

1 **Rapid invisible frequency tagging reveals nonlinear integration of auditory and visual**
2 **information**

3
4 **Linda Drijvers^{1,2}, Ole Jensen^{3*} & Eelke Spaak^{4*}**

5
6 ¹ Radboud University, Donders Institute for Brain, Cognition, and Behaviour, Centre for
7 Cognition, Montessorilaan 3, 6525 HR, Nijmegen, The Netherlands

8 ² Max Planck Institute for Psycholinguistics, Wundtlaan 1, 6525 XD, Nijmegen, The Netherlands

9 ³ School of Psychology, Centre for Human Brain Health, University of Birmingham, Birmingham
10 B15 2TT, United Kingdom

11 ⁴ Radboud University, Donders Institute for Brain, Cognition, and Behaviour, Centre for Cognitive
12 Neuroimaging, Kapittelweg 29, 6525 EN, Nijmegen, The Netherlands

13
14 *** Shared senior authorship**

15
16 **Correspondence to:**

17 Linda Drijvers, Radboud University, Donders Institute for Brain, Cognition and Behaviour,
18 Wundtlaan 1, 6525 XD, Nijmegen, The Netherlands. E-mail: linda.drijvers@mpi.nl, telephone:
19 +31 (0) 24 3521591

20
21
22
23

24 **Abstract**

25 During communication in real-life settings, the brain integrates information from auditory and
26 visual modalities to form a unified percept of our environment. In the current
27 magnetoencephalography (MEG) study, we used rapid invisible frequency tagging (RIFT) to
28 generate steady-state evoked fields and investigated the integration of audiovisual information in
29 a semantic context. We presented participants with videos of an actress uttering action verbs
30 (auditory; tagged at 61 Hz) accompanied by a gesture (visual; tagged at 68 Hz, using a projector
31 with a 1440 Hz refresh rate). Integration difficulty was manipulated by lower-order auditory
32 factors (clear/degraded speech) and higher-order visual factors (congruent/incongruent gesture).
33 We identified MEG spectral peaks at the individual (61/68 Hz) tagging frequencies. We
34 furthermore observed a peak at the intermodulation frequency of the auditory and visually tagged
35 signals ($f_{\text{visual}} - f_{\text{auditory}} = 7 \text{ Hz}$), specifically when lower-order integration was easiest because signal
36 quality was optimal. This intermodulation peak is a signature of nonlinear audiovisual integration,
37 and was strongest in left inferior frontal gyrus and left temporal regions; areas known to be
38 involved in speech-gesture integration. The enhanced power at the intermodulation frequency thus
39 reflects the ease of lower-order audiovisual integration and demonstrates that speech-gesture
40 information interacts in higher-order language areas. Furthermore, we provide a proof-of-principle
41 of the use of RIFT to study the integration of audiovisual stimuli, in relation to, for instance,
42 semantic context.

43

44 **Introduction**

45 During communication in real-life settings, our brain needs to integrate auditory input with visual
46 input in order to form a unified percept of the environment. Several magneto- and
47 electroencephalography (M/EEG) studies have demonstrated that integration of non-semantic
48 audiovisual inputs can occur as early as 50-100 ms after stimulus onset (e.g., Giard & Peronnet,
49 1999; Molholm et al., 2002; Talsma et al., 2010), and encompasses a widespread network of
50 primary sensory and higher-order regions (e.g., Beauchamp et al., 2004; Calvert, 2001; Werner &
51 Noppeney, 2010).

52 The integration of these audiovisual inputs has been studied using frequency tagging (Giani
53 et al., 2012; Regan et al., 1995). Here, an auditory or visual stimulus is periodically modulated at
54 a specific frequency, for example by modulating the luminance of a visual stimulus or the
55 amplitude of an auditory stimulus. This produces steady-state evoked potentials (SSEPs, SSEFs
56 for MEG) with strong power at the tagged frequency (for frequency-tagging in the visual domain
57 and steady-state visual evoked responses (SSVEP), see e.g. Norcia et al., 2015; Vialatte et al.,
58 2010; Gulbinaite et al., 2019, for frequency tagging in the auditory domain and auditory steady-
59 state responses (ASSR), see e.g. Baltus & Herrmann, 2015; Picton et al., 2003; Ross et al., 2005;
60 Ross et al., 2003). This technique is especially interesting in the context of studying audiovisual
61 integration, because it enables the tagging of an auditory stimulus and a visual stimulus at two

62 different frequencies (f_{visual} and f_{auditory}) in order to study whether and how these two inputs interact
63 in the brain. Previous work has suggested that when the auditory and visual signals interact, this
64 results in increased power at the intermodulation frequencies of the two stimuli (e.g., $|f_{\text{visual}} - f_{\text{auditory}}|$
65 or $f_{\text{visual}} + f_{\text{auditory}}$) (Regan & Regan, 1989). Such intermodulation frequencies arise from nonlinear
66 interactions of the two oscillatory signals. In the case of audio-visual integration, the
67 intermodulation likely reflects neuronal activity that combines the signals of the two inputs beyond
68 linear summation (Regan & Regan, 1988; Zemon & Ratliff, 1984).

69 However, other authors have reported inconclusive results on the occurrence of such
70 intermodulation frequencies as a signature of nonlinear audiovisual integration in neural signals.
71 Furthermore, this integration has so far only been studied in non-semantic contexts (e.g., the
72 integration of tones and gratings). For example, whereas Regan et al. (1995) identified
73 intermodulation frequencies (i.e., as a result of tagging an auditory and visual stimulus) in an area
74 close to the auditory cortex, Giani et al., (2012) identified intermodulation frequencies within (i.e.,
75 as a result of tagging two signals in the visual domain), but not between modalities (i.e., as a result
76 of tagging both an auditory and a visual signal).

77 In both of these previous studies, frequency tagging was applied at relatively low
78 frequencies (< 30 Hz for visual stimuli, < 40 Hz for auditory stimuli) (Giani et al., 2012; Regan et
79 al., 1995). This might be problematic, considering that spontaneous neuronal oscillations at lower
80 frequencies (e.g., alpha and beta oscillations) are likely entrained by frequency tagging (Keitel et
81 al., 2014; Spaak et al., 2014). In the current study, we use novel projector technology to perform
82 frequency tagging at high frequencies (rapid invisible frequency tagging; RIFT), and in a semantic
83 context. Previous work has demonstrated that neuronal responses to a rapidly flickering LED can
84 be driven and measured up to 100 Hz (Herrmann, 2001), and can successfully be used to study
85 sensory processing in the brain (Herring, 2017; Zhigalov et al., 2019). We here leverage these
86 rapid neural responses in order to circumvent the issue of endogenous rhythms interacting with
87 low-frequency tagging signals.

88 We use speech-gesture integration as a test case for studying rapid invisible frequency
89 tagging in a semantic context. Speech-gesture integration is a form of semantic audiovisual
90 integration that often occurs in natural, face-to-face communication. Previous behavioral and
91 neuroimaging studies have demonstrated that listeners process and integrate speech and gestures
92 at a semantic level, and that this integration relies on a network involving left inferior frontal gyrus
93 (LIFG), left-temporal regions (STS/MTG), motor cortex, and visual cortex (Dick et al., 2014;
94 Drijvers, Ozyürek, et al., 2018; Drijvers, Ozyurek, et al., 2018; Drijvers et al., 2019; Holle et al.,
95 2008, 2010; Kircher et al., 2009; Straube et al., 2012; Willems et al., 2007, 2009; Zhao et al.,
96 2018). Using frequency tagging in such a context to study whether intermodulation frequencies
97 can be identified as a signature of nonlinear audiovisual integration would provide a proof-of-
98 principle for the use of such a technique to study the integration of multiple inputs during complex
99 dynamic settings, such as multimodal language comprehension.

100 In the present study, we set out to explore whether RIFT can be used to identify
101 intermodulation frequencies as a result of the interaction between a visual and auditory tagged
102 signal in a semantic context. Participants watched videos of an actress uttering action verbs (tagged
103 at $f_{\text{auditory}} = 61$ Hz) accompanied by a gesture (tagged at $f_{\text{visual}} = 68$ Hz). Integration difficulty of
104 these inputs was modulated by auditory factors (clear/degraded speech) and visual factors
105 (congruent/incongruent gesture). For the visually tagged input, we expected power to be strongest
106 at 68 Hz in occipital regions. For the auditory tagged input, we expected power to be strongest at
107 61 Hz in auditory regions. We expected the interactions between the visually tagged and auditory
108 tagged signal to be non-linear in nature, resulting in spectral peaks at the intermodulation
109 frequencies of f_{visual} and f_{auditory} (i.e., $f_{\text{visual}} + f_{\text{auditory}}$ and $f_{\text{visual}} - f_{\text{auditory}}$). On the basis of previous
110 work (e.g., Drijvers, Ozyurek & Jensen, 2018a/b, 2019), we expected the locus of the
111 intermodulation frequencies to occur in LIFG and left-temporal regions such as pSTS/MTG, areas
112 known to be involved in speech-gesture integration.

113

114 **Methods**

115

116 **Participants**

117 Twenty-nine right-handed native Dutch-speaking adults (age range = 19 - 40, mean age = 23.68,
118 $SD = 4.57$, 18 female) took part in the experiment. All participants reported normal hearing, normal
119 or corrected-to-normal vision, no neurophysiological disorders and no language disorders. All
120 participants were recruited via the Max Planck Institute for Psycholinguistics participant database
121 and the Radboud University participant database, and gave their informed consent preceding the
122 experiment. Three participants (2 females) were excluded from the experiment due to unreported
123 metal in dental work (1) or excessive motion artifacts ($>75\%$ of trials affected) (2). The final data
124 set included the data of 26 participants.

125

126 **Stimulus materials**

127 Participants were presented with 160 video clips showing an actress uttering a highly-frequent
128 action verb accompanied by a matching or a mismatching iconic gesture (see for a detailed
129 description of pre-tests on recognizability and iconicity of the gestures, (Drijvers & Ozyürek,
130 2017)). All gestures used in the videos were rated as potentially ambiguous when viewed without
131 speech, which allowed for mutual disambiguation of speech and gesture (Habets et al., 2011).

132 In all videos, the actress was standing in front of a neutrally colored background, in
133 neutrally colored clothes. We predefined the verbs that would form the ‘mismatching gesture’, in
134 the sense that we asked the actress to utter the action verb, and depict the other verb in her gesture.
135 This approach was chosen because we included the face and lips of the actress in the videos, and

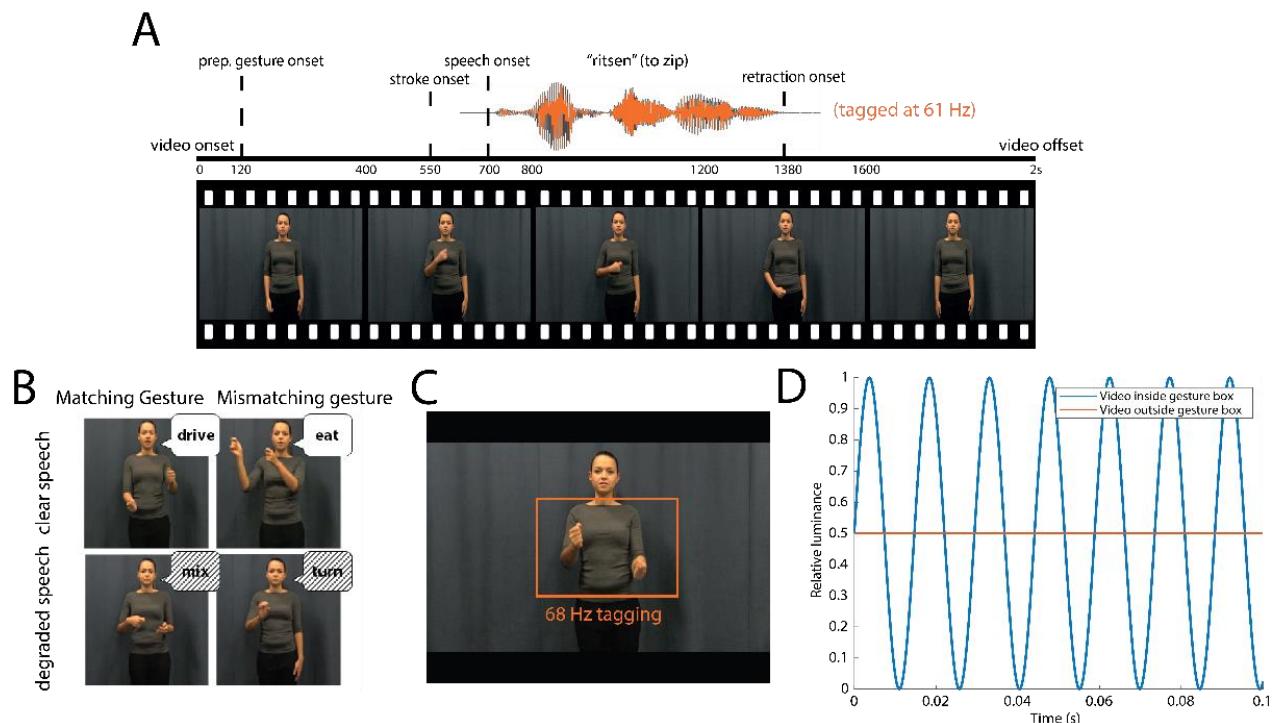


Figure 1 A. Illustration of the structure of the videos. Speech was amplitude-modulated at 61 Hz. B. Illustration of the different conditions. C. Area used for visual frequency tagging at 68 Hz. D. Illustration of luminance manipulation for visual-frequency tagging. D. Frequency tagging was achieved by multiplying the luminance of the pixels with a 68 Hz sinusoid. Modulation signal was equal to 0.5 at sine wave zero-crossing to preserve the mean luminance of the video, and was phase-locked across trials.

136 we did not want to recombine a mismatching audio track to a video to create the mismatch
 137 condition. Videos were on average 2000 ms long ($SD = 21.3$ ms). After 120 ms, the preparation
 138 (i.e., the first frame in which the hands of the actress moved) of the gesture started. On average, at
 139 550 ms ($SD = 74.4$ ms), the meaningful part of the gesture (i.e., the stroke) started, followed by
 140 speech onset at 680 ms ($SD = 112.54$ ms), and average speech offset at 1435 ms ($SD = 83.12$ ms)
 141 None of these timings differed between conditions. None of the iconic gestures were prescribed.
 142 All gestures were performed by the actress on the fly.

143 All audio files were intensity-scaled to 70 dB and denoised using *Praat* (Boersma &
 144 Weenink, 2015), before they were recombined with their corresponding video files using Adobe
 145 Premiere Pro. For 80 of the 160 sound files, we created noise-vocoded versions using *Praat*. Noise-
 146 vocoding pertains the temporal envelope of the audio signal, but degrades the spectral content
 147 (Shannon et al., 1995). We used 6-band noise-vocoding, as we demonstrated in previous work that
 148 this is the noise-vocoding level where the auditory signal is reliable enough for listeners to still be
 149 able to use the gestural information for comprehension (Drijvers & Ozyürek, 2017). To achieve
 150 this, we band-pass filtered the sound files between 50 and 8000 Hz in 6 logarithmically spaced
 151 frequency bands with cut-off frequencies at 50, 116.5, 271.4, 632.5, 1473.6, 3433.5 and 8000 Hz.
 152 These frequencies were used to filter white noise and obtain six noise bands. We extracted the
 153 amplitude envelope of each band using half-wave rectification and multiplied the amplitude

154 envelope with the noise bands. These bands were then recombined. Sound was presented to
155 participants using MEG-compatible air tubes.

156 We manipulated integration strength in the videos by auditory (clear/degraded) and visual
157 (congruent/incongruent) factors (see Figure 1). This resulted in four conditions: clear speech +
158 matching gesture (CM), clear speech + mismatching gesture (CMM), degraded speech + matching
159 gesture (DM) and degraded speech + mismatching gesture (DMM). These stimuli have been
160 thoroughly pretested and used in previous work on speech-gesture integration (e.g., Drijvers &
161 Ozyurek, 2017; Drijvers, Ozyurek & Jensen, 2018). All of the conditions contained 40 videos. All
162 verbs and gestures were only presented once. Participants were asked to pay attention to the videos
163 and identify what verb they heard in the videos in a 4-alternative forced choice identification task.

164

165 **Procedure**

166 Participants were tested in a dimly-lit magnetically shielded room and seated 70 cm from the
167 projection screen. All stimuli were presented using MATLAB 2016b (Mathworks Inc, Natick,
168 USA) and the Psychophysics Toolbox, version 3.0.11 (Brainard, 1997; Kleiner et al., 2007; Pelli,
169 1997). To achieve rapid invisible frequency tagging, we used a GeForce GTX960 2GB graphics
170 card with a refresh rate of 120 Hz, in combination with a PROPIxx DLP LED projector (VPixx
171 Technologies Inc., Saint-Bruno-de-Montarville, Canada), which can achieve a presentation rate up
172 to 1440 Hz. This high presentation rate is achieved by the projector interpreting the four quadrants
173 and three colour channels of the GPU screen buffer as individual smaller, grayscale frames, which
174 it then projects in rapid succession, leading to an increase of a factor 12 (4 quadrants * 3 colour
175 channels * 120 Hz = 1440 Hz) (User Manual for ProPixx, VPixx Technologies Inc., Saint-Bruno-
176 de-Montarville, Canada).

177 *Frequency tagging*

178 The area of the video that would be frequency-tagged was defined by the rectangle in which
179 all gestures occurred, which measured 10.0 by 6.5 degrees of visual angle (width by height). The
180 pixels within that area were always tagged at 68 Hz. This was achieved by multiplying the
181 luminance of the pixels within that square with a 68 Hz sinusoid (modulation depth = 100 %;
182 modulation signal equal to 0.5 at sine wave zero-crossing, in order to preserve the mean luminance
183 of the video), phase-locked across trials (see Figure 1D). For the auditory stimuli, frequency
184 tagging was achieved by multiplying the amplitude of the signal with a 61 Hz sinusoid, with a
185 modulation depth of 100 % (following (Lamminmäki et al., 2014)). In a pretest, we presented 11
186 native Dutch speakers with half of the stimuli containing the amplitude modulation, and half of
187 the stimuli not containing the amplitude modulation in both clear and degraded speech.
188 Participants were still able to correctly identify the amplitude modulated stimuli in clear speech
189 (mean % correct without amplitude modulation: 99.54, with amplitude modulation: 99.31) and in
190 degraded speech (mean % correct without amplitude modulation: 72.74, with amplitude

191 modulation: 70.23) and did not suffer more compared to when the signal was not amplitude
192 modulated.

193 Participants were asked to attentively watch and listen to the videos. Every trial started
194 with a fixation cross (1000 ms), followed by the video (2000 ms), a short delay period (1500 ms),
195 and a 4-alternative forced choice identification task (max 3000 ms, followed by the fixation cross
196 of the next trial as soon as a participant pressed one of the 4 buttons). In the 4-alternative forced
197 choice identification task, participants were presented with four written options, and had to identify
198 which verb they heard in the video by pressing one of 4 buttons on an MEG-compatible button
199 box. This task ensured that participants were attentively watching the videos, and to check whether
200 the verbs were understood. Participants were instructed not to blink during video presentation.

201 Throughout the experiment, we presented all screens at a 1440 Hz presentation rate. Brain
202 activity was measured using MEG, and was recorded throughout the experiment. The stimuli were
203 presented in four blocks of 40 trials each. The whole experiment lasted approximately 30 minutes,
204 and participants were allowed to take a self-paced break after every block. All stimuli were
205 presented in a randomized order per participant.

206

207 **MEG data acquisition**

208 MEG was recorded using a 275-channel axial gradiometer CTF MEG system (CTF MEG systems,
209 Coquitlam, Canada). We used an online low-pass filter at 300 Hz and digitized the data at 1200
210 Hz. All participants' eye gaze was recorded by an SR Research Eyelink 1000 eye tracker for
211 artifact rejection purposes. The head position of the participants was tracked in real time by
212 recording markers on the nasion, and left and right periauricular points (Stolk et al., 2013). This
213 enabled us to readjust the head position of participants relative to their original starting position
214 whenever the deviation was larger than 5 mm. After the experiment, T1-weighted structural
215 magnetic resonance images (MRI) were collected from 24 out of 26 participants using a Siemens
216 3T MAGNETOM Skyra system.

217

218 **MEG data analysis**

219 *Preprocessing*

220 All MEG data were analyzed using the FieldTrip toolbox (version 20180221) (Oostenveld et al.,
221 2011) running in a Matlab environment (2017b). All data were segmented into trials starting 1 s
222 before and ending 3 s after the onset of the video. The data were demeaned and line noise was
223 attenuated using a discrete Fourier transform approach at 50, 100 and 150 Hz. All trials that
224 contained jump artifacts or muscle artifacts were rejected using a semi-automatic routine. The data
225 were then down-sampled to 400 Hz. Independent component analysis (Bell & Sejnowski, 1995;
226 Jung et al., 2000) was used to remove residual eye movements and cardiac-related activity (average

227 number of components removed: 6.05). All data were then inspected on a trial-by-trial basis to
228 remove artifacts that were not identified using these rejection procedures. These procedures
229 resulted in rejection of 8.3 % of the trials. The number of rejected trials did not differ significantly
230 between conditions.

231 *Frequency tagging analyses - Sensor-level*

232 To investigate the response in auditory and visual regions to the frequency-tagged signal, we first
233 calculated event-related fields by averaging time-locked gradiometer data over trials, over
234 conditions, and over participants. All tagged stimuli were presented phase-locked over trials. We
235 used an approximation of planar gradiometer data to facilitate interpretation of the MEG data, as
236 planar gradient maxima are thought to be located above the neuronal sources that may underlie
237 them (Bastiaansen & Knösche, 2000). This was achieved by converting the axial gradiometer data
238 to orthogonal planar gradiometer pairs, which were combined by using root-mean-square (RMS)
239 for the ERFs. For the power analyses, we computed the power separately for the two planar
240 gradient directions, and combined the power data by averaging the two. To visualize the responses
241 per tagging frequency (Figure 3), we used a notch (i.e. band-stop) filter between 60 and 62 Hz to
242 display the ERF at 68 Hz, and a notch filter between 67 and 69 Hz to display the ERF at 61 Hz.

243 We then performed a spectral analysis on an individual's ERF data pooled over conditions,
244 in the time window in which both the auditory and visual stimulus unfolded (0.5 - 1.5 s), and a
245 post-stimulus baseline (2.0 - 3.0s). We chose this post-stimulus time window as a baseline because,
246 contrary to the pre-stimulus time window, it is not affected by the button press of the 4-alternative
247 forced choice identification task. We chose the 0.5-1.5 s time window to focus our analysis on,
248 because this time window captures both the meaningful part of the gesture and the full speech
249 signal. We computed power spectra in frequencies ranging from 1 to 130 Hz for both the baseline
250 and stimulus window using fast Fourier transform and a single Hanning taper of the 1s segments.
251 This data was then averaged over conditions, and the stimulus window was compared to the
252 baseline window.

253 *Frequency tagging analyses - Source-level*

254 To reconstruct activity at the tagging frequencies, we calculated coherence between a pure sine
255 wave at either 61 Hz or 68 Hz, reflecting the tagged stimulus, and the observed MEG signal at
256 those frequencies. Although the phase of the tagging was designed to be identical over trials, the
257 projector that we used occasionally experienced a brief delay in presenting the video material (in
258 16 of the 26 participants). We corrected for this by translating any observed delays between video
259 onset and offset markers (recorded in a stimulus trigger channel) into a phase-difference, which
260 was then subtracted from the tagging signal. Note that this correction only uses information in the
261 stimulus marker channel and the length of the original video files, and does not rely on any
262 information in the measured MEG signal.

263 We performed source analysis to identify the neuronal sources that were coherent with the
264 modulation signal at either 61 Hz or 68 Hz, and compared the difference in coherence in the
265 stimulus and post-stimulus window. This was done pooled over conditions. Source analyses on
266 coherence values (for 61 and 68 Hz) and power values (for the intermodulation frequency at 7 Hz,
267 see results), was performed using dynamic imaging of coherent sources (DICS; (Gross et al.,
268 2001)) as a beamforming approach. We computed a common spatial filter per subject from the
269 lead field matrix and the cross-spectral density matrix (CSD) that was the same for all conditions.
270 An individual's leadfield was obtained by spatially co-registering an individual's anatomical MRI
271 to the MEG data by the anatomical markers at the nasion and left and right periauicular points.
272 Then, for each participant, a single-shell head model was constructed on the basis of the MRI
273 (Nolte, 2003). A source model was created for each participant by warping a 10 mm spaced grid
274 defined in MNI space to the individual participant's segmented MRI. The MNI template brain was
275 used for those participants (2/26) for which an individual MRI scan was not available.

276 After establishing regions that showed elevated coherence with the tagged stimuli, we
277 proceeded to test the effect of the experimental conditions (clear versus degraded speech; matching
278 versus mismatching gesture) within these regions-of-interest (ROIs). The ROIs for the auditory
279 and visual tagged signals were defined by taking the grid points that exceeded 80 percent of the
280 peak coherence difference value between stimulus and baseline, across all conditions. For these
281 ROIs, coherence difference values were extracted per condition. Analogously, the ROI for the
282 intermodulation frequency at 7 Hz was defined by taking those grid points that exceeded 80 percent
283 of the peak power difference value between stimulus and baseline. The 80 percent threshold was
284 chosen as an exploratory threshold.

285 *Statistical comparisons*

286 As we predefined our frequencies of interest and have specific regions of interest for analysis, we
287 compared the differences between conditions using 2x2 repeated measures ANOVAs, with the
288 factors Speech (clear/degraded) and Gesture (matching/mismatching).

289

290 **Results**

291 Participants watched videos of an actress uttering action verbs in clear or degraded speech,
292 accompanied by a matching or mismatching gesture. After the video, participants were asked to
293 identify the verb they heard in a 4-alternative forced choice identification task, presented on the
294 screen in written form. Video presentation was manipulated by tagging the gesture space in the
295 video by 68 Hz flicker, while the sound in the videos was tagged by 61 Hz amplitude modulation
296 (see Figure 1).

297

298 **Behavioral results**

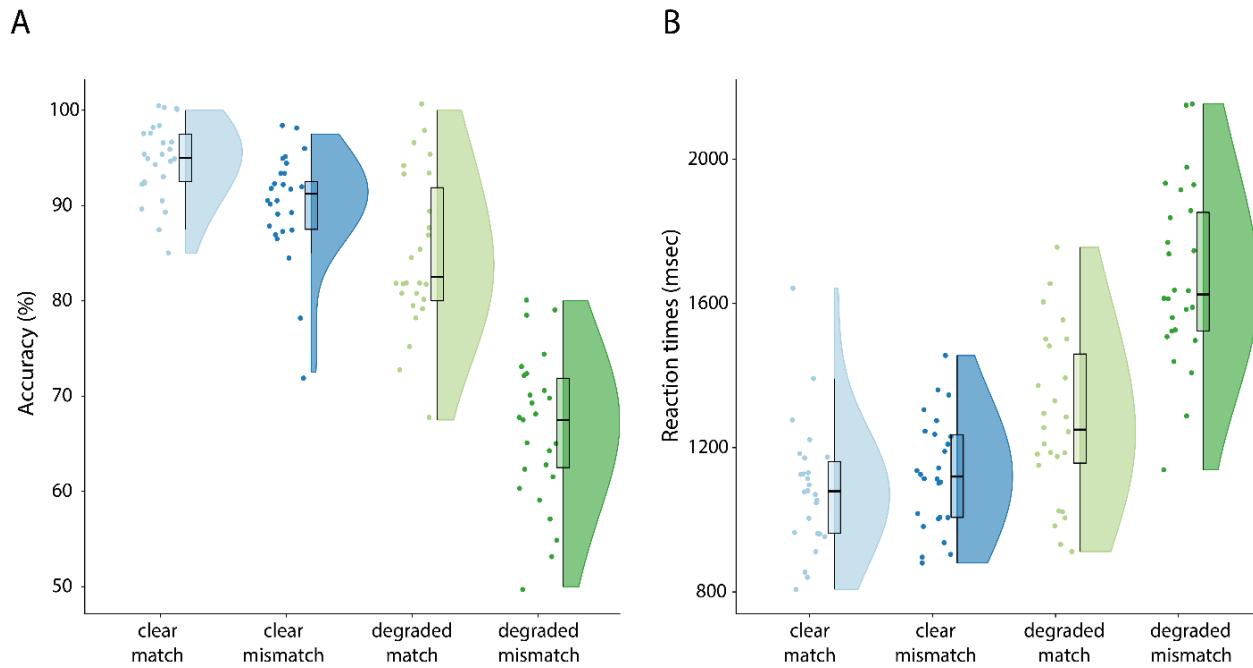


Figure 2 A: Accuracy results per condition. Response accuracy is highest for clear speech conditions, and when a gesture matches the speech signal. B: Reaction times per condition. Reaction times are faster in clear speech and when a gesture matches the speech signal. Raincloud plots reveal raw data, density and boxplots for coherence change.

299 In our behavioral task we replicated previous results (see Drijvers, Özyürek, et al., 2018; Drijvers
300 & Özyürek, 2018) and observed that when the speech signal was clear, response accuracy was
301 higher than when speech was degraded ($F(1, 25) = 301.60, p < .001$, partial $\eta^2 = .92$) (mean scores
302 and SDs: CM: 94.7% (SD = 4.0%), CMM: 90.2% (SD = 5.6%), DM: 85.0% (SD = 8.2%), DMM:
303 66.5% (SD = 7.8%)). Similarly, response accuracy was higher when a gesture matched compared
304 to mismatched the speech signal ($F(1, 25) = 184.29, p < .001$, partial $\eta^2 = .88$). The difference in
305 response accuracy was larger in degraded speech than in clear speech ($F(1, 25) = 4.87, p < .001$,
306 partial $\eta^2 = .66$) (see raincloud plots (Allen et al., 2019), Figure 2).

307 We observed similar results in the reaction times (RTs). Participants were faster to identify the
308 verbs when speech was clear, compared to when speech was degraded ($F(1, 25) = 198.06, p <$
309 $.001$, partial $\eta^2 = .89$) (mean RTs and SDs: CM: 1086.3 ms, SD = 177.1 ms, CMM: 1127.92 ms,
310 SD = 153.84 ms, DM: 1276.96 ms, SD = 230.13 ms, DMM: 1675.77 ms, SD = 246.69 ms).
311 Participants were faster to identify the verbs when the gesture matched the speech signal, compared
312 to when the gesture mismatched the speech signal ($F(1, 25) = 105.42, p < .001$, partial $\eta^2 = .81$).
313 This difference in reaction times was larger in degraded speech than in clear speech ($F(1, 25) =$
314 187.78, $p < .001$, partial $\eta^2 = .88$).

315 In sum, these results demonstrate that gestures facilitate speech comprehension when the
316 actress performed a matching gesture, but hindered comprehension when she performed a
317 mismatching gesture. This effect was larger in degraded speech than in clear speech.

318

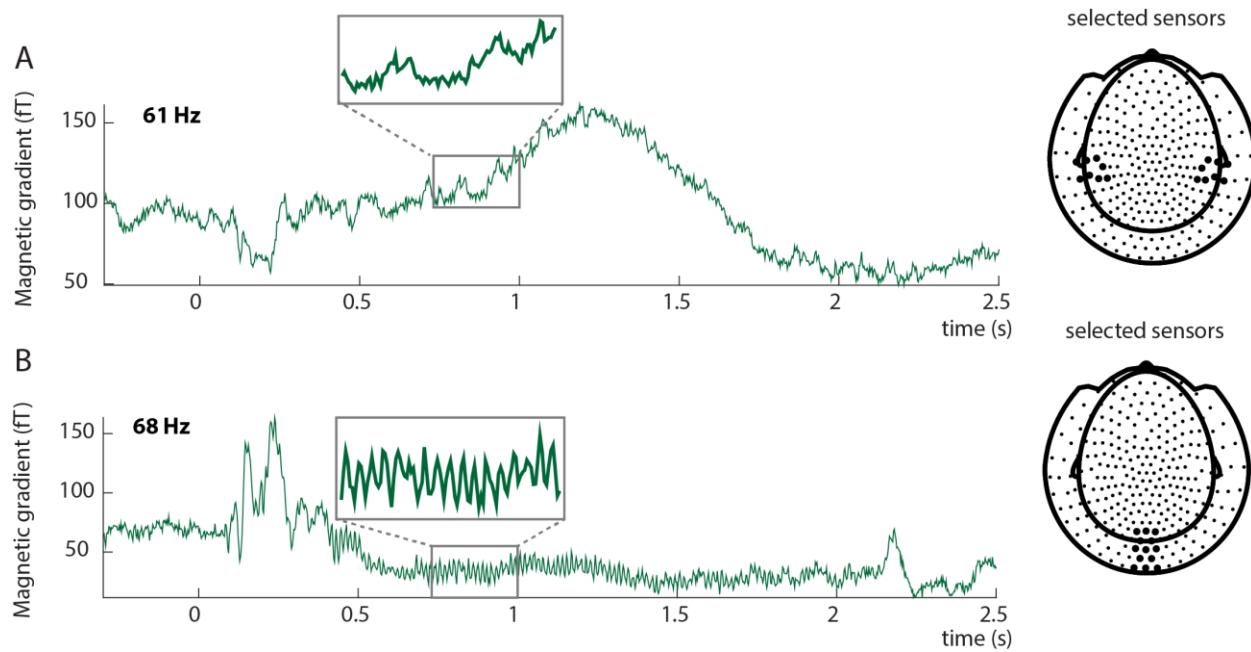


Figure 3: Event-related fields show clear responses at the tagged frequencies. Auditory input was tagged by 61 Hz amplitude modulation (A), Visual input was tagged by 68 Hz flicker (B). The insets reflect an enlarged part of the signal to clearly demonstrate the effect of the tagging on the event-related fields. Tagging was phase-locked over trials. A: Average ERF for a single subject at selected sensors overlying the left and right temporal lobe. The highlighted sensors in the right plot reflect the sensors for which the ERF is plotted. B: Average ERF for 68 Hz for a single subject at selected sensors overlying occipital cortex. The highlighted locations in the right plot reflect the sensors for which the ERF is plotted. ERFs show combined planar gradient data.

319 **MEG results - Frequency tagging**

320 *Both visual and auditory frequency tagging produce a clear steady-state response that is larger*
321 *than baseline*

322 As a first step, we calculated the time-locked averages of the event-related fields pooled over
323 conditions. Auditory frequency tagging at 61 Hz produced an auditory steady-state response over
324 left and right-temporal regions (see Figure 3A), and visual frequency tagging at 68 Hz produced a
325 clear visual steady-state response at occipital regions (see Figure 3B).

326 To explicitly compare the tagged signals between stimulus (0.5 – 1.5 s) and post-stimulus baseline
327 (2.0 – 3.0 s) periods, we plotted the difference in spectral power calculated from the ERF (i.e.
328 power of the time-locked average) in Figure 4. We observe that both visual and auditory responses
329 at the tagged frequency were reliable larger in the stimulus period than in the baseline (see below
330 for statistical assessment at the source level). Note that the visual tagged signal at 68 Hz seems to
331 be more focal and strong than the auditory tagged signal at 61 Hz (see Figure 4). These analyses
332 confirm that we were able to induce high-frequency steady-state responses simultaneously for both
333 auditory and visual stimulation.

334

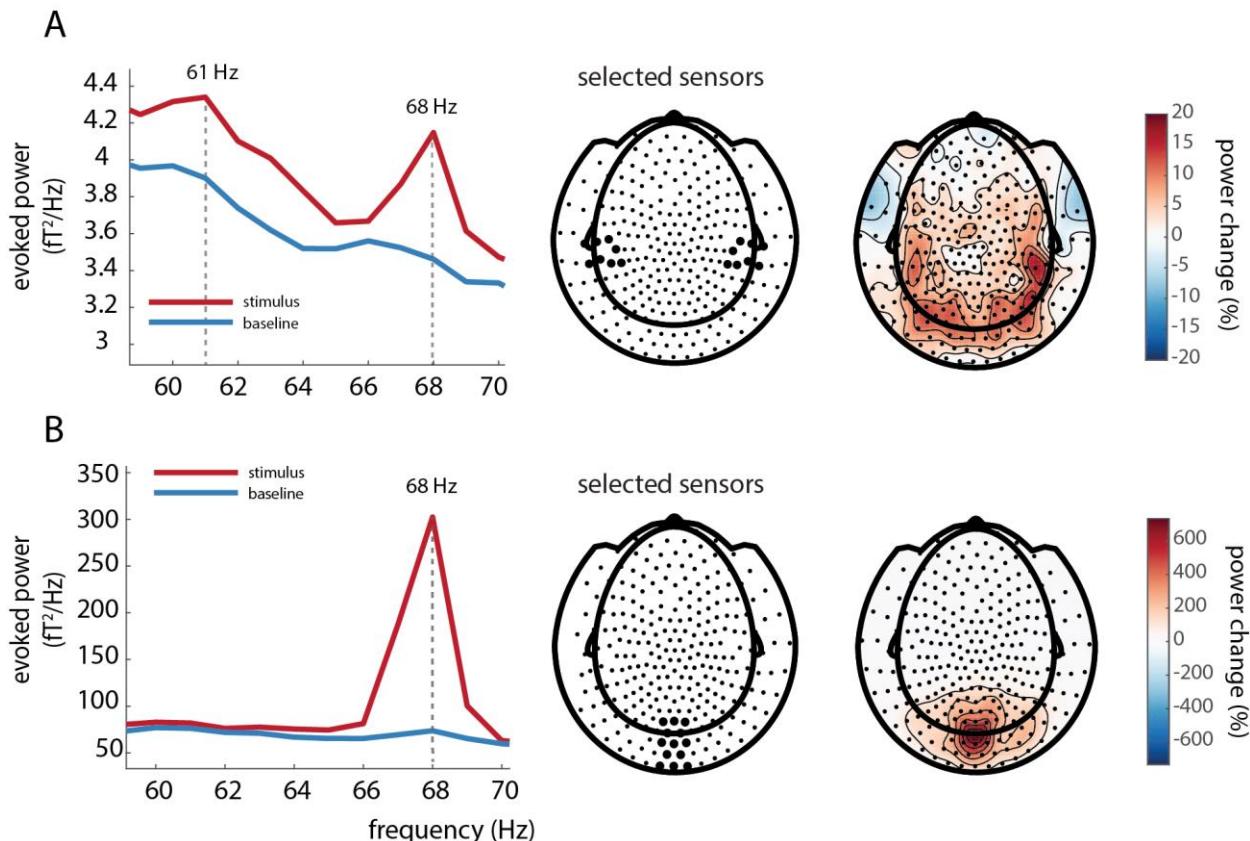


Figure 4: A: Power over auditory sensors peaks at the tagged frequency of the auditory stimulus (61 Hz). Note the visual 68 Hz tagged signal is still observable at left- and right-temporal sensors of interest. 61 Hz power is stronger in the stimulus interval than in the baseline interval, and is widely spread over posterior regions, with maxima at right-temporal regions. B: A power increase is observed at the tagged frequency (68 Hz) for the visual stimuli. 68 Hz power is larger in the stimulus than in the baseline window and is strongest over occipital regions.

335 *Coherence is strongest at occipital regions for the visually tagged signal (68 Hz) and strongest*
 336 *when speech is clear*

337 We proceeded to identify the neural generators of the tagged signals using beamformer source
 338 analysis. We computed source-level coherence coefficients for all conditions pooled together. This
 339 was done by computing coherence between a visual dummy 68 Hz modulation signal and the
 340 observed MEG data. The relative coherence increase between stimulus and baseline was largest in
 341 occipital regions (see Figure 5A), in an area consistent with early visual cortex.

342 To compare conditions, we then formed a visual ROI by selecting those grid points
 343 exceeding an exploratory threshold of 80 % of the peak coherence increase. For each participant,
 344 the percentage of change in coherence between stimulus and baseline was computed in that ROI
 345 per condition and compared in a 2x2 (Speech: clear/degraded, Gesture: matching/mismatching)
 346 RM-ANOVA (see Figure 5B). Coherence change was larger for videos containing clear speech

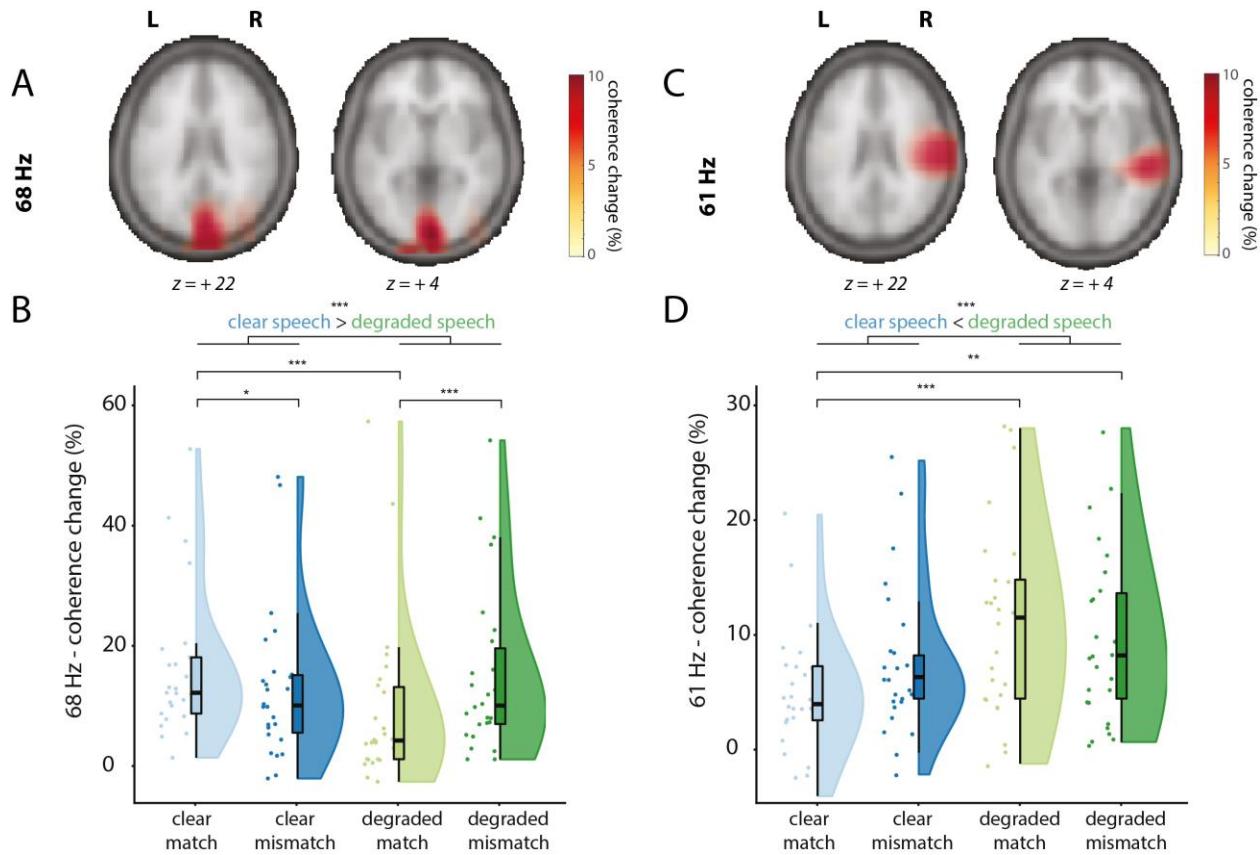


Figure 5: Sources of the visually tagged signal at 68 Hz (A/B) and sources of the auditory tagged signal at 61 Hz (C/D), and individual scores in the respective ROI per condition (clear match/clear mismatch/degraded match/degraded mismatch). Z-coordinates of slices are in mm and in MNI space. A: Coherence change in percentage when comparing coherence values in the stimulus window to a post-stimulus baseline for 68 Hz (the frequency of the visual tagging), pooled over conditions. Only positive coherence change values are plotted (>80% of peak maximum). Coherence change is largest over occipital regions for the visually tagged signal. B: Coherence change values in percentage extracted from the 68 Hz ROI. Raincloud plots reveal raw data, density and boxplots for coherence change. C: Coherence change in percentage when comparing coherence values in the stimulus window to a post-stimulus baseline for 61 Hz (the frequency of the auditory tagging), pooled over conditions. Only positive coherence values are plotted (>80% of peak maximum). Coherence change is largest over right-temporal regions. D: Coherence change values in percentage extracted from the 61 Hz ROI. Raincloud plots reveal raw data, density and boxplots for coherence change.

than videos containing degraded speech ($F(1, 25) = 17.14, p < .001$, partial $\eta^2 = .41$), but did not differ between matching or mismatching trials ($F(1, 25) = 0.025, p = .87$, partial $\eta^2 = .001$). We observed a significant interaction between Speech and Gesture ($F(1, 25) = 26.87, p < .001$, partial $\eta^2 = .52$). Post-hoc pairwise comparisons revealed a stronger coherence change in videos containing clear speech and a matching gesture (CM) than clear speech and a mismatching gesture (CMM) ($t(25) = 3.26, p = .015$), and a stronger coherence change in videos containing degraded speech and a mismatching gesture (DMM) than in videos containing degraded speech and a matching gesture (DM) ($t(25) = -4.03, p < .001$). Coherence change was larger in CM than in DM ($t(25) = 6.59, p < .001$), in CMM than DM ($t(25) = 2.93, p = .04$), but not larger in CM than in DMM ($t(25) = 2.02, p = .27$), and not larger in CMM compared to DMM ($t(26) = -1.74, p = .48$).

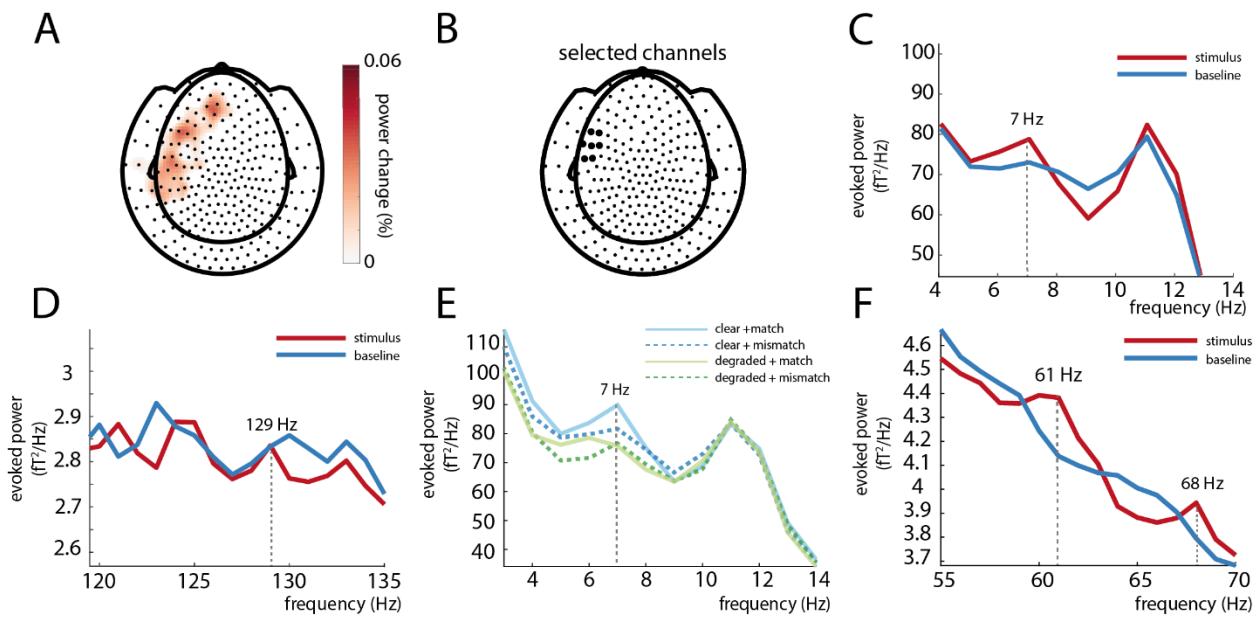


Figure 6: An intermodulation frequency could be observed at 7 Hz ($|f_{\text{visual}} - f_{\text{auditory}}|$) (A/C/E) but not 129 Hz ($f_{\text{visual}} + f_{\text{auditory}}$). (D). A: 7 Hz power in the stimulus window is larger than baseline over left-temporal and left-frontal sensors. Only positive values are plotted. B: Selected sensors (based on visual inspection). The black highlighted sensors represent the sensors at which the power spectra of the ERFs was calculated. C: Power spectra of 7 Hz (stimulus>baseline). D: No difference could be observed at 129 Hz between stimulus and baseline. E: Power spectra per condition. 7 Hz power peaks strongest in the clear+match condition. F: Power spectra of 61 Hz and 68 Hz over selected channels of 7 Hz power peak (see B).

357

358 These results thus indicate that visual regions responded stronger to the frequency-tagged gestural
 359 signal when speech was clear than when speech was degraded. This suggests that when speech is
 360 clear, participants allocate more visual attention to gestures than when speech is degraded,
 361 especially when a gesture matched the speech signal. When speech is degraded, participants
 362 allocate more attention to mismatching than to matching gestures.

363 *Coherence is strongest at right-temporal regions for the auditory tagged signal (61 Hz) and*
 364 *strongest when speech is degraded*

365 Similar to the visually tagged signal, we first computed coherence coefficients for all conditions
 366 pooled together. This was done by computing source-level coherence between a dummy 61 Hz
 367 modulation signal (reflecting the auditory tagging drive) and the observed MEG data. The
 368 coherence difference between stimulus and baseline peaked at right temporal regions (Figure 5C),
 369 in an area consistent with (right) early auditory cortex.

370 To compare conditions, we then formed the auditory ROI by selecting those grid points
 371 exceeding an exploratory threshold of 80 % of peak coherence change. Again, coherence change
 372 values per condition and per participant were compared in a 2x2 RM-ANOVA (see Figure 5D).
 373 Coherence change was larger in degraded speech conditions than in clear speech conditions ($F(1, 25) = 12.87, p = .001$, partial $\eta^2 = .34$), but did not differ between mismatching and matching
 374 conditions ($F(1, 25) = 0.09, p = .77$, partial $\eta^2 = .04$). No interaction effect was observed ($F(1, 25)$

376 = 3.13, $p = .089$, partial $\eta^2 = .11$). Post-hoc pairwise comparisons revealed that there was no
377 difference in coherence change when comparing CM and CMM ($t(25) = -1.44, p = .81$), or between
378 DM and DMM ($t(25) = 1.38, p = .90$). Coherence change was larger in DM than in CM ($t(25) = -$
379 4.24, $p < .001$), and in DMM than in CM ($t(25) = -3.90, p < .01$) but not when comparing CMM
380 to DMM ($t(25) = -1.40, p = .87$). These results thus indicate that right-lateralized auditory regions
381 processed the frequency-tagged auditory signal more strongly when speech was degraded than
382 when speech was clear. This suggests that when speech is degraded, participants allocate more
383 auditory attention to speech than when speech is clear.

384 *An intermodulation frequency was observed at 7 Hz ($|f_{\text{visual}} - f_{\text{auditory}}|$), but not at 129 Hz ($f_{\text{visual}} +$
385 f_{auditory})*

386 To test whether intermodulation frequencies ($|f_{\text{visual}} - f_{\text{auditory}}|$, $f_{\text{visual}} + f_{\text{auditory}}$) could be observed,
387 we then calculated power spectra of the ERFs in the stimulus time window and the post-stimulus
388 time window at 7 Hz and 129 Hz. Only for 7 Hz a difference between stimulus and baseline was
389 observed at left frontal and left temporal sensors (Figure 6A/C). No reliable differences were
390 observed for 129 Hz (Figure 6D). Interestingly, the spectral peak at 7 Hz during stimulus was most
391 pronounced for the clear/match condition (Figure 6E).

392 As a next step, we then took a similar approach as for the visual and auditory tagged stimuli
393 and calculated the coherence difference between stimulus and baseline at 7 Hz, pooled over
394 conditions. This was done by computing source-level coherence between a dummy 7 Hz
395 modulation signal (the intermodulation frequency of our 61 and 68 Hz tagging signals, specified
396 as the multiplication of the 61 and 68 Hz dummy signal) and the observed MEG data. The
397 coherence analysis did not reveal any differences between stimulus and baseline (see Figure 7A).
398 It should be noted here that our frequency-tagged signals at f_{auditory} and f_{visual} were exactly phase-
399 consistent across trials, since the phase was uniquely determined by the stimuli themselves.
400 However, it is possible that the phase of the intermodulation signal has a much weaker phase
401 consistency across trials, since it depends not only on the stimuli but also on the nature of the
402 nonlinear neural interaction. If this is the case, we might still observe an effect on the *power* at the
403 intermodulation frequency, rather than the coherence. We therefore performed source analysis on
404 the power of the combined conditions versus baseline. Here, we observed a power change at 7 Hz
405 in left frontal and temporal regions that mirrored the effect we observed at sensor level (Figure
406 7B).

407 The condition-averaged effect at the intermodulation frequency of 7 Hz is less striking than
408 at the primary tagged frequencies of 61 and 68 Hz, potentially due to it being driven mainly by
409 one of the four conditions only (see Figure 6E). Note that the 61 and 68 Hz signal were still present
410 over the left-frontotemporal sensors where we observed the 7Hz effect (see Figure 6F). As a next
411 step, and sticking to our a priori defined hypotheses and analysis plan, we again proceeded by
412 comparing conditions within an ROI defined by the condition-averaged contrast in source space.
413 As before, the ROI was defined as those grid points exceeding an exploratory threshold of 80 %

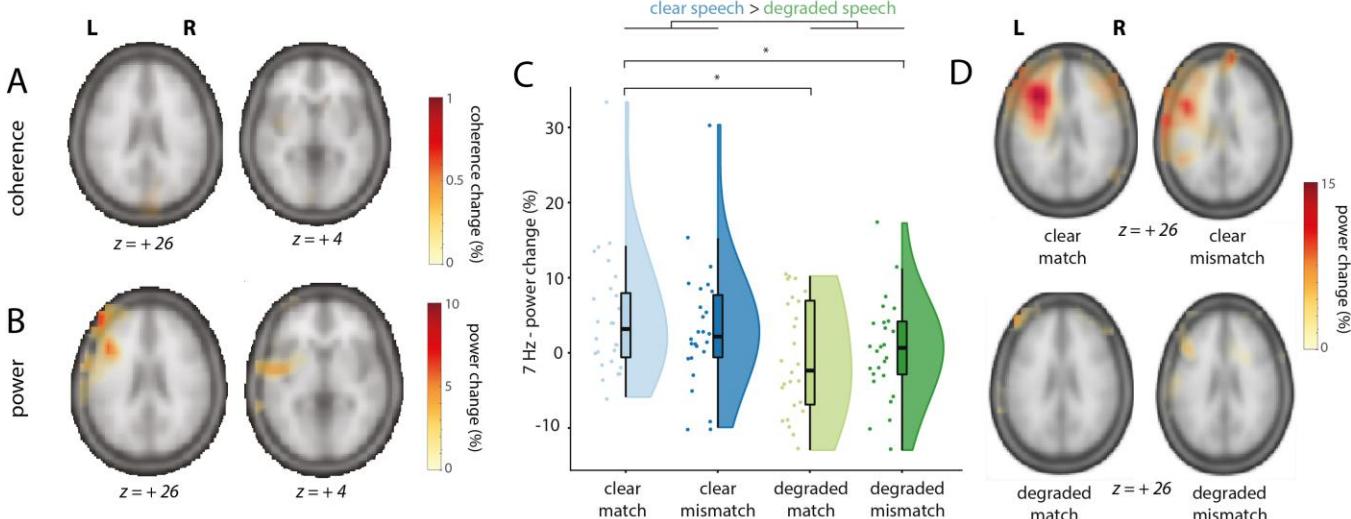


Figure 7: Sources of the intermodulation frequency ($f_{\text{visual}} - f_{\text{auditory}}$) at 7 Hz and individual scores in the left-frontotemporal ROI per condition (clear match/clear mismatch/degraded match/degraded mismatch). Z-coordinates of slices are in mm and in MNI space. A: Coherence change in percentage when comparing coherence values in the stimulus window to a post-stimulus baseline for 7 Hz (intermodulation frequency, $f_{\text{visual}} - f_{\text{auditory}}$), pooled over conditions. Only positive coherence values are plotted (> 80 % of maximum). No differences could be observed. B: Power change in percentage when comparing power values in the stimulus window to a post-stimulus baseline for 7 Hz, pooled over conditions. Power changes were largest in left-frontal and left-temporal regions. Highest peak value was at MNI coordinates -44, 24, 22, and extended from LIFG to pSTS/MTG. Only positive coherence values are plotted (> 80 % of maximum). C: Power change values in percentage extracted from the 7 Hz ROI in source space. Raincloud plots reveal raw data, density and boxplots for power change per condition. D: Power change in percentage when comparing power values in the stimulus window to a post-stimulus baseline for 7Hz, per condition.

414 of the peak power change from baseline to stimulus epochs. We compared the strength of the 7 Hz
 415 signal at source level between conditions by using a 2x2 RM-ANOVA (Figure 7C). Power change
 416 was larger in clear speech conditions than in degraded speech conditions ($F(1, 25) = 10.26, p =$
 417 $.004$, partial $\eta^2 = .29$), but did not differ between matching and mismatching trials ($F(1, 25) =$
 418 $0.01, p = .91$, partial $\eta^2 = .001$), suggesting an effect of speech degradation, but not of semantic
 419 congruency. No interaction effect was observed ($F(1, 25) = 1.27, p = .27$, partial $\eta^2 = .05$). Post-
 420 hoc pairwise comparisons revealed that 7 Hz power was not different for CM compared to CMM
 421 ($t(25) = 1.14, p = 1$), and not different for DM compared to DMM ($t(25) = -.67, p = 1$). However,
 422 7 Hz power was larger in CM than in DM ($t(25) = 3.01, p = .025$), and larger in CM than in DMM
 423 ($t(25) = 2.82, p = .045$). No difference was observed between CMM and DMM ($t(25) = 1.61, p =$
 424 $.6$). To rule out that these differences in 7 Hz power were due to general power differences in the
 425 theta band, we compared the strength of 6 Hz and 8 Hz between conditions, using two 2x2 RM-
 426 ANOVA's. Here, no differences between conditions were observed (all $p > 0.05$), suggesting this
 427 was specific to the 7 Hz signal. These results are also in line with previous MEG studies on speech-
 428 gesture integration, where no differences in theta power were observed (Drijvers, Ozyürek, et al.,
 429 2018; Drijvers, Ozyürek, et al., 2018b; Drijvers, van der Plas, et al., 2019).

430 In addition to our ROI-based analysis, we present the full beamformer source maps of 7
 431 Hz power (stimulus versus baseline) for the four conditions in Figure 7D. These reveal results fully
 432 compatible with the aforementioned RM-ANOVA. Furthermore, they show that our ROI selection
 433 on the condition-averaged response versus baseline was likely suboptimal, since the source map

434 for CM shows a more clearly elevated intermodulation cluster than the average (in line with the
435 sensor-level results shown in Figure 6A).

436 These results thus demonstrate that we could reliably observe an intermodulation signal
437 when speech was clear and a gesture matched the speech signal. Left-frontotemporal regions
438 showed a stronger intermodulation peak (reflecting the lower-order interaction between the
439 auditory and visually tagged signal) when speech was clear than when speech was degraded. This
440 suggests that the interaction between the auditory and visual tagged signal is strongest when signal
441 quality was optimal and speech was clear.

442

443 **Discussion**

444 In the current MEG study we provide a proof-of-principle that rapid invisible frequency tagging
445 (RIFT) can be used to estimate task-dependent neuronal excitability in visual and auditory areas,
446 as well as the auditory-visual interaction. Coherence was strongest over occipital regions for the
447 visual-tagged input, and strongest when speech was clear. Coherence was strongest over right-
448 temporal regions for the auditory-tagged input and strongest when speech was degraded.
449 Importantly, we identified an intermodulation frequency at 7 Hz ($f_{\text{visual}} - f_{\text{auditory}}$) as a result of the
450 interaction between a visual frequency-tagged signal (gesture; 68 Hz) and an auditory frequency-
451 tagged signal (speech; 61 Hz). In line with our hypotheses, power at this intermodulation frequency
452 was strongest in LIFG and left-temporal regions (pSTS/MTG), and was strongest when the lower-
453 order integration of auditory and visual information was optimal (i.e., when speech was clear).
454 Below we provide interpretations of these results.

455

456 **Clear speech enhances visual attention to gestural information**

457 In occipital regions, we observed a stronger drive by the 68 Hz visual modulation signal when
458 speech was clear than when speech was degraded. We speculate that this effect reflects that
459 listeners allocate more visual attention to gestures when speech is clear. This speculative
460 interpretation is in line with previous eye-tracking work that demonstrated that when speech is
461 degraded, listeners gaze more often to the face and mouth than to gestures to extract phonological
462 information to aid comprehension (Drijvers, Vaitonytè, et al., 2019), as well as previous work that
463 revealed that the amplitude of SSVEPs was enhanced by visual attention, irrespective of whether
464 the stimuli were task-relevant (Morgan et al., 1996; Müller et al., 2006). Note that gestural
465 information is often processed in the periphery of a listener's visual field (Gullberg & Holmqvist,
466 1999, 2002, 2006; Gullberg & Kita, 2009). As listeners do not necessarily need to extract the
467 phonological information conveyed by the lips when speech is clear, overt visual attention might
468 be directed to a 'resting' position in the middle of the screen during clear speech processing,
469 resulting in stronger coherence with the visual drive when speech is clear than when speech is
470 degraded. Pairwise comparisons of the conditions revealed that in clear speech, coherence was

471 larger when the gesture matched, rather than mismatched, the signal. In line with the interpretation
472 above, a listener might have reconsidered the auditory input when noticing that the gesture
473 mismatched the perceived auditory input, and might have directed their attention to the face/lips
474 of the actress, which, in turn, reduces visual attention to the gesture.

475 However, we observed the opposite effect when speech was degraded; i.e. a stronger
476 coherence when the gesture mismatched, rather than matched, the degraded speech signal. We
477 speculate that when speech is degraded and a gesture matches the signal, a listener might more
478 strongly allocate visual attention to the information conveyed by the face/lips, so that information
479 conveyed by the lips and the information conveyed by the gesture can jointly aid in disambiguating
480 the degraded speech signal (Drijvers & Ozyürek, 2017). However, when speech is degraded and
481 a gesture mismatches the signal, the uncertainty of both inputs may result in a reconsideration of
482 both inputs, and thus a less fixed locus of attention (see also Nath & Beauchamp, 2011 for work
483 on perceptual reliability weighting in clear and degraded speech). These interpretations are rather
484 speculative, and further work is needed to disambiguate different interpretations. For example,
485 future work could consider tagging the mouth-region to further investigate how a listener allocates
486 visual attention to these two visual articulators during comprehension

487

488 **Degraded speech enhances auditory attention to speech information**

489 In line with our hypotheses, we observed stronger drive by the 61 Hz amplitude modulation signal
490 in temporal areas overlapping with auditory cortex when speech was degraded than when speech
491 was clear. This response was strongest at right-temporal regions, which is in line with previous
492 work that demonstrated that for speech stimuli, the ASSR is often localized to right-lateralized
493 sources (Lamminmäki et al., 2014; Ross et al., 2005). Although both left- and right-hemispheres
494 process speech, a right-lateralized dominance is often observed because right-lateralized regions
495 are sensitive to spectral changes and prosodic information, and processing of low-level auditory
496 cues (Zatorre & Gandour, 2008; Scott et al., 2000).

497 Previous work has reported enhanced ASSR responses to amplitude-modulated multi-
498 speech babble when attention to this input increases (Keitel et al., 2011; Ross et al., 2004; Saupe
499 et al., 2009; Talsma et al., 2010; Tiitinen et al., 1993). The enhanced ASSR which we observed in
500 the degraded compared to clear speech conditions could thus reflect an increase in attention to the
501 speech signal when speech is degraded. Note that no differences in coherence were observed when
502 comparing matching and mismatching gestures in either clear or degraded speech. As the gesture
503 congruency manipulation is a visual manipulation, this indicates that modulation of the ASSR is
504 modality-specific (Parks et al., 2011; Rees et al., 2001).

505

506 **The auditory tagged speech signal and visual tagged gesture signal interact in left- 507 frontotemporal regions**

508 We set out to study whether intermodulation frequencies could be identified in a multimodal,
509 semantic context as a result of the interaction of the visual and auditory tagged signals. In contrast
510 to previous work by (Giani et al., 2012) using lower frequencies, we did observe an
511 intermodulation frequency at 7 Hz ($f_{\text{visual}} - f_{\text{auditory}}$), but not at 129 Hz ($f_{\text{visual}} + f_{\text{auditory}}$). As responses
512 in lower frequencies tend to be stronger than in higher frequencies, the higher-frequency
513 intermodulation frequency might not have been identifiable as neurons cannot be driven in this
514 fast range.

515 Note that although we observed a stronger 7 Hz power peak at sensor level in the stimulus
516 interval compared to the baseline interval, we did not observe stronger coherence between a 7 Hz
517 dummy signal and the observed MEG data at source level. This indicates that the phase of the
518 intermodulation signal is not as consistent over trials as the f_{visual} and f_{auditory} signals, which in turn
519 might imply that the time point of interaction of the two signals differs across trials. This could
520 explain why we observed a clear difference between stimulus and baseline when we reconstructed
521 the sources of the intermodulation frequency on the basis of power, but not coherence.

522 We observed a reliable peak at 7 Hz power during stimulation when integration of the
523 lower-order auditory and visual input was optimal, i.e., when speech was clear. In line with our
524 hypotheses, the source of the intermodulation frequency was localized in LIFG and left-temporal
525 (pSTS/MTG) regions. It has been shown that these areas are involved in the integration of speech
526 and gestures (Dick et al., 2014; Drijvers, Ozyurek, et al., 2018; Drijvers, Ozyurek, et al., 2018;
527 Drijvers, van der Plas, et al., 2019; Holle et al., 2008, 2010; Kircher et al., 2009; Straube et al.,
528 2012; Willems et al., 2007, 2009; Zhao et al., 2018). There are, however, important differences
529 between the interpretation of the intermodulation frequency in this work, and the results observed
530 in response to higher-order speech-gesture integration in previous work.

531 First, although previous work has observed effects related to higher-order integration in
532 LIFG and pSTS/MTG, the observed intermodulation frequency in the current work is most likely
533 related to lower-order integration. Specifically, we observed that power at the intermodulation
534 frequency was stronger in clear speech conditions than in degraded speech conditions, but we did
535 not observe an effect of gesture congruency. We therefore propose that, contrary to our hypotheses,
536 power at the intermodulation frequency does not reflect the integration of higher-order semantic
537 audiovisual integration, but rather is a direct reflection of the non-linear integration of lower-order
538 speech and gesture information. This difference might be explained by considering that the
539 intermodulation frequency is unable to capture higher-order effects that result from lexical access
540 on the basis of the auditory and visual input. Second, the current work is not able to dissociate
541 between the different roles of the LIFG and pSTS/MTG in the speech-gesture integration process.
542 The accuracy of source modeling using MEG should be considered in the light of the inverse
543 problem (Baillet, 2017). This limits our ability to make precise claims about the exact locus of the
544 observed effect when comparing to fMRI (see e.g., Papeo et al., 2019, for a functional distinction
545 of different subregions of the MTG in the speech-gesture integration process). Furthermore, fMRI
546 is sensitive to modulations in the BOLD signal whereas MEG detects changes in neuronal

547 synchronization. As such, these techniques provide complementary but not necessarily
548 overlapping information on neuronal activation.

549 **Proof of principle: using RIFT to study the integration of complex and dynamic audiovisual
550 stimuli in a semantic context.**

551 The current MEG study provides a proof of principle of the use of rapid invisible frequency tagging
552 (RIFT) to study the integration of audiovisual stimuli, and is the first study to identify
553 intermodulation frequencies as a result of the lower-order interaction between auditory and visual
554 stimuli in a semantic context. Note that although previous work has reported the occurrence of
555 intermodulation frequencies in a non-semantic context (Regan et al., 1995), other studies have
556 failed to identify between-modality intermodulation frequencies (Giani et al., 2012). This could be
557 due to the fact that lower frequencies were used for tagging. Another possibility is that this was
558 due to the nature of the stimuli used in these studies. As Giani et al., (2012) suggest, the occurrence
559 of intermodulation frequencies resulting from audiovisual integration of non-semantic inputs such
560 as tones and gratings might reflect low-level spatiotemporal coincidence detection that is
561 prominent for transient stimuli, but less so for sustained steady-state responses. Similarly, previous
562 fMRI work that investigated the difference between transient and sustained BOLD responses
563 revealed that primary auditory and visual regions were only involved in the integration of rapid
564 transient stimuli at stimulus onset. However, integration for sustained responses did involve
565 higher-order areas (Werner & Noppeney, 2011). The observed 7 Hz intermodulation frequency in
566 response to our semantic audiovisual stimuli was also localized to higher-order areas, rather than
567 early sensory regions. This again underlines the possibility that the observed intermodulation
568 frequency in the current study reflects the ease of lower-order integration of these audiovisual
569 stimuli in certain higher-order regions.

570 An important advantage of using RIFT is that spontaneous neuronal oscillations in lower
571 frequencies were not entrained by our tagging frequencies. This might explain why a clear
572 intermodulation frequency was observed in the current study, but was less easy to identify in
573 previous work. Future studies might consider exploiting this feature and using RIFT to study the
574 interaction of these endogenous lower frequency oscillations with the tagged signals, in order to
575 elucidate their role in sensory processing. However, future work should also consider that high-
576 frequency tagging might entrain spontaneous neuronal oscillations at higher frequencies. Although
577 this was not directly relevant for the identification of the intermodulation frequency in this study,
578 and we did not observe any gamma band modulations in response to the stimuli used in this study
579 in earlier work (Drijvers, Ozyurek & Jensen, 2018b), it should be noted that gamma band
580 modulations have been observed in other work related to linguistic semantic processing (e.g., in
581 the 30-50 Hz range in Mellem et al., 2013; Wang et al., 2018).

582 **Conclusion**

583 First of all, we provided a proof of principle that RIFT can be used to tag visual and auditory inputs
584 at high frequencies, resulting in clear spectral peaks in the MEG signal, localized to early sensory

585 cortices. Second, we demonstrated that RIFT can be used to identify intermodulation frequencies
586 in a multimodal, semantic context. The observed intermodulation frequency was the result of the
587 nonlinear interaction between visual and auditory tagged stimuli. Third, the intermodulation signal
588 was localized to LIFG and pSTS/MTG, areas known to be involved in speech-gesture integration.
589 The strength of this intermodulation frequency was strongest when lower-order signal quality was
590 optimal. In conclusion, we thus propose that the strength of this intermodulation frequency reflects
591 the ease of lower-order audiovisual integration, that RIFT can be used to study both unimodal
592 sensory signals as well as their multimodal interaction in downstream higher-order areas, and that
593 RIFT has many use cases for future work.

594

595 **Acknowledgements**

596 This work was supported by Gravitation Grant 024.001.006 of the Language in Interaction
597 Consortium from Netherlands Organization for Scientific Research (NWO). OJ was supported by
598 James S. McDonnell Foundation Understanding Human Cognition Collaborative Award
599 [220020448] and the Royal Society Wolfson Research Merit Award. LD was supported by the
600 European Research Council (grant #773079-CoAct awarded to J.Holler). ES was supported by
601 NWO (Veni grant 016.Veni.198.065). We are very grateful to Nick Wood, for helping us in editing
602 the video stimuli, and to Gina Ginos, for being the actress in the videos.

603

604 **References**

605 Allen, M., Poggiali, D., Whitaker, K., Marshall, T. R., & Kievit, R. A. (2019). Raincloud plots: a
606 multi-platform tool for robust data visualization. *Wellcome open research*, 4.

607 Baillet, S. (2017). Magnetoencephalography for brain electrophysiology and imaging. *Nature
608 Neuroscience*, 20(3), 327–339. <https://doi.org/10.1038/nn.4504>

609 Baltus, A., & Herrmann, C. S. (2015). Auditory temporal resolution is linked to resonance
610 frequency of the auditory cortex. *International Journal of Psychophysiology*, 98(1), 1–7.
611 <https://doi.org/10.1016/j.ijpsycho.2015.08.003>

612 Bastiaansen, M. C. M., & Knösche, T. R. (2000). Tangential derivative mapping of axial MEG
613 applied to event-related desynchronization research. *Clinical Neurophysiology*, 111(7),
614 1300–1305. [https://doi.org/10.1016/S1388-2457\(00\)00272-8](https://doi.org/10.1016/S1388-2457(00)00272-8)

615 Beauchamp, M. S., Argall, B. D., Bodurka, J., Duyn, J. H., & Martin, A. (2004). Unraveling
616 multisensory integration: patchy organization within human STS multisensory cortex.
617 *Nature Neuroscience*, 7(11), 1190–1192. <https://doi.org/10.1038/nn1333>

618 Bell, A. J., & Sejnowski, T. J. (1995). An Information-Maximization Approach to Blind
619 Separation and Blind Deconvolution. *Neural Computation*, 7(6), 1129–1159.
620 <https://doi.org/10.1162/neco.1995.7.6.1129>

621 Boersma, P., & Weenink, D. (2015). *Praat: doing phonetics by computer*. Praat: Doing
622 Phonetics by Computer [Computer Program].

623 Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, 10, 433–436.

624 Calvert, G. A. (2001). Crossmodal Processing in the Human Brain : Insights from Functional
625 Neuroimaging Studies. *Cerebral Cortex*, 11, 1110–1123.

626 Dick, A. S., Mok, E. H., Raja Beharelle, A., Goldin-Meadow, S., & Small, S. L. (2014). Frontal
627 and temporal contributions to understanding the iconic co-speech gestures that
628 accompany speech. *Human Brain Mapping*, 35(3), 900–917.
629 <https://doi.org/10.1002/hbm.22222>

630 Drijvers, L, Ozyürek, A., & Jensen, O. (2018). Alpha and Beta Oscillations Index Semantic
631 Conguency between Speech and Gestures in Clear and Degraded Speech. *Journal of*
632 *Cognitive Neuroscience*, 30(8), 1086–1097. <https://doi.org/10.1162/jocn>

633 Drijvers, L, Ozyurek, A., & Jensen, O. (2018). Hearing and seeing meaning in noise: Alpha,
634 beta, and gamma oscillations predict gestural enhancement of degraded speech
635 comprehension. *Human Brain Mapping*, January, 1–13.
636 <https://doi.org/10.1002/hbm.23987>

637 Drijvers, L, & Ozyürek, A. (2017). Visual Context Enhanced: The Joint Contribution of Iconic
638 Gestures and Visible Speech to Degraded Speech Comprehension. *Journal of Speech,
639 Language & Hearing Research*, 60, 212–222. [https://doi.org/10.1044/2016_JSLHR-H-16-0101](https://doi.org/10.1044/2016_JSLHR-H-
640 16-0101)

641 Drijvers, L, & Özyürek, A. (2018). Native language status of the listener modulates the neural
642 integration of speech and iconic gestures in clear and adverse listening conditions. *Brain
643 and Language*, 177–178, 7–17. <https://doi.org/10.1016/j.bandl.2018.01.003>

644 Drijvers, L, Vaitonytè, J., & Özyürek, A. (2019). Degree of Language Experience Modulates
645 Visual Attention to Visible Speech and Iconic Gestures During Clear and Degraded
646 Speech Comprehension. *Cognitive Science*, 43(10). <https://doi.org/10.1111/cogs.12789>

647 Drijvers, L, van der Plas, M., Özyürek, A., & Jensen, O. (2019). Native and non-native listeners
648 show similar yet distinct oscillatory dynamics when using gestures to access speech in
649 noise. *NeuroImage*, 194, 55–67. <https://doi.org/10.1016/j.neuroimage.2019.03.032>

650 Giani, A. S., Ortiz, E., Belardinelli, P., Kleiner, M., Preissl, H., & Noppeney, U. (2012). Steady-
651 state responses in MEG demonstrate information integration within but not across the
652 auditory and visual senses. *NeuroImage*, 60(2), 1478–1489.
653 <https://doi.org/10.1016/j.neuroimage.2012.01.114>

654 Giard, M. H., & Peronnet, F. (1999). Auditory-Visual Integration during Multimodal Object
655 Recognition in Humans: A Behavioral and Electrophysiological Study. *Journal of
656 Cognitive Neuroscience*, 11(5), 473–490. <https://doi.org/10.1162/08989299563544>

657 Gross, J., Kujala, J., Hamalainen, M., Timmermann, L., Schnitzler, A., & Salmelin, R. (2001).
658 Dynamic imaging of coherent sources: Studying neural interactions in the human brain.

659 *Proceedings of the National Academy of Sciences of the United States of America*, 98(2),

660 694–699. <https://doi.org/10.1073/pnas.98.2.694>

661 Gullberg, M., & Holmqvist, K. (1999). Keeping an eye on gestures: Visual perception of

662 gestures in face-to-face communication. *Pragmatics & Cognition*, 7(1), 35–63.

663 <https://doi.org/10.1075/pc.7.1.04gul>

664 Gullberg, M., & Holmqvist, K. (2002). Visual Attention towards Gestures in Face-to-Face

665 Interaction vs . on Screen *. *International Gesture Workshop*, 206–214.

666 Gullberg, M., & Holmqvist, K. (2006). What speakers do and what addressees look at: Visual

667 attention to gestures in human interaction live and on video. *Pragmatics & Cognition*,

668 14(1), 53–82. <https://doi.org/10.1075/pc.14.1.05gul>

669 Gullberg, M., & Kita, S. (2009). Attention to Speech-Accompanying Gestures: Eye Movements

670 and Information Uptake. *Journal of Nonverbal Behavior*, 33(4), 251–277.

671 <https://doi.org/10.1007/s10919-009-0073-2>

672 Habets, B., Kita, S., Shao, Z., Ozyurek, A., & Hagoort, P. (2011). The role of synchrony and

673 ambiguity in speech-gesture integration during comprehension. *Journal of Cognitive*

674 *Neuroscience*, 23(8), 1845–1854. <https://doi.org/10.1162/jocn.2010.21462>

675 Herring, J. D. (2017). *Driving visual cortex to study neuronal oscillations*.

676 Herrmann, C. S. (2001). Human EEG responses to 1-100 Hz flicker: Resonance phenomena in

677 visual cortex and their potential correlation to cognitive phenomena. *Experimental Brain*

678 *Research*, 137(3–4), 346–353. <https://doi.org/10.1007/s002210100682>

679 Holle, H., Gunter, T. C., Ruschemeyer, S. A., Hennenlotter, A., & Iacoboni, M. (2008). Neural

680 correlates of the processing of co-speech gestures. *NeuroImage*, 39(4), 2010–2024.

681 <https://doi.org/10.1016/j.neuroimage.2007.10.055>

682 Holle, H., Obleser, J., Rueschemeyer, S.-A., & Gunter, T. C. (2010). Integration of iconic
683 gestures and speech in left superior temporal areas boosts speech comprehension under
684 adverse listening conditions. *NeuroImage*, 49(1), 875–884.
685 <https://doi.org/10.1016/j.neuroimage.2009.08.058>

686 Jung, T.-P. P., Makeig, S., Humphries, C., Lee, T.-W. W., McKeown, M. J., Iragui, V., &
687 Sejnowski, T. J. (2000). Removing electroencephalographic artifacts by blind source
688 separation. *Psychophysiology*, 37(2), 163–178. <https://doi.org/10.1111/1469->
689 8986.3720163

690 Keitel, C., Quigley, C., & Ruhnau, P. (2014). Stimulus-Driven Brain Oscillations in the Alpha
691 Range: Entrainment of Intrinsic Rhythms or Frequency-Following Response? *Journal of*
692 *Neuroscience*, 34(31), 10137–10140. <https://doi.org/10.1523/JNEUROSCI.1904-14.2014>

693 Keitel, C., Schröger, E., Saupe, K., & Müller, M. M. (2011). Sustained selective intermodal
694 attention modulates processing of language-like stimuli. *Experimental Brain Research*,
695 213(2–3), 321–327. <https://doi.org/10.1007/s00221-011-2667-2>

696 Kircher, T., Straube, B., Leube, D., Weis, S., Sachs, O., Willmes, K., Konrad, K., & Green, A.
697 (2009). Neural interaction of speech and gesture: Differential activations of metaphoric
698 co-verbal gestures. *Neuropsychologia*, 47(1), 169–179.
699 <https://doi.org/10.1016/j.neuropsychologia.2008.08.009>

700 Kleiner, M., Brainard, D., & Pelli, D. (2007). What's new in Psychtoolbox-3? *Perception* 36
701 *ECVP Abstract Supplement*.

702 Lamminmäki, S., Parkkonen, L., & Hari, R. (2014). Human Neuromagnetic Steady-State
703 Responses to Amplitude-Modulated Tones, Speech, and Music. *Ear and Hearing*, 35(4),
704 461–467.

705 Molholm, S., Ritter, W., Murray, M. M., Javitt, D. C., Schroeder, C. E., & Foxe, J. J. (2002).
706 Multisensory auditory–visual interactions during early sensory processing in humans: a
707 high-density electrical mapping study. *Cognitive Brain Research*, 14(1), 115–128.
708 [https://doi.org/10.1016/S0926-6410\(02\)00066-6](https://doi.org/10.1016/S0926-6410(02)00066-6)

709 Morgan, S. T., Hansen, J. C., Hillyard, S. A., & Posner, M. (1996). Selective attention to
710 stimulus location modulates the steady-state visual evoked potential. *Neurobiology*,
711 93(May), 4770–4774. <https://doi.org/10.1073/pnas.93.10.4770>

712 Müller, M. M., Andersen, S., Trujillo, N. J., Valdés-Sosa, P., Malinowski, P., & Hillyard, S. a.
713 (2006). Feature-selective attention enhances color signals in early visual areas of the
714 human brain. *Proceedings of the National Academy of Sciences of the United States of
715 America*, 103(38), 14250–14254. <https://doi.org/10.1073/pnas.0606668103>

716 Nath, A. R., & Beauchamp, M. S. (2011). Dynamic Changes in Superior Temporal Sulcus
717 Connectivity during Perception of Noisy Audiovisual Speech. *Journal of Neuroscience*,
718 31(5), 1704–1714. <https://doi.org/10.1523/JNEUROSCI.4853-10.2011>

719 Norcia, A. M., Appelbaum, L. G., Ales, J. M., Cottreau, B. R., & Rossion, B. (2015). The
720 steady-state visual evoked potential in vision research: A review. *Journal of Vision*,
721 15(6), 4. <https://doi.org/10.1167/15.6.4>

722 Oostenveld, R., Fries, P., Maris, E., & Schoffelen, J.-M. (2011). FieldTrip: Open source software
723 for advanced analysis of MEG, EEG, and invasive electrophysiological data.
724 *Computational Intelligence and Neuroscience*, 2011, 156869.
725 <https://doi.org/10.1155/2011/156869>

726 Parks, N. A., Hilimire, M. R., & Corballis, P. M. (2011). Steady-state signatures of visual
727 perceptual load, multimodal distractor filtering, and neural competition. *Journal of*
728 *Cognitive Neuroscience*, 23(5), 1113–1124. <https://doi.org/10.1162/jocn.2010.21460>

729 Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming
730 numbers into movies. *Spatial Vision*, 10, 437–442.

731 Picton, T. W., John, M. S., Dimitrijevic, A., Purcell, D., Picton, T. W., John, M. S., Dimitrijevic,
732 A., & Purcell, D. (2003). Human auditory steady-state responses : Respuestas auditivas
733 de estado estable en humanos. *International Journal of Audiology*, 42(4), 177–219.
734 <https://doi.org/10.3109/14992020309101316>

735 Rees, G., Frith, C., & Lavie, N. (2001). Processing of irrelevant visual motion during
736 performance of an auditory attention task. *Neuropsychologia*, 39, 937–949.

737 Regan, M. P., He, P., & Regan, D. (1995). An audio-visual convergence area in the human brain.
738 *Exp Brain Res*, July, 485–487.

739 Regan, M. P., & Regan, D. (1988). A Frequency Domain Technique for Characterizing
740 Nonlinearities in Biological Systems. *J. Theor. Biol.*, 133, 293–317.

741 Regan, M. P., & Regan, D. (1989). Objective Investigation of Visual Function Using a
742 Nondestructive Zoom-FFT Technique for Evoked Potential Analysis. *Canadian Journal*
743 *of Neurological Sciences / Journal Canadien Des Sciences Neurologiques*, 16(2), 168–
744 179. <https://doi.org/10.1017/S0317167100028845>

745 Ross, B., Herdman, A. T., & Pantev, C. (2005). Right hemispheric laterality of human 40 Hz
746 auditory steady-state responses. *Cerebral Cortex*, 15(12), 2029–2039.
747 <https://doi.org/10.1093/cercor/bhi078>

748 Ross, B, Picton, T. W., Herdman, A. T., & Pantev, C. (2004). The effect of attention on the
749 auditory steady-state response. *Neurology & Clinical Neurophysiology*, 2004(3), 22.
750 <https://doi.org/10.1016/j.eclnm.2010.02.002>

751 Ross, B., Draganova, R., Picton, T. W., & Pantev, C. (2003). Frequency specificity of 40-Hz
752 auditory steady-state responses. *Hearing Research*, 186(1), 57–68.
753 [https://doi.org/10.1016/S0378-5955\(03\)00299-5](https://doi.org/10.1016/S0378-5955(03)00299-5)

754 Saupe, K., Widmann, A., Bendixen, A., Müller, M. M., & Schröger, E. (2009). Effects of
755 intermodal attention on the auditory steady-state response and the event-related potential.
756 *Psychophysiology*, 46(2), 321–327. <https://doi.org/10.1111/j.1469-8986.2008.00765.x>

757 Shannon, R., Zeng, F.-G., Kamath, V., Wygonski, J., & Ekelid, M. (1995). Speech Recognition
758 with Primarily Temporal Cues. *Science*, 270(5234), 303–304.

759 Spaak, E., de Lange, F. P., & Jensen, O. (2014). Local Entrainment of Alpha Oscillations by
760 Visual Stimuli Causes Cyclic Modulation of Perception. *Journal of Neuroscience*,
761 34(10), 3536–3544. <https://doi.org/10.1523/JNEUROSCI.4385-13.2014>

762 Stolk, A., Todorovic, A., Schoffelen, J. M., & Oostenveld, R. (2013). Online and offline tools for
763 head movement compensation in MEG. *NeuroImage*, 68, 39–48.
764 <https://doi.org/10.1016/j.neuroimage.2012.11.047>

765 Straube, B., Green, A., Weis, S., & Kircher, T. (2012). A supramodal neural network for speech
766 and gesture semantics: an fMRI study. *PloS One*, 7(11), e51207.
767 <https://doi.org/10.1371/journal.pone.0051207>

768 Talsma, D., Senkowski, D., Soto-Faraco, S., & Woldorff, M. G. (2010). The multifaceted
769 interplay between attention and multisensory integration. *Trends in Cognitive Sciences*,
770 14(9), 400–410. <https://doi.org/10.1016/j.tics.2010.06.008>

771 Tiitinen, H., Sinkkonen, J., Reinikainen, K., Alho, K., Lavaikainen, J., & Naatanen, R. (1993).

772 Selective attention enhanced the auditory 40-Hz transient response in humans. *Nature*,
773 364, 59–60.

774 Vialatte, F.-B., Maurice, M., Dauwels, J., & Cichocki, A. (2010). Steady-state visually evoked
775 potentials: focus on essential paradigms and future perspectives. *Progress in
776 Neurobiology*, 90(4), 418–438. <https://doi.org/10.1016/j.pneurobio.2009.11.005>

777 Werner, S., & Noppeney, U. (2010). Superadditive responses in superior temporal sulcus predict
778 audiovisual benefits in object categorization. *Cerebral Cortex*, 20(8), 1829–1842.
779 <https://doi.org/10.1093/cercor/bhp248>

780 Werner, S., & Noppeney, U. (2011). The contributions of transient and sustained response codes
781 to audiovisual integration. *Cerebral Cortex*, 21(4), 920–931.
782 <https://doi.org/10.1093/cercor/bhq161>

783 Willems, R. M., Özyürek, A., & Hagoort, P. (2007). When language meets action: The neural
784 integration of gesture and speech. *Cerebral Cortex*, 17(10), 2322–2333.
785 <https://doi.org/10.1093/cercor/bhl141>

786 Willems, R. M., Özyürek, A., & Hagoort, P. (2009). Differential roles for left inferior frontal and
787 superior temporal cortex in multimodal integration of action and language. *NeuroImage*,
788 47(4), 1992–2004. <https://doi.org/10.1016/j.neuroimage.2009.05.066>

789 Zemon, V., & Ratliff, F. (1984). Biological Cybernetics Intermodulation Components of the
790 Visual Evoked Potential : *Biological Cybernetics*, 408, 401–408.
791 <https://doi.org/10.1007/BF00335197>

792 Zhao, W., Riggs, X. K., Schindler, X. I., & Holle, X. H. (2018). *Transcranial Magnetic
793 Stimulation over Left Inferior Frontal and Posterior Temporal Cortex Disrupts Gesture-*

794 *Speech Integration*. 38(8), 1891–1900. <https://doi.org/10.1523/JNEUROSCI.1748->

795 17.2017

796 Zhigalov, A., Herring, J. D., Herpers, J., Bergmann, T. O., & Jensen, O. (2019). Probing cortical

797 excitability using rapid frequency tagging. *NeuroImage*, 195, 59–66.

798 <https://doi.org/10.1016/j.neuroimage.2019.03.056>

799