in our regions of interest (based on the The Human Brainnetome Atlas; (Fan et al. 2016). Robust tagging responses were found over auditory cortex (58 Hz; Figure 3C) and visual cortex (65 Hz, 63 Hz; Figure 3D), reflecting the neural resources associated with auditory and visual processing. Our initial visualizations encompassing all visual channels and covering the entire visual cortex, give the impression of a stronger response to the attended frequency (65 Hz) as compared to the unattended frequency (63 Hz). However, this wasn't statistically significant and we observed great variations in individual tagging responses. Similarly, there was no significant difference between the attended and unattended tagging responses in auditory cortex.

Auditory and visual sensory regions as the neural sources of the tagging signals

Then, we proceeded to identify the neural sources of the tagged signals using beamformer source analysis. To compare conditions, we formed ROIs by selecting those grid points exceeding a threshold of 80% of peak power change (based on all conditions pooled together). First, we conducted a full-factorial analysis of Speech (clear/degraded), Gesture (matching/mismatching), and Attention (attended/unattended). The results revealed not only a main effect of Gesture but also interaction effects between Speech and Attention ($F(1,37) = 6.89$, $p = 0.0125$), and Gesture and Attention ($F(1,37) = 5.75$, $p = 0.02$). There was no three-way interaction. Therefore, we continued to analyze the power change per condition separately for attended and unattended frequencies. Power change values per condition and per participant were compared in a 2×2 Repeated Measures ANOVA.

Listeners engage their auditory system most when speech is degraded

For the auditory tagging frequency (58 Hz) power was strongest in right-temporal regions, and stronger when speech was degraded compared to when speech was clear ($F[1, 33] = 14.1429$, $p < .001$, partial $\eta 2 = 0.30$). There was no main effect of gesture (matching/mismatching; ($F[1, 33] = 0.88$, $p = 0.36$, partial $\eta 2 = 0.026$) and no interaction effect ($F[1, 33] = 0.16$, $p = .69$, partial $\eta 2 = 0.005$).

Degraded speech enhances covert attention to the gestural information (65 hz)

Similarly, power at the attended visual tagging frequency (65 Hz) was stronger when speech was degraded, compared to when speech was clear ($F[1, 33] = 9.14$, $p = .005$, partial $\eta 2 = 0.217$). Again, there was no main effect of gesture (matching/mismatching; ($F[1, 33] = 0.26$, $p = 0.62$, partial $\eta 2 = 0.008$) and no interaction effect ($F[1, 33] = 0.68$, $p = .42$, partial $\eta 2 = 0.020$).

Mismatching gestures enhance processing of the unattended side (63 Hz)

For the unattended visual tagging frequency (63 Hz), power was stronger when gestures mismatched the speech, compared to when the gestures matched the speech ($F[1, 33] = 15.25$, $p < .001$, partial $\eta 2 = 0.316$).

7 Hz power peak was strongest when speech was degraded and a gesture matched the speech signal

To evaluate whether intermodulation frequencies (5 and 7 Hz in our experiment) could be observed, we then calculated the power spectra at sensor- and source-level in the stimulus time window and the post-stimulus time window. Based on previous work (Drijvers et al., 2021) we focused on left frontal sensors and LIFG. Apart from a peak at 7 Hz for the DM condition, we visually did not observe clear peaks at 5Hz, nor for the other conditions at 7Hz. (Figure 5A/B). Note that the 58 and 65 Hz signal were still present over the frontal regions where we observed the 7 Hz effect. We refined our analyses with direct contrasts between conditions, focusing on power spectra in LIFG and evaluating relative power changes for conditions permitting audiovisual integration (CM vs. CMM; DM vs. DMM). For statistical evaluation see next section. Contrasting DM and DMM, a peak was observed at 7 Hz (Figure 6A).

Frontotemporal and frontal regions as the neural sources of the intermodulation signals

Beamformer source analysis confirmed left frontotemporal regions as the neural sources of the intermodulation signals. Additionally, activity in frontal regions (left/right) and in the right hemisphere was observed. To evaluate whether power at the intermodulation frequencies in LIFG was increased during the stimulus window compared to the post-stimulus baseline window, 1 sample permutation tests against zero were performed. At 7 Hz there was a significant increase in power for both conditions in which gestures matched the speech (CM: $p = 0.043$; DM: $p = 0.004$). A non-parametric Friedman test differentiated % power change across the four conditions (CM, CMM, DM, DMM), Friedman's $Q(3) = 9.071$, $p = .028$. Post-hoc analyses with Wilcoxon signed-rank tests indicated increased power for the degraded match (DM) condition, compared to the degraded mismatch (DMM) condition, $W = 113$, $p = .01$, after Benjamini/Hochberg correction for multiple comparisons. This suggests that activity in LIFG is increased for those conditions that benefit most from integration (with a matching gesture). To exclude the possibility that unreliable participants' (outliers) confound our findings, we detected participants with any observation that was classified as a suspected outlier using the interquartile range (IQR) criterion (2.5*IQR). This resulted in one outlier. We repeated the analyses without this participant and again found the same patterns of results. There were no differences between conditions at 7 Hz in the Left Postcentral Gyrus (A1/2/3), which was taken as a control region as it is not typically associated with audiovisual integration, attention, or 5-7 Hz activity related to cognitive tasks, Friedman' s $Q(3) = 2.576$, $p = 0.462$.

Beamformer source analysis, contrasting CM and CMM, revealed enhanced activity in the temporal lobe, particularly the STS, a region associated with multisensory processing. When comparing DM and DMM, we observed increased activity in LIFG, left parietal regions (SPL), temporal areas (STG), and the occipital cortex (Figure 6B).

Discussion

In the present MEG study we used RIFT to investigate how covert attention affects the processing of auditory (speech) and visual information (iconic gestures), as well as their integration, during multimodal communication. Our results showed that attention selectively modulates the processing of sensory information, depending on the congruence (matching vs. mismatching gestures) and quality (clear vs. degraded speech) for the task at hand. Specifically, we observed enhanced processing of auditory information when speech was degraded. In line with previous studies (Drijvers et al., 2021) we observed a stronger drive by the 58 Hz amplitude modulation signal in auditory regions when speech was degraded compared to when speech was clear. In visual regions, we observed a stronger drive by the attended visual modulation signal (65 Hz) when speech was degraded. For the unattended visual modulation signal (63 Hz), we observed enhanced processing when gestures were mismatching. We observed enhanced activity in LIFG at the attended intermodulation frequency (7 Hz, $f_{visual\_attended} - f_{auditory}$) for those conditions that benefitted from integration (i.e. conditions with a matching gesture: CM. DM). Together, our results suggest that attention can modulate audiovisual processing and interaction, depending on the relevance and quality of the sensory input.

Degraded speech enhances attention to auditory information

The current study provides evidence that degraded speech enhances attention to auditory information when compared to clear speech. We observed a stronger drive by the 58 Hz amplitude modulation signal in auditory cortex when speech was degraded, compared to when speech was clear. This finding is consistent with previous studies that have reported enhanced attention to degraded speech (Helfer and Freyman, 2008; Drijvers et al., 2021). The increase in attention to auditory information in the degraded speech condition may be due to increased effort needed to understand the speech, leading to a greater allocation of attentional resources to the auditory signal (Wild et al., 2012).

Degraded speech enhances processing of the attended gestural information

Additionally, degraded speech enhanced processing of the attended visual information. In occipital regions, we observed a stronger drive by the 65 Hz visual modulation signal when speech was degraded compared to when speech was clear. The enhanced attention to the attended gestural information in the degraded speech condition may be due to a compensatory mechanism, where participants rely more heavily on visual information in the presence of degraded auditory information (Drijvers & Ozyurek, 2017; Holle et al., 2010; Holle & Gunter, 2007; Ross et al., 2007; Obermeier, Dolk & Gunter, 2012; Erber, 1975; Sumby and Pollack, 1954).

In previous work, the opposite pattern was found; that is, a stronger drive when speech was clear, rather than degraded (Drijvers et al., 2021). However, in that study participants were presented with only one video in the center of the screen. This allowed for more room for participants to explore the different planes and parts of the visual information (away from the gestures). Because listeners gaze

more often to the face and mouth than to gestures when speech is degraded (Drijvers, Vaitonytė, & Özyürek, 2019), this could have resulted in lower power at the visual tagging frequency when speech was degraded.

Mismatching gestures enhance processing of the unattended gestural information

The processing of gestures during audiovisual integration has been shown to be influenced by the congruency between speech and gestural information. Our findings support this idea and suggest that the presence of mismatching gestures can reduce visual attention to the attended gestural information and enhance processing of the unattended side. This finding is consistent with previous studies showing that processing of a task-relevant stimulus can be reduced in the presence of task-irrelevant information (Lavie et al., 2004). In the current study, it is possible that the inability to integrate the mismatching gestural information led participants to allocate less attentional resources to the attended side and instead attend to the unattended side. In other words, subjects may have shown less focused attention during mismatching gestures, leading to less suppression of visual information on the unattended side of the screen. In our study, we presented the same video on both sides to ensure a controlled comparison, minimizing saliency-related effects. While this might have led to cross-video auditory and visual integration, it reduced potential confounds from variations between videos. We acknowledge the limitations in our design but believe our choices effectively addressed the research question. Future studies can further refine these task designs based on our findings

Flexible allocation of neurocognitive resources

Overall, these findings suggest that the recruitment of sensory resources is not static, but dynamic. The ability to flexibly allocate neurocognitive resources allows listeners to rapidly adapt to speech processing under a wide variety of conditions (Peelle, 2018). For example, degraded speech enhances attentional allocation to both auditory and gestural information, potentially reflecting a compensatory mechanism to overcome the challenges of processing degraded speech. On the other hand, attention may be diverted when the audiovisual information does not match and therefore becomes irrelevant. These findings also highlight the importance of considering both lower-order and higher-order factors when investigating audiovisual integration and attention. The manipulation of degradation in the auditory modality allowed us to investigate the role of lower-order factors (i.e., the quality of the sensory input), while the manipulation of gesture congruence allowed us to investigate the role of higher-order factors (i.e., the semantic relationship between the auditory and visual information). Future studies could build on this by manipulating a wider range of factors such as the complexity, familiarity or timing, to better understand how different types of information interact during audiovisual processing.

The auditory tagged speech signal and attended gestural information interact in left-frontotemporal regions

Our findings also shed light on the role of *top-down* attention in audiovisual integration. At the attended intermodulation frequency (7 Hz), we found that power in LIFG was enhanced for degraded speech with a matching gesture compared to degraded speech with a mismatching gesture. This is in line with earlier work showing an influence of the quality or relevance of sensory input in modulating audiovisual integration. For example, studies have shown that manipulations of sensory congruence can affect the degree of audiovisual integration, with greater integration occurring when stimuli are congruent across modalities (e.g., Vatakis & Spence, 2007; Welch & Warren, 1980; Talsma et al., 2010).

Moreover, our results showed stronger power at 7 Hz when speech was degraded and the gesture was matching compared to when the gesture was mismatching. This suggests that when the auditory signal was weaker due to the degradation of speech, attention was shifted more strongly towards the visual modality when this was relevant, resulting in enhanced neural processing of the visual stimulus at the attended frequency. In simple audiovisual perceptual tasks, inverse effectiveness is often observed, which holds that the weaker the unimodal stimuli, or the poorer their signal-to-noise ratio, the stronger the audiovisual benefit (Kayser et al., 2005; Meredith and Stein, 1983, 1986b; Perrault et al., 2005; Stanford et al., 2005). A similar pattern has been observed for more complex audiovisual speech stimuli, where results show an enhanced benefit of adding information from visible speech to the speech signal at moderate levels of noise-vocoding (Drijvers & Ozyürek, 2017) or an enhanced benefit from bimodal presentation for words that were less easily recognized through the visual input (van de Rijt et al., 2019). In our study, in line with this idea, we observe enhanced power at the attended intermodulation frequency (7 Hz) for the degraded match condition compared to the degraded mismatch condition.

The observed effects on the intermodulation frequencies are different from earlier work (Drijvers et al., 2021) that observed a reliable peak at 7 Hz power during stimulation when integration of the lower-order auditory and visual input was optimal, that is, when speech was clear and a gesture was matching. These previous results suggested that the strength of the intermodulation frequency reflected the ease of lower-order audiovisual integration. However, results from the current study indicating an effect of gesture congruence (enhanced activity for the DM condition compared to DMM), suggest otherwise. We speculate that this discrepancy might be due to differences in task demand. The current study utilized smaller videos displayed outside participants' fixation, in contrast to Drijvers et al. (2021) where a single central video was presented. In that study, participants could freely explore visuals due to a centrally positioned video. Conversely, our design constrained visual exploration. In the current study, we did not counterbalance the frequencies for attended and unattended conditions, which is an acknowledged limitation. This design aspect potentially introduces ambiguity about whether observed effects at 7 Hz are indeed representative of intermodulation processes or rather specific endogenous activities associated with this frequency. There is a distinct possibility that effects we attribute to intermodulation could be conflated with inherent oscillatory behavior at 7 Hz. In future work, the attended and unattended tagging frequencies should be counterbalanced. This would also allow for a direct comparison between power at the attended vs. the unattended frequency.

Because we selected specific tagging frequencies that resulted in intermodulation frequencies at 5 and 7 Hz, our effects of integration are manifested in the theta range. Theta oscillations have been implicated in both attentional selection and audiovisual integration processes. For example, theta activity seems to be related to cognitive control in cross-modal visual attention paradigms (Wang et al., 2016), multisensory divided attention (Keller et al., 2017) and theta oscillations have been shown to modulate attentional search performance (Dugue & VanRullen 2015). Thus, parts of the 7 Hz power may reflect a combination of attentional and integrative processes. For example, enhanced theta power in response to clear speech may reflect the presence of more attentional resources (driven by the simplicity of the trial). On the other hand, enhanced theta power in response to degraded speech for the attended stimulus may reflect both increased attentional demands due to the degraded speech and increased integration demands due to the need to compensate for the degraded auditory information. However, mostly mid-frontal and central brain areas, and not LIFG, have been shown to be involved in allocating and controlling the direction of attention (Yantis & Serences, 2003; Woldorff et al., 2004; Corbetta & Shulman 2002, Moore, 2003). Future studies could use different tagging

frequencies (and thus different intermodulation frequencies) to try to disentangle these effects of both integration and attention. Moreover, time-resolved measures could be a valuable avenue for future investigations to elucidate when these effects occur in time.

Conclusion

This study provides insights into the neural mechanisms underlying attentional modulation of audiovisual processing and integration during communication. By utilizing RIFT and MEG, we were able to identify the neural sources associated with sensory processing and integration, and their involvement during different requirements for audiovisual integration. Our findings highlight the critical role of degraded speech in enhancing attention to both auditory and attended gestural information, and the potential role of mismatching gestural information in shifting visual attention away from the attended side. Overall, our results demonstrate the complex interplay between different sensory modalities and attention during audiovisual integration and the importance of considering both lower- and higher-order factors in understanding these processes. The role of attention may be context-dependent. Understanding the factors that modulate audiovisual speech-gesture integration is crucial for developing a more comprehensive understanding of how humans communicate in daily life.

References

Ahmed F, Nidiffer AR, O'Sullivan AE, Zuk NJ, Lalor EC (2021) The integration of continuous audio and visual speech in a cocktail-party environment depends on attention. Cold Spring Harbor Laboratory:2021.02.10.430634 Available at: https://www.biorxiv.org/content/10.1101/2021.02.10.430634v1

Alsius A, Möttönen R, Sams ME, Soto-Faraco S, Tiippana K (2014) Effect of attentional load on audiovisual speech perception: evidence from ERPs. Front Psychol 5:727 Available at: http://dx.doi.org/10.3389/fpsyg.2014.00727.

Alsius A, Navarra J, Campbell R, Soto-Faraco S (2005) Audiovisual integration of speech falters under high attention demands. Curr Biol 15:839–843 Available at: http://dx.doi.org/10.1016/j.cub.2005.03.046.

Alsius A, Navarra J, Soto-Faraco S (2007) Attention to touch weakens audiovisual speech integration. Exp Brain Res 183:399–404 Available at: http://dx.doi.org/10.1007/s00221-007-1110-1.

Alsius A, Soto-Faraco S (2011) Searching for audiovisual correspondence in multiple speaker scenarios. Exp Brain Res 213:175–183 Available at: http://dx.doi.org/10.1007/s00221-011-2624-0.

Andersen TS, Tiippana K, Laarni J, Kojo I, Sams M (2009) The role of visual spatial attention in audiovisual speech perception. Speech Commun 51:184–193 Available at: https://www.sciencedirect.com/science/article/pii/S016763930800126X.

Bastiaansen MC, Knösche TR (2000) Tangential derivative mapping of axial MEG applied to event-related desynchronization research. Clin Neurophysiol 111:1300–1305 Available at: http://dx.doi.org/10.1016/s1388-2457(00)00272-8.

Bell A, Jung TP, Sejnowski TJ (1995) Independent component analysis of electroencephalographic data. Adv Neural Inf Process Syst Available at:

https://proceedings.neurips.cc/paper/1995/hash/754dda4b1ba34c6fa89716b85d68532b-Abstract.html.

Bertelson P, Vroomen J, de Gelder B, Driver J (2000) The ventriloquist effect does not depend on the direction of deliberate visual attention. Percept Psychophys 62:321–332 Available at: http://dx.doi.org/10.3758/bf03205552.

Boersma P, Weenink D (n.d.) Praat [computer program]. Version 6.0. 05. URL http://www praat org.

Brainard DH (1997) The Psychophysics Toolbox. Spat Vis 10:433–436 Available at: https://www.ncbi.nlm.nih.gov/pubmed/9176952.

Brickwedde M, Limachya R, Markiewicz R, Sutton E, Shapiro K, Jensen O, Mazaheri A (2022) Cross-modal alterations of Alpha Activity do not reflect inhibition of early sensory processing: A frequency tagging study. bioRxiv:2022.04.19.488727 Available at: https://www.biorxiv.org/content/10.1101/2022.04.19.488727v1 [Accessed April 20, 2022].

Buchan JN, Munhall KG (2011) The influence of selective attention to auditory and visual speech on the integration of audiovisual speech information. Perception 40:1164–1182 Available at: http://dx.doi.org/10.1068/p6939.

Buchan JN, Munhall KG (2012) The effect of a concurrent working memory task and temporal offsets on the integration of auditory and visual speech information. Seeing Perceiving 25:87–106 Available at: http://dx.doi.org/10.1163/187847611X620937.

Corbetta M, Shulman GL (2002) Control of goal-directed and stimulus-driven attention in the brain. Nat Rev Neurosci 3:201–215 Available at: http://dx.doi.org/10.1038/nrn755.

Dick AS, Mok EH, Raja Beharelle A, Goldin-Meadow S, Small SL (2014) Frontal and temporal contributions to understanding the iconic co-speech gestures that accompany speech. Hum Brain Mapp 35:900–917 Available at: https://onlinelibrary.wiley.com/doi/abs/10.1002/hbm.22222?casa_token=CjvdzLNCqisAAAAA:yBt_dz94D4FyLa2vRIktE8SJ-crh4Xb1fpfKWvNNu2HMq9mAfoXoJNA7YBAmFn_XG2CY_0BFjKkY.

Drijvers L, Jensen O, Spaak E (2021) Rapid invisible frequency tagging reveals nonlinear integration of auditory and visual information. Hum Brain Mapp Available at: http://dx.doi.org/10.1002/hbm.25282.

Drijvers L, Özyürek A (2017) Visual Context Enhanced: The Joint Contribution of Iconic Gestures and Visible Speech to Degraded Speech Comprehension. J Speech Lang Hear Res 60:212–222 Available at: http://dx.doi.org/10.1044/2016_JSLHR-H-16-0101.

Drijvers, L., Özyürek, A., & Jensen, O. (2018). Hearing and seeing meaning in noise: Alpha, beta, and gamma oscillations predict gestural enhancement of degraded speech comprehension. *Human brain mapping, 39*(5), 2075-2087.

Drijvers L, Vaitonytė J, Özyürek A (2019) Degree of Language Experience Modulates Visual Attention to Visible Speech and Iconic Gestures During Clear and Degraded Speech Comprehension. Cogn Sci 43:e12789 Available at: http://dx.doi.org/10.1111/cogs.12789.

Driver J (1996) Enhancement of selective listening by illusory mislocation of speech sounds due to lip-reading. Nature 381:66–68 Available at: http://dx.doi.org/10.1038/381066a0.

637 Duecker K, Gutteling TP, Herrmann CS, Jensen O (2021) No Evidence for Entrainment:
638 Endogenous Gamma Oscillations and Rhythmic Flicker Responses Coexist in Visual Cortex. J
639 Neurosci 41:6684–6698 Available at: http://dx.doi.org/10.1523/JNEUROSCI.3134-20.2021.

640 Erber NP (1975) Auditory-visual perception of speech. J Speech Hear Disord 40:481–492
641 Available at: http://dx.doi.org/10.1044/jshd.4004.481.

642 Fairhall SL, Macaluso E (2009) Spatial attention can modulate audiovisual integration at multiple
643 cortical and subcortical sites. Eur J Neurosci 29:1247–1257 Available at:
644 http://dx.doi.org/10.1111/j.1460-9568.2009.06688.x.

645 Fan L, Li H, Zhuo J, Zhang Y, Wang J, Chen L, Yang Z, Chu C, Xie S, Laird AR, Fox PT, Eickhoff SB,
646 Yu C, Jiang T (2016) The Human Brainnetome Atlas: A New Brain Atlas Based on Connectional
647 Architecture. Cereb Cortex 26:3508–3526 Available at:
648 http://dx.doi.org/10.1093/cercor/bhw157.

649 Ferrante O, Zhigalov A, Hickey C, Jensen O (2023) Statistical Learning of Distractor Suppression
650 Down-regulates Pre-Stimulus Neural Excitability in Early Visual Cortex. J Neurosci Available at:
651 http://dx.doi.org/10.1523/JNEUROSCI.1703-22.2022.

652 Foxe JJ, Morocz IA, Murray MM, Higgins BA, Javitt DC, Schroeder CE (2000) Multisensory
653 auditory–somatosensory interactions in early cortical processing revealed by high-density
654 electrical mapping. Cognitive Brain Research 10:77–83 Available at:
655 https://www.sciencedirect.com/science/article/pii/S0926641000000240.

656 Fujisaki W, Koene A, Arnold D, Johnston A, Nishida S 'ya (2006) Visual search for a target
657 changing in synchrony with an auditory signal. Proc Biol Sci 273:865–874 Available at:
658 http://dx.doi.org/10.1098/rspb.2005.3327.

659 Hartcher-O'Brien J, Soto-Faraco S, Adam R (2017) A Matter of Bottom-Up or Top-Down
660 Processes: The Role of Attention in Multisensory Integration. Frontiers Media SA. Available at:
661 https://play.google.com/store/books/details?id=UUswDwAAQBAJ.

662 Holle H, Gunter TC (2007) The role of iconic gestures in speech disambiguation: ERP evidence. J
663 Cogn Neurosci 19:1175–1192 Available at: http://dx.doi.org/10.1162/jocn.2007.19.7.1175.

664 Holle H, Obleser J, Rueschemeyer S-A, Gunter TC (2010) Integration of iconic gestures and
665 speech in left superior temporal areas boosts speech comprehension under adverse listening
666 conditions. Neuroimage 49:875–884 Available at:
667 http://dx.doi.org/10.1016/j.neuroimage.2009.08.058.

668 Jones E, Oliphant T, Peterson P, Others (2001) SciPy: Open source scientific tools for Python.

669 Jung T-P, Makeig S, McKeown MJ, Bell AJ, Lee T-W, Sejnowski TJ (2001) Imaging Brain Dynamics
670 Using Independent Component Analysis. Proc IEEE Inst Electr Electron Eng 89:1107–1122
671 Available at: http://dx.doi.org/10.1109/5.939827.

672 Kayser C, Petkov CI, Augath M, Logothetis NK (2005) Integration of touch and sound in auditory
673 cortex. Neuron 48:373–384 Available at: http://dx.doi.org/10.1016/j.neuron.2005.09.018.

674 Kleiner M, Brainard D, Pelli D (n.d.) What's new in Psychtoolbox-3? Available at:
675 https://pure.mpg.de/rest/items/item_1790332/component/file_3136265/content [Accessed
676 March 20, 2023].

677 Koelewijn T, Bronkhorst A, Theeuwes J (2010) Attention and the multiple stages of multisensory
678 integration: A review of audiovisual studies. Acta Psychol 134:372–384 Available at:
679 http://dx.doi.org/10.1016/j.actpsy.2010.03.010.

680 Lamminmäki S, Parkkonen L, Hari R (2014) Human neuromagnetic steady-state responses to
681 amplitude-modulated tones, speech, and music. Ear Hear 35:461–467 Available at:
682 http://dx.doi.org/10.1097/AUD.0000000000000033.

683 Lavie N, Hirst A, de Fockert JW, Viding E (2004) Load theory of selective attention and cognitive
684 control. J Exp Psychol Gen 133:339–354 Available at: http://dx.doi.org/10.1037/0096-
685 3445.133.3.339.

686 Macaluso E, Noppeney U, Talsma D, Vercillo T, Hartcher-O'Brien J, Adam R (2016) The Curious
687 Incident of Attention in Multisensory Integration: Bottom-up vs. Top-down. Multisensory
688 Research 29:557–583 Available at: https://brill.com/view/journals/msr/29/6-7/article-
689 p557_3.xml?casa_token=wrkurwb7qToAAAAA:jg5D9c5wVFPy1p6cVbfPmMAiXEBuFC0HZiXjGm
690 1zxKQxiEHTOfjJMg5jOHbUgrpwW2Ugs8Y [Accessed March 20, 2023].

691 Marshall TR, Ruesseler M, Hunt LT, O'Reilly JX (2021) Computational specialization within the
692 cortical eye movement system. bioRxiv:2021.05.03.442155 Available at:
693 https://www.biorxiv.org/content/10.1101/2021.05.03.442155v1.abstract [Accessed August 19,
694 2021].

695 Mc Kinney W (n.d.) Pandas: A foundational python library for data analysis and statistics.
696 Available at:
697 https://www.dlr.de/sc/portaldata/15/resources/dokumente/pyhpc2011/submissions/pyhpc20
698 11_submission_9.pdf [Accessed March 20, 2023].

699 Meredith MA, Stein BE (1983) Interactions among converging sensory inputs in the superior
700 colliculus. Science 221:389–391 Available at: http://dx.doi.org/10.1126/science.6867718.

701 Minarik T, Berger B, Jensen O (2022) Optimal parameters for Rapid Invisible Frequency Tagging
702 using MEG. bioRxiv:2022.12.21.521401 Available at:
703 https://www.biorxiv.org/content/10.1101/2022.12.21.521401v1 [Accessed January 4, 2023].

704 Moore T, Armstrong KM, Fallah M (2003) Visuomotor origins of covert spatial attention. Neuron
705 40:671–683 Available at: http://dx.doi.org/10.1016/s0896-6273(03)00716-5.

706 Navarra J, Alsius A, Soto-Faraco S, Spence C (2010) Assessing the role of attention in the
707 audiovisual integration of speech. Inf Fusion 11:4–11 Available at:
708 https://www.sciencedirect.com/science/article/pii/S1566253509000347.

709 Norcia AM, Appelbaum LG, Ales JM, Cottereau BR, Rossion B (2015) The steady-state visual
710 evoked potential in vision research: A review. J Vis 15:4 Available at:
711 http://dx.doi.org/10.1167/15.6.4.

712 Obermeier C, Dolk T, Gunter TC (2012) The benefit of gestures during communication: evidence
713 from hearing and hearing-impaired individuals. Cortex 48:857–870 Available at:
714 http://dx.doi.org/10.1016/j.cortex.2011.02.007.

715 Oliphant TE (2006) A guide to NumPy. Trelgol Publishing USA. Available at:
716 https://ecs.wgtn.ac.nz/foswiki/pub/Support/ManualPagesAndDocumentation/numpybook.pdf.

717 Oostenveld R, Fries P, Maris E, Schoffelen J-M (2011) FieldTrip: Open source software for

718 advanced analysis of MEG, EEG, and invasive electrophysiological data. Comput Intell Neurosci
719 2011:156869 Available at: http://dx.doi.org/10.1155/2011/156869.

720 Pan Y, Frisson S, Jensen O (2021) Neural evidence for lexical parafoveal processing. Nat
721 Commun 12:5234 Available at: http://dx.doi.org/10.1038/s41467-021-25571-x.

722 Peelle JE (2018) Speech Comprehension: Stimulating Discussions at a Cocktail Party. Current
723 Biology 28:R68–R70 Available at: http://dx.doi.org/10.1016/j.cub.2017.12.005.

724 Perrault TJ Jr, Vaughan JW, Stein BE, Wallace MT (2005) Superior colliculus neurons use distinct
725 operational modes in the integration of multisensory stimuli. J Neurophysiol 93:2575–2586
726 Available at: http://dx.doi.org/10.1152/jn.00926.2004.

727 Ross LA, Saint-Amour D, Leavitt VM, Javitt DC, Foxe JJ (2007) Do you see what I am saying?
728 Exploring visual enhancement of speech comprehension in noisy environments. Cereb Cortex
729 17:1147–1153 Available at: http://dx.doi.org/10.1093/cercor/bhl024.

730 Sawaki R, Luck SJ, Raymond JE (2015) How Attention Changes in Response to Incentives. J Cogn
731 Neurosci 27:2229–2239 Available at: http://dx.doi.org/10.1162/jocn_a_00847.

732 Seabold S, Perktold J (2010) Statsmodels: Econometric and statistical modeling with python. In:
733 Proceedings of the 9th Python in Science Conference, pp 61. Scipy. Available at:
734 https://www.researchgate.net/profile/Josef_Perktold/publication/264891066_Statsmodels_Ec
735 onometric_and_Statistical_Modeling_with_Python/links/5667ca9308ae34c89a0261a8/Statsmo
736 dels-Econometric-and-Statistical-Modeling-with-Python.pdf.

737 Seijdel N, Marshall TR, Drijvers L (2022) Rapid invisible frequency tagging (RIFT): a promising
738 technique to study neural and cognitive processing using naturalistic paradigms. Cereb Cortex
739 Available at: http://dx.doi.org/10.1093/cercor/bhac160.

740 Senkowski D, Talsma D, Herrmann CS, Woldorff MG (2005) Multisensory processing and
741 oscillatory gamma responses: effects of spatial selective attention. Exp Brain Res 166:411–426
742 Available at: http://dx.doi.org/10.1007/s00221-005-2381-z.

743 Shannon RV, Zeng FG, Kamath V, Wygonski J, Ekelid M (1995) Speech recognition with primarily
744 temporal cues. Science 270:303–304 Available at:
745 http://dx.doi.org/10.1126/science.270.5234.303.

746 Stanford TR, Quessy S, Stein BE (2005) Evaluating the operations underlying multisensory
747 integration in the cat superior colliculus. J Neurosci 25:6499–6508 Available at:
748 http://dx.doi.org/10.1523/JNEUROSCI.5095-04.2005.

749 Sumby WH, Pollack I (1954) Visual Contribution to Speech Intelligibility in Noise. J Acoust Soc
750 Am 26:212–215 Available at: https://doi.org/10.1121/1.1907309.

751 Talsma D, Senkowski D, Soto-Faraco S, Woldorff MG (2010) The multifaceted interplay between
752 attention and multisensory integration. Trends Cogn Sci 14:400–410 Available at:
753 http://dx.doi.org/10.1016/j.tics.2010.06.008.

754 Tiippana K, Puharinen H, Möttönen R, Sams M (2011) Sound location can influence audiovisual
755 speech perception when spatial attention is manipulated. Seeing Perceiving 24:67–90 Available
756 at: http://dx.doi.org/10.1163/187847511X557308.

757 Toffanin P, de Jong R, Johnson A, Martens S (2009) Using frequency tagging to quantify

attentional deployment in a visual divided attention task. Int J Psychophysiol 72:289–298 Available at: http://dx.doi.org/10.1016/j.ijpsycho.2009.01.006.

van de Rijt LPH, Roye A, Mylanus EAM, van Opstal AJ, van Wanrooij MM (2019) The Principle of Inverse Effectiveness in Audiovisual Speech Perception. Front Hum Neurosci 13:335 Available at: http://dx.doi.org/10.3389/fnhum.2019.00335.

Vallat, R. (2018). Pingouin: statistics in Python. *Journal of Open Source Software, 3*(31), 1026, https://doi.org/10.21105/joss.01026

Vatakis A, Spence C (2007) Crossmodal binding: evaluating the "unity assumption" using audiovisual speech stimuli. Percept Psychophys 69:744–756 Available at: http://dx.doi.org/10.3758/bf03193776.

Vialatte F-B, Maurice M, Dauwels J, Cichocki A (2010) Steady-state visually evoked potentials: focus on essential paradigms and future perspectives. Prog Neurobiol 90:418–438 Available at: http://dx.doi.org/10.1016/j.pneurobio.2009.11.005.

Vroomen J, Bertelson P, de Gelder B (2001a) The ventriloquist effect does not depend on the direction of automatic visual attention. Percept Psychophys 63:651–659 Available at: http://dx.doi.org/10.3758/bf03194427.

Vroomen J, Driver J, de Gelder B (2001b) Is cross-modal integration of emotional expressions independent of attentional resources? Cogn Affect Behav Neurosci 1:382–387 Available at: http://dx.doi.org/10.3758/cabn.1.4.382.

Welch RB, Warren DH (1980) Immediate perceptual response to intersensory discrepancy. Psychol Bull 88:638–667 Available at: https://www.ncbi.nlm.nih.gov/pubmed/7003641.

Wild CJ, Yusuf A, Wilson DE, Peelle JE, Davis MH, Johnsrude IS (2012) Effortful listening: the processing of degraded speech depends critically on attention. J Neurosci 32:14010–14021 Available at: http://dx.doi.org/10.1523/JNEUROSCI.1528-12.2012.

Willems RM, Ozyürek A, Hagoort P (2007) When language meets action: the neural integration of gesture and speech. Cereb Cortex 17:2322–2333 Available at: http://dx.doi.org/10.1093/cercor/bhl141.

Willems RM, Ozyürek A, Hagoort P (2009) Differential roles for left inferior frontal and superior temporal cortex in multimodal integration of action and language. Neuroimage 47:1992–2004 Available at: http://dx.doi.org/10.1016/j.neuroimage.2009.05.066.

Woldorff MG, Hazlett CJ, Fichtenholtz HM, Weissman DH, Dale AM, Song AW (2004) Functional parcellation of attentional control regions of the brain. J Cogn Neurosci 16:149–165 Available at: http://dx.doi.org/10.1162/089892904322755638.

Yantis S, Serences JT (2003) Cortical mechanisms of space-based and object-based attentional control. Curr Opin Neurobiol 13:187–193 Available at: http://dx.doi.org/10.1016/s0959-4388(03)00033-3.

Zhigalov A, Duecker K, Jensen O (2021) The visual cortex produces gamma band echo in response to broadband visual flicker. PLoS Comput Biol 17:e1009046 Available at: http://dx.doi.org/10.1371/journal.pcbi.1009046.

Zhigalov A, Herring JD, Herpers J, Bergmann TO, Jensen O (2019) Probing cortical excitability

using rapid frequency tagging. Neuroimage 195:59–66 Available at: http://dx.doi.org/10.1016/j.neuroimage.2019.03.056.

Zhigalov A, Jensen O (2020) Alpha oscillations do not implement gain control in early visual cortex but rather gating in parieto-occipital regions. Hum Brain Mapp 41:5176–5186 Available at: http://dx.doi.org/10.1002/hbm.25183.

Figure Legends

Figure 1. Experimental paradigm. Participants were asked to attend to one of the videos, indicated by a cue. The attended video was frequency-tagged at 65 Hz, the unattended video at 63 Hz. Speech was frequency-tagged at 58 Hz. Participants were asked to attentively watch and listen to the videos. After the video, participants were presented with four written options, and had to identify which verb they heard in the video by pressing one of 4 buttons on an MEG-compatible button box. This task ensured that participants were attentively watching the videos, and to check whether the verbs were understood. Participants were instructed not to blink during the video presentation. In addition to the normal trials, "attention trials" were included in which participants were asked to detect a change in brightness.

Figure 2. Verb categorization behavior A) Accuracy results per condition. Response accuracy is highest for clear speech conditions, and when a gesture matches the speech signal. **B)** Reaction times per condition. Reaction times are faster in clear speech and when a gesture matches the speech signal.

Figure 3. Power at temporal and occipital sensors and corresponding source regions (% increased compared to a post-stimulus baseline) averaged across conditions. A) Average ERF for a single subject at selected sensors overlying the left and right temporal lobe. Auditory input was tagged by 58 Hz amplitude modulation. Tagging was phase-locked over trials. ERFs show combined planar gradient data. **B)** Average ERF for a single subject at selected sensors overlying the occipital lobe. Visual input was tagged by 65 Hz and a 63 Hz flicker. **C)** power increase in temporal sensors at the tagged frequency of the auditory stimulus (58 Hz) **D)** power increases in occipital sensors are observed at the visual tagging frequencies (63 Hz: unattended; 65 Hz: attended). **E)** power increase in auditory cortex at the tagged frequency of the auditory stimulus (58 Hz). **F)** power increases in visual cortex observed at the visual tagging frequencies (63 Hz: unattended; 65 Hz: attended). Shaded error bars represent the Standard Error.

Figure 4. Sources of power at the auditory tagged signal at 58 Hz and the visually tagged signals at 65 Hz and 63 Hz. A). Power change in percentage when comparing power values in the stimulus window to a post stimulus baseline for the different tagging frequencies, pooled over conditions. Power change is the largest over temporal regions for the auditory tagging frequency, and largest over occipital regions for the visually tagged signals. B) Power change values in percentage extracted from the ROIs. Raincloud plots reveal raw data, density, and boxplots for power change in the different conditions. CM = clear speech with a matching gesture, CMM = clear speech with a mismatching gesture, DM = degraded speech with a matching gesture and DMM = degraded speech with a mismatching gesture.

Figure 5. Power at the intermodulation frequencies (f_{visual}-$f_{auditory}$). A) Power over left frontal sensors (% increased compared to a post-stimulus baseline). B) Power over LIFG source region (% increased compared to a

post-stimulus baseline) C) Sources of power at 7 Hz **D)** Power change values in percentage extracted from the Left Inferior Frontal Gyrus (LIFG) in source space. Raincloud plots reveal raw data, density, and boxplots for power change per condition.

Figure 6. A) Power over LIFG source region (% increased compared to the mismatching gesture conditions). Shaded error bars represent the Standard Error. B) Power was higher in the CM condition compared to the CMM condition across the temporal lobe. Comparing DM and DMM, we observed enhanced activity in LIFG, left parietal regions and occipital cortex.