

1 **Age-related decreases of cortical visuo-phonological transformation of unheard**
2 **spectral fine-details**

3
4 Nina Suess*¹, Anne Hauswald¹, Patrick Reisinger¹, Sebastian Rösch², Anne Keitel³ [†]&
5 Nathan Weisz^{1,4} [†]

6
7 ¹Centre for Cognitive Neuroscience and Department of Psychology,
8 University of Salzburg, Austria

9 ²Department of Otorhinolaryngology, Head and Neck Surgery, Paracelsus Medical University
10 Salzburg, University Hospital Salzburg, Salzburg, Austria

11 ³School of Social Sciences, University of Dundee

12 ⁴Neuroscience Institute, Christian Doppler University Hospital, Paracelsus Medical
13 University, Salzburg, Austria

14
15
16
17 Address for correspondence:

18 Nina Suess

19 University of Salzburg

20 Centre for Cognitive Neuroscience

21 Hellbrunner Straße 34

22 A-5020 Salzburg

23 Austria/Europe

24 Tel.: 0043 662 8044 5161

25 E-Mail: nina.suess@sbq.ac.at

26
27 [†]shared senior authors

28
29 *Preprint server:* Biorxiv, <https://doi.org/10.1101/2021.04.13.439628>

30
31 *Classification:* Biological sciences, Neuroscience

32
33 *Keywords:* MEG, visual speech processing, low-frequency speech tracking, multisensory
34 processing

35 **Abstract**

36 The integration of visual and auditory cues is crucial for successful processing of speech,
37 especially under adverse conditions. Recent reports have shown that when participants watch
38 muted videos of speakers, the phonological information about the acoustic speech envelope
39 is tracked by the visual cortex. However, the speech signal also carries much richer acoustic
40 details, e.g. about the fundamental frequency and the resonant frequencies, whose visuo-
41 phonological transformation could aid speech processing. Here, we investigated the neural
42 basis of the visuo-phonological transformation processes of these more fine-grained acoustic
43 details and assessed how they change with ageing. We recorded whole-head
44 magnetoencephalography (MEG) data while participants watched silent intelligible and
45 unintelligible videos of a speaker. We found that the visual cortex is able to track the unheard
46 intelligible modulations of resonant frequencies and the pitch linked to lip movements.
47 Importantly, only the processing of intelligible unheard formants decreases significantly with
48 age in the visual and also in the cingulate cortex. This is not the case for the processing of the
49 unheard speech envelope, the fundamental frequency or the purely visual information carried
50 by lip movements. These results show that unheard spectral fine-details (along with the
51 unheard acoustic envelope) are transformed from a mere visual to a phonological
52 representation. Aging affects especially the ability to derive spectral dynamics at formant
53 frequencies. Since listening in noisy environments should capitalize on the ability to track
54 spectral fine-details, our results provide a novel focus on compensatory processes in such
55 challenging situations.

56 **Significance statement**

57 The multisensory integration of speech cues from visual and auditory modalities is crucial for
58 optimal speech perception in noisy environments or for elderly individuals with progressive
59 hearing loss. It has already been shown that the visual cortex is able to extract global acoustic
60 information like amplitude modulations from silent visual speech, but whether this extends to
61 fine-detailed spectral acoustic information remains unclear. Here, we demonstrate that the
62 visual cortex is indeed able to extract fine-detailed phonological cues just from watching silent
63 lip movements. Furthermore, this tracking of acoustic fine-details is deteriorating with age.
64 These results suggest that the human brain is able to transform visual information into useful
65 phonological information, and this process might be crucially affected in ageing individuals.

66 **1 Introduction**

67 Speech understanding is a multisensory process that requires diverse modalities to work
68 together for an optimal experience. Congruent audiovisual input is especially crucial for
69 understanding speech in noise (Crosse et al., 2016; Sumbly & Pollack, 1954), highlighting the
70 importance of visual cues in speech processing studies. One hypothesis is that activation from
71 visual speech directly modulates activation in auditory cortex, although the results have been
72 mixed and a lot of questions remain unanswered (Bernstein & Liebenthal, 2014; Keitel et al.,
73 2020). One important question regards the nature of the representation in the visual cortex,
74 and whether it is strictly visual or already tracks acoustic information that is associated with
75 the visual input (for non-speech stimuli see e.g. Escoffier et al., 2015). A first approach to
76 address this showed that occipital activation elicited by silent lip reading also reflects dynamics
77 of the acoustic envelope (O'Sullivan et al., 2017). Further evidence that the visual cortex is
78 able to track certain aspects of speech by visual cues alone comes from a recent study by
79 Hauswald et al. (2018). Evidently, it has been shown that visual speech contributes
80 substantially to audiovisual speech processing in the sense that the visual cortex is able to
81 extract phonological information from silent lip movements in the theta-band (4-7 Hz).
82 Crucially, this tracking is dependent on the intelligibility of the silent speech, with absent
83 tracking when the silent visual speech is unintelligible. Another study supports the former
84 findings and extends the present framework by providing evidence that the visual cortex
85 passes information to the angular gyrus, which extracts slow features (below 1 Hz) from lip
86 movements, which are then mapped onto auditory features and passed on to auditory cortices
87 for better speech comprehension (Bourguignon et al., 2020). These findings underline the
88 importance of slow frequency properties of visual speech for enhanced speech
89 comprehension from both the delta (0.5-3 Hz) and theta-band (4-7 Hz), especially due to
90 frequencies between 1-7 Hz being crucial for comprehension (Giraud & Poeppel, 2012).
91 Moreover, the spectral profile of lip movements is also settled within this range (Park et al.,
92 2016).

93 Recent behavioural evidence describes that spectral fine details can also be extracted by
94 observation of lip movements (Plass et al., 2020). This raises the interesting question whether
95 this information is also represented at the level of the visual cortex, analogous to the envelope
96 as shown previously (Hauswald et al., 2018). Particularly relevant spectral fine details are
97 formant peaks around 2500 Hz, which are indicated to be modulated in the front cavity (Badin
98 et al., 1990). This corresponds to expansion and contraction of the lips (Plass et al., 2020),
99 thus having a relationship with certain lip movements and could therefore be extracted for
100 important phonological cues.

101 Furthermore, not only resonant frequencies, but also the fundamental frequency (or pitch
102 contour) plays an important role in speech understanding in noisy environments (Hopkins et

103 al., 2008), and could potentially be extracted from silent lip movements. Whether the visual
104 cortex is able to track formant and pitch information in (silent) visual speech, has not been
105 investigated to date.

106 Knowledge on how the brain is processing speech is also vital when it comes to ageing,
107 potentially with regards to age-related hearing loss (Lieberman, 2017). Several studies have
108 investigated the influence of age on speech comprehension, with results that signify ageing
109 is, in most cases, accompanied by listening difficulties, especially in noise (Tun & Wingfield,
110 1999; Wong et al., 2009). Furthermore, while the auditory tracking of a speech-paced
111 stimulation (~ 3 Hz) is less consistent in older adults compared to younger adults, alpha
112 oscillations are enhanced in younger adults during attentive listening, suggesting declined top-
113 down inhibitory processes that support selective suppression of irrelevant information (Henry
114 et al., 2017). Older adults also indicate a compensatory mechanism when processing
115 degraded speech especially in anterior cingulate cortex (ACC) and middle frontal gyrus (Erb
116 & Obleser, 2013). Additionally, the temporal processing of auditory information is altered in
117 the ageing brain, pointing to decreased selectivity for temporal modulations in primary auditory
118 areas (Erb et al., 2020). Those studies reinforce a distinctive age-related alteration in
119 processing auditory speech. This raises the question whether we also see an impact of age
120 on audiovisual speech processing, an issue that has not been addressed so far.

121 Combining the important topics mentioned above, this study aims to answer two critical
122 questions regarding audiovisual speech processing: First, we ask if the postulated visuo-
123 phonological transformation process in visual cortex mainly represents global energy
124 modulations (i.e. speech envelope) or if it also entails spectral fine details (like formant or pitch
125 curves). Second, we question if visuo-phonological transformation is subject to age-related
126 decline. To the best of our knowledge, this study presents first neurophysiological evidence
127 that the visual cortex is not only able to extract the unheard speech envelope, but also unheard
128 formant and pitch information from lip movements. Crucially, we observed an age-related
129 decline that mainly affects tracking of the formants (and to some extent the envelope and the
130 fundamental frequency). Interestingly, we observed different tracking properties for different
131 brain regions and frequencies: While tracking intelligible formants declines reliably in occipital
132 and cingulate cortex for both delta and theta, we observed a decline of theta-tracking just in
133 occipital cortex, suggesting different age-related effects in different brain regions. Our results
134 suggest that the ageing brain deteriorates in deriving spectral fine-details linked to the visual
135 input, a process that could contribute to perceptual difficulties in challenging listening
136 situations.

137 **2 Materials and methods**

138

139 *2.1 Participants*

140 We recruited 50 participants (28 females; 2 left-handed; mean age: 37.96 years; SD: 13.33
141 years, range: 19-63 years) for the experiment. All participants had normal or corrected-to-
142 normal eyesight, self-reported normal hearing and no neurological disorders. All participants
143 received either a reimbursement of €10 per hour or course credits for their participation. All
144 participants signed an informed consent form. The experimental procedure was approved by
145 the Ethics Committee of the University of Salzburg.

146

147 *2.2 Stimuli*

148 Videos were recorded with a digital camera (Sony NEX FS100) at a rate of 50 frames per
149 second, the corresponding audio files were recorded at a sampling rate of 48 kHz. The videos
150 were spoken by two female native German speakers. The stimuli were taken from the book
151 “Das Wunder von Bern” (“The Miracle of Bern”; [https://www.aktion-](https://www.aktion-mensch.de/inklusion/bildung/bestellservice/materialsuche/detail?id=62)
152 [mensch.de/inklusion/bildung/bestellservice/materialsuche/detail?id=62](https://www.aktion-mensch.de/inklusion/bildung/bestellservice/materialsuche/detail?id=62)) which was provided
153 in an easy language. The easy language does not include any foreign words, has a coherent
154 verbal structure and is facile to understand. We used simple language to avoid that limited
155 linguistic knowledge is interfering with possible lip reading abilities. 24 pieces of text were
156 chosen from the book and recorded from each speaker, lasting between 33 and 62 seconds,
157 thus resulting in 24 videos. Additionally, all videos were reversed, which resulted in 24 forward
158 videos and 24 corresponding backward videos. Forward and backward audio files were
159 extracted from the videos and used for the data analysis. Half of the videos were randomly
160 selected to be presented forward and the remaining half to be presented backward. The videos
161 were back-projected on a translucent screen in the centre of the screen by a Propixx DLP
162 projector (VPixx technologies, Canada) with a refresh rate of 120 Hz per second and a screen
163 resolution of 1920 x 1080 pixels. The translucent screen was placed ~110 cm in front of the
164 participant and had a screen diagonal of 74 cm. One speaker was randomly chosen per
165 subject and kept throughout the experiment, so each participant only saw one speaker.

166

167 *2.3 Procedure*

168 Participants were first instructed to take part in an online study, in which their behavioural lip
169 reading abilities were tested, and in which they were asked about their subjective hearing
170 impairment. This German lip reading test is available as SaLT (Salzburg Lipreading Test)
171 (Suess et al., 2021). Participants were presented with silent videos of numbers, words and
172 sentences and could watch every video twice. They then had to write down the words they
173 thought they had understood from the lip movements. This online test lasted approximately 40

174 minutes and could be conducted at home or right before the experiment in the MEG-lab. After
175 completing the behavioural experiment, the MEG experiment started. Participants were
176 instructed to pay attention to the lip movements of the speakers and passively watch the mute
177 videos. They were presented with 6 blocks of videos, and in each block, 2 forward and 2
178 backward videos were presented in a random order. The experiment lasted about an hour
179 including preparation. The experimental procedure was programmed in Matlab with the
180 Psychtoolbox-3 (Brainard, 1997) and an additional class-based abstraction layer
181 (https://gitlab.com/thht/o_ptb) programmed on top of the Psychtoolbox (Hartmann & Weisz,
182 2020).

183

184 *2.4 Data acquisition*

185 Brain activity was measured using a 306-channel whole head MEG system with 204 planar
186 gradiometers and 102 magnetometers (Neuromag TRIUX, Elekta), a sampling rate of 1000
187 Hz and an online highpass-filter of 0.1 Hz. Before entering the magnetically shielded room
188 (AK3B, Vakuumschmelze, Hanau, Germany), the head shape of each participant was
189 acquired using approximately 500 digitized points on the scalp, including fiducials (nasion, left
190 and right pre-auricular points) with a Polhemus Fastrak system (Polhemus, Vermont, USA).
191 The head position of each individual participant relative to the MEG sensors was controlled
192 once before each experimental block. Vertical and horizontal eye movements and
193 electrocardiographic data was also recorded, but not used for preprocessing. The continuous
194 MEG data was then preprocessed off-line with the signal space separation method from the
195 Maxfilter software (Elekta Oy, Helsinki, Finland) to correct for different head positions across
196 blocks and to suppress external interference (Taulu et al., 2005).

197

198 *2.5. Data analysis*

199

200 *2.5.1 Preprocessing*

201 Acquired datasets were analysed using the Fieldtrip toolbox (Oostenveld et al., 2011). The
202 maxfiltered MEG data were highpass-filtered at 1 Hz using a finite impulse response (FIR)
203 filter (Kaiser window, order 440). For extracting physiological artefacts from the data, 60
204 principal components were calculated. Via visual inspection, the components displaying eye
205 movements, heartbeat and external power noise from the nearby train tracks (16.67 Hz) were
206 removed from the data. We removed on average 2.24 components per participant ($SD = 0.65$).
207 The data were then lowpass-filtered at 30 Hz and corrected for the delay between the stimulus
208 computer and the screen inside the chamber (9 ms for each video). We then resampled the
209 data to 150 Hz and segmented them in 2-second trials to increase the signal-to-noise ratio.

210

211 2.5.2 Source projection of MEG data

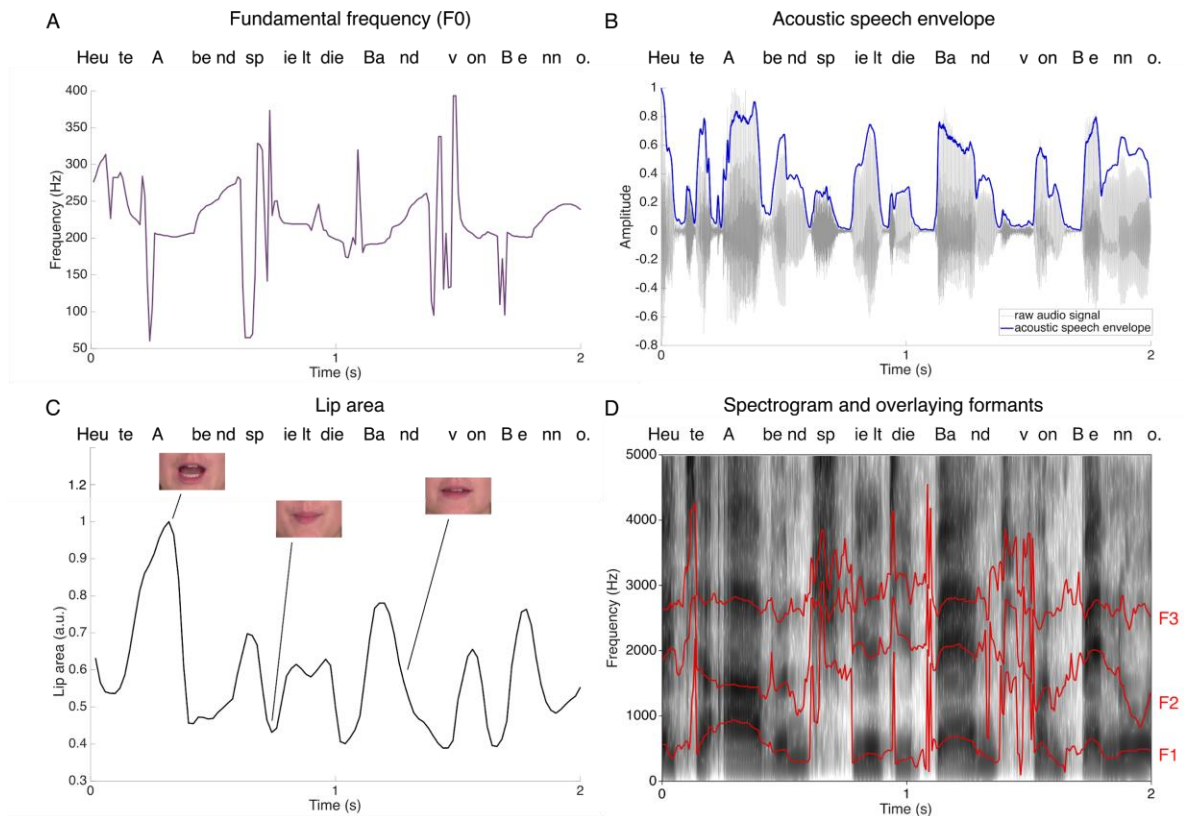
212 We used either a standard structural brain from the Montreal Neurological Institute (MNI,
213 Montreal, Canada) or, where possible, the individual structural MRI (20 participants) and
214 warped it to match the individual's fiducials and head shape as accurately as possible (Mattout
215 et al., 2007). A 3-D grid with 1-cm resolution and 2982 voxels based on an MNI template brain
216 was morphed into the brain volume of each participant. This allows group-level averaging and
217 statistical analysis as all the grid points in the warped grid belong to the same brain region
218 across subjects. These aligned brain volumes were also used for computing single-shell head
219 models and leadfields (Nolte, 2003). By using the leadfields and the common covariance
220 matrix (pooling data from all blocks), a common LCMV beamformer spatial filter was computed
221 (Veen et al., 1997).

222

223 2.5.3 Extraction of lip area, acoustic speech envelope, formants and pitch

224 The lip area of the visual speech was extracted using a MATLAB script adapted from Park et
225 al. (2016). This data was then upsampled to 150 Hz to match the downsampled preprocessed
226 MEG signal. The acoustic speech envelope was extracted with the Chimera toolbox from the
227 audio files corresponding to the videos which constructs nine frequency bands in the range of
228 100-10000 Hz as equidistant on the cochlear map (Smith et al., 2002). Then the sound stimuli
229 were band-pass filtered in these bands with a 4th-order Butterworth filter to avoid edge
230 artefacts. For each of the frequency bands, the envelopes were calculated as absolute values
231 of the Hilbert transform and then averaged to get the full-band envelope for coherence analysis
232 (Gross et al., 2013; Keitel et al., 2017). This envelope was then downsampled to 150 Hz to
233 match the preprocessed MEG signal. The resonant frequencies (or formants) were extracted
234 using the Burg method implemented in Praat 6.0.48 (Boersma & Weenink, 2019). Up to 5
235 formants were extracted from each audio file to make sure that the relevant formants were
236 extracted thoroughly. For analysis purposes, just F2 and F3 were averaged and used. Those
237 two formants fluctuate around 2500 Hz and tend to merge into a single peak when pronouncing
238 certain consonant-vowel combinations (Badin et al., 1990). The mentioned merging process
239 is taking place in the front region of the oral cavity and can therefore also be seen by observing
240 lip movements (Plass et al., 2020). The formants were extracted at a rate of 200 Hz for the
241 sake of simplicity and then downsampled to 150 Hz. The pitch (or fundamental frequency, f_0)
242 was extracted using the Matlab Audio Toolbox function *pitch.m* with default options (extraction
243 between 50 and 400 Hz) at a rate of 100 Hz and then upsampled to 150 Hz.

244



245

246 *Figure 1: Example time series for a 2 second forward section of all the parameters used for*
 247 *coherence calculation. A) Example time series of the fundamental frequency extracted with*
 248 *the pitch.m MATLAB function. B) Example audio signal and the acoustic speech envelope (in*
 249 *blue). C) Lip area extracted from the video frames with the MATLAB script adapted from Park*
 250 *et al. (2016). D) Example spectrogram with overlaying formants (F1-F3, red lines) extracted*
 251 *with Praat.*

252

253 2.5.4 Coherence calculation

254 We calculated the cross-spectral density between the lip area, the unheard acoustic speech
 255 envelope, the averaged F2 and F3 formants and the pitch and every virtual sensor with a multi-
 256 taper frequency transformation (1-25 Hz in 0.5 Hz steps, 3 Hz smoothing). Then we calculated
 257 the coherence between the activity of every virtual sensor and the lip area, the acoustic speech
 258 envelope, the averaged formant curve of F2 and F3 and the pitch curve, which we will refer to
 259 as lip-brain coherence, envelope-brain coherence, formant-brain coherence and pitch-brain
 260 coherence, respectively, in the manuscript.

261

262 2.6 Statistical analysis

263 To test for differences in source space in occipital cortex for forward and backward coherence
 264 values, we extracted all voxels labeled as “occipital cortex” in the Automated Anatomical
 265 Labeling (AAL) atlas (Tzourio-Mazoyer et al., 2002) for a predefined region-of-interest analysis

266 (Hauswald et al., 2018). We then contrasted forward and backward conditions using two-tailed
267 dependent-samples t -tests on the averaged coherence values for the frequency bands of
268 interest (1-7 Hz). This was done separately for the lip-brain coherence, the envelope-brain
269 coherence, the formant-brain coherence and the pitch-brain coherence. In a first step, we
270 decided to average over the delta (1-3 Hz) and theta (4-7 Hz) frequency bands since they
271 carry important information in general on speech processing (phrasal and syllabic processing,
272 respectively) (Giraud & Poeppel, 2012). Moreover, previous studies investigated lip movement
273 related activity either in the delta-band (Bourguignon et al., 2020; Park et al., 2016) or the
274 theta-band (Hauswald et al., 2018), leading us to also do a follow-up analysis separately for
275 the different frequency bands (described later in this section).

276 To generate a normalized contrast between processing of forward (intelligible) and backward
277 (unintelligible) lip movements, we subtracted the backward coherence values from the forward
278 coherence values for our respective measures (lip-brain coherence, unheard speech
279 envelope-brain coherence, unheard formant-brain coherence and unheard pitch-brain
280 coherence). From now on, we refer to this normalized contrast as “Intelligibility index”, which
281 quantifies the differences in coherence between intelligible and unintelligible visual speech.

282 For testing the relationship between the four different intelligibility indices (lip-brain, envelope-
283 brain, formant-brain and pitch-brain) and age, we conducted a voxelwise correlation with age.
284 To control for multiple comparisons, we used a non-parametric cluster-based permutation test
285 (Maris & Oostenveld, 2007). Here, clusters of correlation coefficients being significantly
286 different from zero (showing p -values < 0.05) were identified and their respective t -values were
287 extracted and summed up to get a cluster-level test statistic. Random permutations of the data
288 were then drawn by reordering the behavioural data (in our case age) across participants.
289 After each permutation, the maximum cluster level t -value was recorded, generating a
290 reference distribution of cluster-level t -values (using a Monte Carlo procedure with 1000
291 permutations). Cluster p -values were estimated as the proportion of cluster t -values in the
292 reference distribution exceeding the cluster t -values observed in the actual data. Significant
293 voxels (which were only found in the correlation between the formant-brain index and age)
294 were then extracted and averaged for data-driven ROIs (occipital cortex and cingulate cortex)
295 which were defined using the Automated Anatomical Labeling (AAL) atlas (Tzourio-Mazoyer
296 et al., 2002). These data-driven ROIs were then applied to all intelligibility indices to make the
297 ROI analysis comparable. We then fitted four linear models using the function lm from the
298 stats package in R to investigate if age could predict the change in the calculated intelligibility
299 indices and to visualize the statistical effects of the whole brain analysis. To further clarify the
300 relationship between age and the processing of intelligible and unintelligible lip movements
301 and to unravel the dynamics in our whole brain correlation analysis, we split our participants
302 into two groups by the median (young: people < 37 , $N=25$, older: people > 37 , $N=25$). We then

303 calculated a repeated-measures ANOVA with 2 conditions: age (young vs. older) and
304 intelligibility (forward vs. backward visual speech) for our data-driven ROIs separately
305 (occipital cortex and cingulate cortex) using the *stats* package in R. To further investigate the
306 effects between age and intelligibility, we conducted post-hoc tests with Bonferroni correction
307 using the function *PostHocTest*. The last step consisted of a follow-up analysis where we
308 decided to separate the averaged frequency-bands (delta and theta) again to unravel possible
309 differences of our effect dependent on the frequency-band. We again conducted a voxelwise
310 correlation with age separately for the delta-band (1-3 Hz) and for the theta-band (4-7 Hz) with
311 the already described non-parametric cluster-based permutation test for all described
312 intelligibility indices. Finally, we extracted the values from the voxel with the lowest *t*-value (for
313 the delta and theta-band, respectively) and fitted a linear model again to investigate if age
314 could predict the change in the intelligibility indices and to visualize the statistical effects of the
315 whole brain analysis.

316 **3 Results**

317

318 **3.1 Behavioural results**

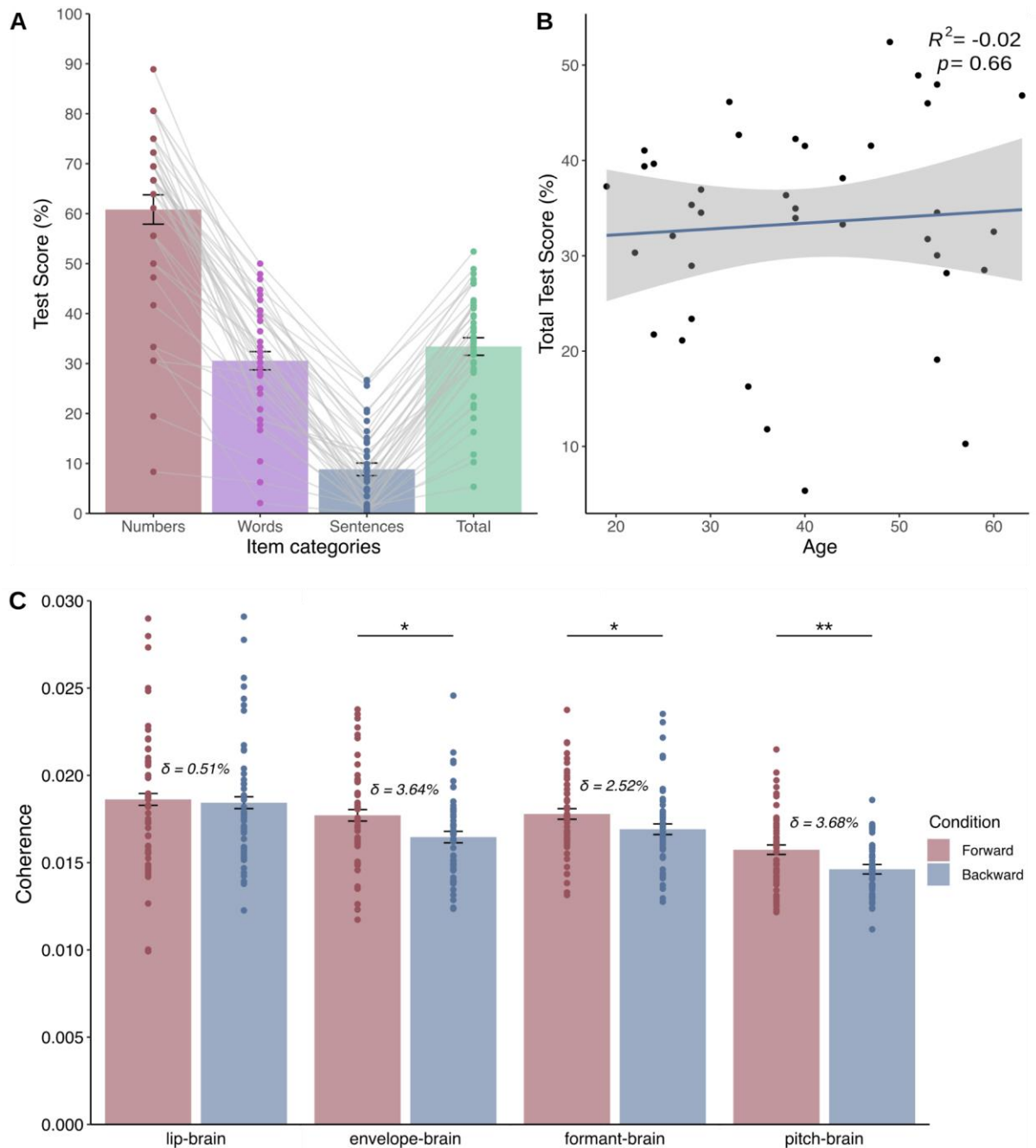
319 We investigated participants' lip reading abilities in a separate experiment that was conducted
320 before the MEG session. They were presented with silent videos of spoken numbers, words,
321 and sentences, and the task was to write down what they had understood just from the lip
322 movements alone. A detailed description of the behavioural task will be published in a
323 separate paper (Suess et al., 2021). 43 of the 50 participants completed the behavioural
324 experiment. 4 people had to be excluded because there were problems with the data
325 acquisition and their answers were not saved. While the recognition rate for the numbers were
326 high ($M = 60.83\%$, $SE = 2.93\%$), lip reading abilities for complex stimuli (words and sentences)
327 were low in general (words: $M = 30.57\%$, $SE = 1.82\%$; sentences: $M = 8.83\%$, $SE = 1.26\%$).
328 Participants had an average total score of 33.41% ($SE = 1.75\%$, Figure 2A). Investigating if
329 age could predict the total test score revealed that those two variables were uncorrelated ($F(1,$
330 $37) = .191$, $p = .664$, $R^2 = -0.021$), Figure 2B), showing that in our sample, behavioural lip
331 reading abilities are not changing with age. This is consistent with our study on general lip
332 reading abilities in the German language (Suess et al., 2021), but different to other studies
333 indicating higher lip reading abilities in younger individuals (Feld & Sommers, 2009; Tye-
334 Murray et al., 2007b). Participants also completed a questionnaire on subjective hearing
335 impairment (APHAB, Löhler et al., (2014)). Further investigating the relationship between
336 subjective hearing impairment and test score also revealed no significant effect ($F(1, 37) =$
337 $.104$, $p = .75$, $R^2 = -0.024$) in the current sample. This is in line with studies investigating
338 hearing impairment in older adults (Tye-Murray et al., 2007a), but not supporting our own
339 results which show a relationship between self-reported hearing impairment and lip reading
340 abilities (Suess et al., 2021). However, as the current study was aiming to test normal hearing
341 individuals with restricted variance in hearing impairment, those results cannot be compared
342 directly to Suess et al. (2021), which also included individuals with severe hearing loss as well
343 as prelingually deaf individuals.

344

345 **3.2 Visuo-phonological transformation is carried by both tracking of global envelope** 346 **and spectral fine-details during presentation of intelligible silent lip movements**

347 We calculated the coherence between the MEG data and the lip envelope, the unheard
348 acoustic speech envelope, the unheard resonant frequencies and the unheard pitch (from now
349 on called lip-brain coherence, envelope-brain coherence, formant-brain coherence, and pitch-
350 brain coherence, respectively). As the visuo-phonological transformation process is likely
351 taking place in visual areas (Hauswald et al., 2018), we defined the occipital cortex using the
352 AAL atlas (Tzourio-Mazoyer et al., 2002) as a predefined region-of-interest and averaged over

353 all voxels from this ROI. We then compared the mean for the coherence of the presented
354 forward videos (intelligible lip movements) with the mean of the presented backward videos
355 (unintelligible lip movements) separately for the lip-brain coherence, the envelope-brain
356 coherence, the formant-brain coherence and the pitch-brain coherence. While there was no
357 significant difference in lip-brain coherence for intelligible and unintelligible visual speech ($t(49)$
358 = 0.396, $p = 0.694$, $d = 0.056$), we found a significant difference in unheard envelope-brain
359 coherence for intelligible and unintelligible visual speech ($t(49) = 2.679$, $p = 0.01$, $d = 0.379$).
360 Most importantly, we found a significant difference also for the unheard formant-brain
361 coherence ($t(49) = 2.039$, $p = 0.047$, $d = 0.288$) and for the unheard pitch-brain coherence for
362 intelligible and unintelligible visual speech ($t(49) = 2.91$, $p = 0.005$, $d = 0.411$, all in Figure 2C).
363 The results on the tracking of lip movements are in line with former findings, showing that the
364 visual cortex tracks these regardless of intelligibility, but point to different tracking properties
365 dependent on the intelligibility of the unheard speech envelope. Interestingly, we show here
366 that the visual cortex is also able to distinguish between unheard intelligible and unintelligible
367 formants (or resonant frequencies) and pitch (or F0) modulations extracted from the
368 spectrogram, showing that also unheard intelligible spectral details are extracted from visual
369 speech and represented at the level of the visual cortex.



370

371

372 *Figure 2: Behavioural data and comparison of information tracking in visual cortex. A)*
373 *Behavioural lip reading abilities. Participants recognized numbers the most, followed by words*
374 *and sentences. B) Correlation between age and total test score revealed no significant*
375 *correlation ($p = 0.66$), suggesting that lip reading abilities do not change with age. Blue line*
376 *depicts regression line, shaded areas depict standard error of mean (SE). C) Mean values*
377 *extracted from all voxels in occipital cortex showing no significant differences in lip-brain*
378 *coherence ($p = 0.694$), but showing significant differences in unheard envelope-brain*
379 *coherence ($p = 0.01$), formant-brain coherence ($p = 0.047$) and unheard pitch-brain coherence*
380 *($p = 0.005$) between forward and backward presentation of visual speech. Error bars represent*

381 1 standard error of mean for within-subject designs (O'Brien & Cousineau, 2014), δ indicates
382 the relative change between forward and backward conditions in percent.

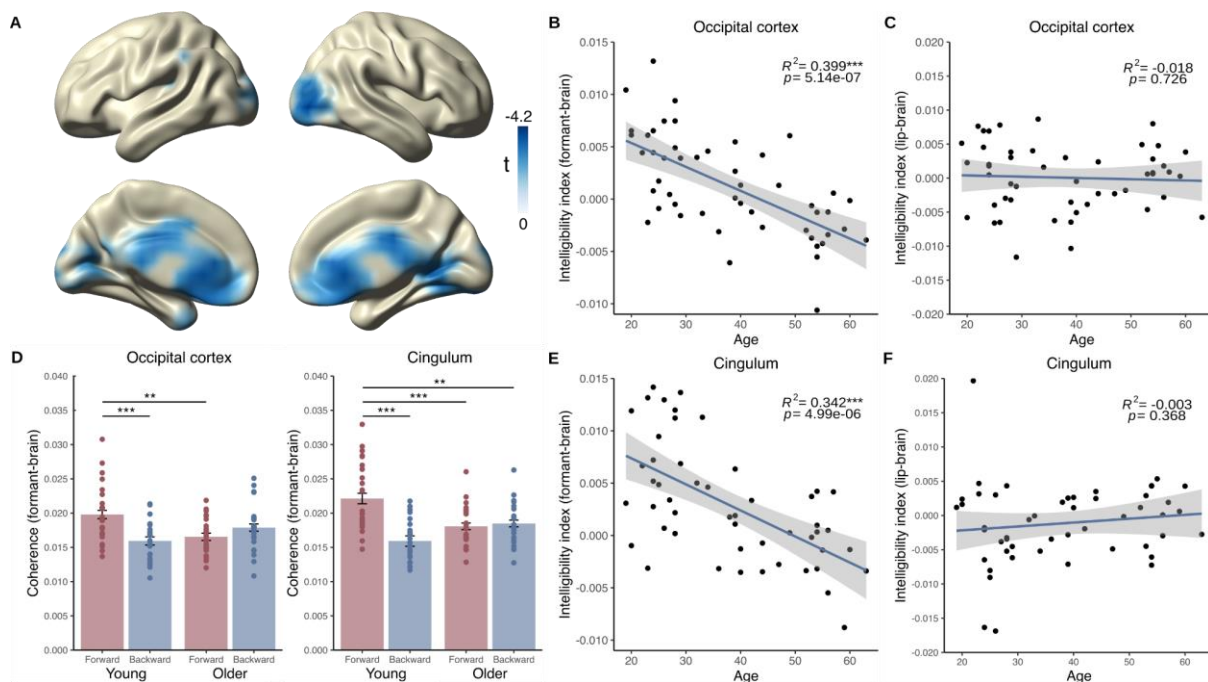
383

384 **3.3 Spectral fine-detail tracking rather than global envelope tracking is altered in the** 385 **ageing population**

386 We were then interested in how the visuo-phonological transformation process is influenced
387 by age. So we calculated a voxelwise correlation between the intelligibility index (difference
388 between coherence for forward videos and coherence for backward videos) separately for our
389 coherence indices (lip-brain, envelope-brain, formant-brain and pitch-brain) and the age of the
390 participants. We neither found a significant correlation between the intelligibility index of the
391 lip-brain coherence and age ($p = 1$, cluster-corrected) nor between the intelligibility index of
392 the unheard envelope-brain coherence and age ($p = 0.09$, cluster-corrected). Also, the
393 correlation between the intelligibility index of the unheard pitch-brain coherence was
394 statistically not significant ($p = 0.07$, cluster-corrected). However, the overall trend for the
395 envelope-brain and the pitch-brain coherence was to decline with age. Interestingly, we did
396 find a significant negative correlation between the intelligibility index of the unheard formant-
397 brain coherence and age ($p = 0.002$, cluster-corrected), strongest in occipital cortex and
398 cingulate cortex (lowest t -value: -4.124 , MNI [40 -90 0], Figure 3A). To further investigate the
399 effects, we extracted the voxels showing a statistical effect in our whole brain analysis (Figure
400 3A) and divided them into occipital voxels and voxels from the cingulate cortex using the AAL
401 atlas (Tzourio-Mazoyer et al., 2002).

402 To investigate how strong the relationship between age and the different intelligibility indices
403 is in our ROIs, we fitted four separate linear models. We started with the lip-brain index to
404 exclude the possibility that our effect is due to visual processing. We found that age could not
405 predict the lip-brain intelligibility index in any of the chosen ROIs (occipital cortex: $F(1, 48) =$
406 0.124 , $p = 0.727$, $\eta^2 = 0.002$, Figure 3C; cingulate cortex: $F(1, 48) = 0.825$, $p = 0.368$, $\eta^2 =$
407 0.017 , Figure 3F). On the contrary, we found that age could significantly predict the decrease
408 in the formant-brain intelligibility index in both occipital areas ($F(1, 48) = 33.59$, $p = 5.14e-07$,
409 $\eta^2 = 0.412$, Figure 3B) and cingulate cortex ($F(1, 48) = 26.42$, $p = 4.99e-06$, $\eta^2 = 0.355$, Figure
410 3E), suggesting an altered tracking process for the formants in ageing. Further fitting linear
411 models to investigate the effects in our ROIs for the envelope-brain coherence and the pitch-
412 brain coherence revealed that age could not significantly predict the envelope-brain index in
413 occipital ($F(1, 48) = 1.638$, $p = 0.207$, $\eta^2 = 0.033$) or cingulate cortex ($F(1, 48) = 0.681$, $p =$
414 0.413 , $\eta^2 = 0.014$) and also not the pitch-brain index in occipital cortex ($F(1, 48) = 2.584$, $p =$
415 0.114 , $\eta^2 = 0.051$). However, age could significantly predict the pitch-brain index in cingulate
416 cortex ($F(1, 48) = 6.972$, $p = 0.011$, $\eta^2 = 0.127$). The lack of tracking differences between
417 intelligible and unintelligible lip movements suggests that the visual cortex processes basic

418 visual properties of lip movements, but that there are differential processing strategies for
 419 acoustic information associated with these lip movements. These results also suggest that
 420 processing of the pitch (or fundamental frequency) is altered to some extent in the ageing
 421 population, at least in cingulate cortex. In summary, the correlation between the envelope-
 422 brain index and age and the pitch-brain index and age seem to show a tendency in line with
 423 the relationship between the formant-brain index and age in the whole brain analysis. We see
 424 that effect sizes are biggest for the formant-brain index (occipital $\eta^2 = 0.412$, cingulate $\eta^2 =$
 425 0.355), followed by the pitch-brain index (occipital $\eta^2 = 0.051$, cingulate $\eta^2 = 0.127$). Lower
 426 effect sizes are found for the envelope-brain index (occipital $\eta^2 = 0.033$, cingulate $\eta^2 = 0.014$)
 427 and the lip-brain index (occipital $\eta^2 = 0.002$, cingulate $\eta^2 = 0.017$) after extracting voxels from
 428 the data-driven ROI, adding to the evidence of a differential processing of speech properties
 429 in age.
 430



431
 432 *Figure 3: Correlation between age and intelligibility index (i.e. difference in forward vs.*
 433 *backward tracking) and comparison of age-groups. A) Statistical values of the voxelwise*
 434 *correlation of the intelligibility index (forward formant-brain coherence - backward formant-*
 435 *brain coherence) with age (averaged over 1-7 Hz, $p < 0.05$, cluster-corrected) showing a*
 436 *strong decrease of intelligibility tracking in occipital regions and in cingulate cortex. B)*
 437 *Correlation of formant-brain intelligibility index in significant occipital voxels extracted from A*
 438 *showing a significant correlation with age ($p = 5.14e-07$). C) Correlation of lip-brain intelligibility*
 439 *index in significant occipital voxels extracted from A showing a not significant correlation with*
 440 *age ($p = 0.726$). D) Formant-brain coherence separated for age and for forward and backward*
 441 *presented visual speech for different ROIs. Coherence values from occipital cortex indicating*

442 *significant differences between forward and backward tracking in the young group ($p =$*
443 *0.0004), but not in the older group ($p = 0.467$), and also a difference between forward tracking*
444 *in the young group and forward tracking in the older group ($p = 0.004$). Coherence values from*
445 *cingulum indicating significant differences between forward and backward tracking in the*
446 *young group ($p = 1.1e-07$), but not in the older group ($p = 1.000$), and also a difference*
447 *between forward tracking in the young group and forward tracking in the older group ($p =$*
448 *0.0005). Additional significant effects were observed between the forward tracking in the*
449 *young group and the backward tracking in the older group ($p = 0.002$). E) Correlation of*
450 *formant-brain intelligibility index in significant voxels from cingulate cortex extracted from A*
451 *showing a significant correlation with age ($p = 4.99e-06$). F) Correlation of lip-brain intelligibility*
452 *index in significant voxels from cingulate cortex extracted from A showing a not significant*
453 *correlation with age ($p = 0.368$). Blue lines depict regression lines, shaded areas depict*
454 *standard error of mean (SE).*

455

456 **3.4 Intelligibility effects are mainly carried by young individuals**

457 To unravel the effects explained in sections 3.3, we reassessed the coherence values
458 separately for forward and backward speech with respect to the age of our participants. Thus,
459 we decided to split our sample into two age groups (younger vs. older) and calculated a 2x2
460 ANOVA on the averaged voxels that we extracted from figure 3A for the former calculated
461 formant-brain coherence (for forward and backward coherence, respectively). We again
462 separated them into two ROIs (occipital cortex and cingulate cortex) and calculated for each
463 an ANOVA with the factors age (young vs. older) and intelligibility (forward formant-brain
464 coherence vs. backward formant-brain coherence). We did not find a main effect of age in
465 occipital cortex ($F(1, 49) = 0.981, p = 0.324$), but a main effect close to significance threshold
466 of intelligibility ($F(1, 49) = 3.627, p = 0.059$). We also found a distinct interaction effect between
467 age and intelligibility ($F(1, 49) = 15.723, p = 0.0001$, Figure 3D, occipital cortex). To further
468 investigate the interaction effect, we calculated a post-hoc test with Bonferroni correction,
469 which revealed a significant difference between the forward and backward conditions in the
470 young group ($p = 0.0004$), but not in the older group ($p = 0.467$). Furthermore, we discovered
471 a significant difference between the forward condition in the young group and the forward
472 condition in the older group ($p = 0.004$), exhibiting that the young group is able to track the
473 forward speech stronger than the older group. In cingulate cortex, we also did not find a main
474 effect of age ($F(1, 49) = 1.399, p = 0.239$), but here we found a main effect of intelligibility ($F(1,$
475 $49) = 16.474, p = 0.0001$). We also found a distinct interaction effect between age and
476 intelligibility ($F(1, 49) = 21.536, p = 1.1e-05$, Figure 4D, cingulum). The Bonferroni corrected
477 post-hoc test also revealed a significant difference between the forward and backward
478 conditions in the young group ($p = 1.1e-07$), but not in the older group ($p = 1.000$). Additionally,

479 we found a significant difference between the forward condition in the young group and the
480 forward condition in the older group ($p = 0.0005$), strengthening our observation that the young
481 group is able to distinguish more faithfully between forward and backward speech than the
482 older group. In the cingulate cortex, we also found a significant difference between the forward
483 condition in the young group and the backward condition in the older group ($p = 0.002$). Here,
484 we observe an additional effect in the older group, conveying that the ageing brain fails to
485 distinguish between intelligible and unintelligible speech, and even exhibits a reverse pattern
486 by tracking the presented backward speech more than the young group.

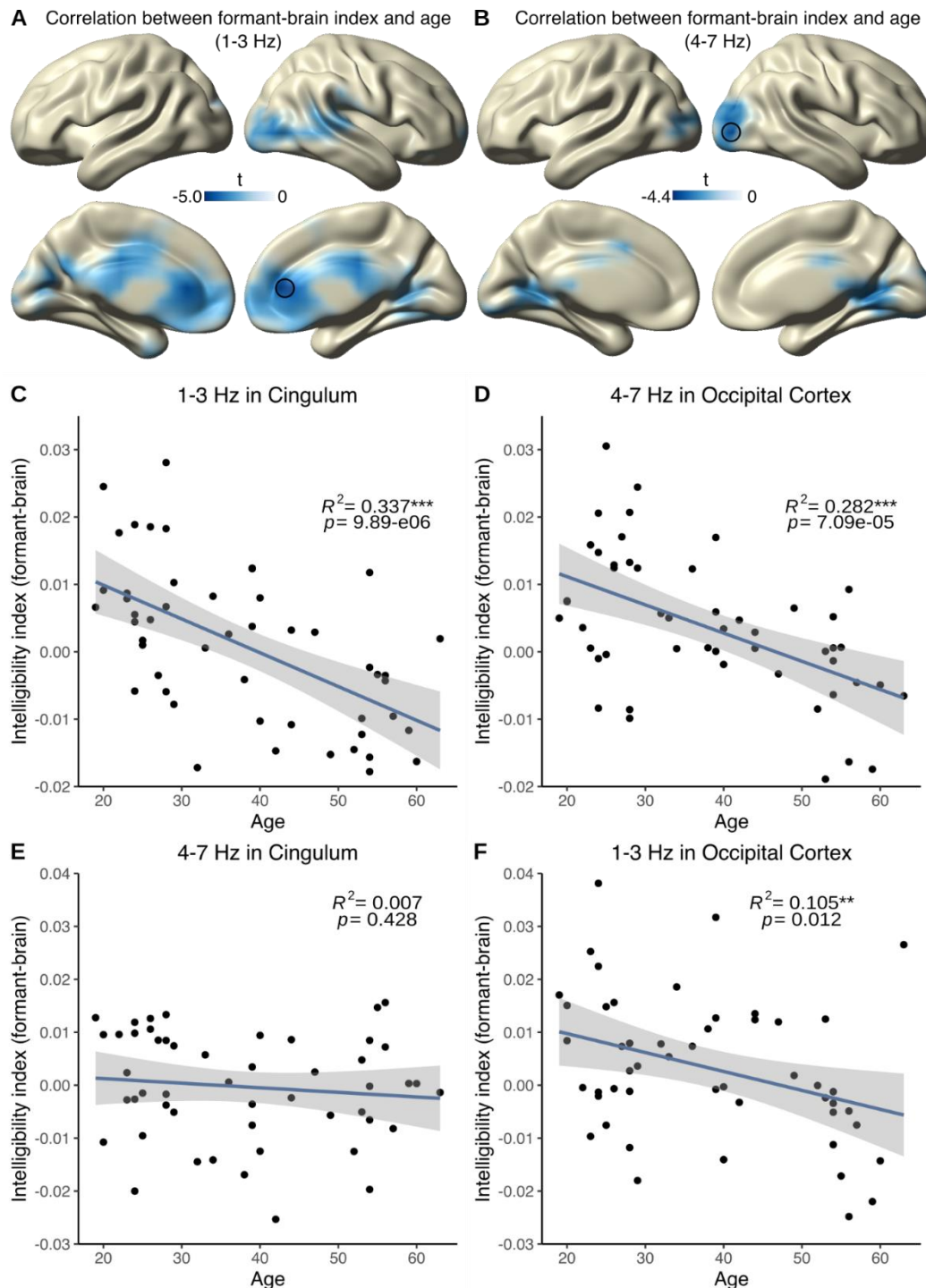
487

488 **3.5 Different frequency-bands show an age-related decline in different brain regions**

489 As a last step, we investigated if different frequency bands are impacted differently by age-
490 related decline. Therefore, we repeated the analysis steps explained in 3.3, meaning that we
491 calculated again a voxelwise correlation between the intelligibility index separately for our
492 coherence conditions (lip-brain, envelope-brain, formant-brain and pitch-brain) and the age of
493 the participants, but this time separately for the delta-band (1-3 Hz) and the theta-band (4-7
494 Hz). For the delta-band, we again found a significant correlation between age and the
495 intelligibility index just for the formant-brain index ($p = 0.002$, cluster-corrected). This effect
496 was strongest in cingulate cortex (lowest t -value: -4.991 , MNI [0 40 10], Figure 4A). No
497 correlation occurred between age and the other indices (lip-brain index: $p = 0.833$; envelope-
498 brain index: $p = 0.268$; pitch-brain index: $p = 0.166$, all cluster-corrected). Repeating the
499 analysis for the theta-band revealed a similar picture: While we could find a significant
500 correlation between the formant-brain index and age ($p = 0.018$, cluster-corrected) which was
501 strongest in visual cortex (lowest t -value: -4.394 , MNI [40 -90 0], Figure 4B), we did not find it
502 for the remaining indices and age (lip-brain index: $p = 1$; envelope-brain index: $p = 0.096$;
503 pitch-brain index: $p = 0.675$, all cluster-corrected). These results display a differential spatial
504 pattern for different frequency bands: While tracking of intelligible speech in the theta-band
505 declines reliably in visual cortex, tracking of the slower delta-band rather declines in cingulate
506 cortex and frontal areas. We then extracted the values from the voxel with the lowest t -value
507 (i.e. the most significant negative one) respectively for both frequency bands (delta-band:
508 cingulate cortex, MNI [0 40 10]; theta-band: visual cortex, MNI [40 -90 0]) and again fitted a
509 linear model for the formant-brain index to further clarify the effects found in the whole brain
510 analysis. Age could significantly predict the formant-brain index in the delta-band in cingulate
511 cortex ($F(1, 48) = 24.4$, $p = 9.885e-06$, $\eta^2 = 0.337$, Figure 4C) and in the theta-band in visual
512 cortex ($F(1, 48) = 18.92$, $p = 7.089e-05$, $\eta^2 = 0.282$, Figure 4D). To further clarify if the tested
513 relationship is specific to a certain frequency band and brain region, we also tested the vice
514 versa relationship (i.e. the relationship between age and theta-band in cingulate cortex and
515 the relationship between age and delta-band in occipital cortex). We found that while age could

516 not significantly predict the formant-brain index in the theta-band in cingulate cortex ($F(1, 48)$
517 $= 0.637$, $p = 0.429$, $\eta^2 = 0.01$, Figure 4E), it could significantly predict the formant-brain index
518 in the delta-band in occipital cortex ($F(1, 48) = 6.757$, $p = 0.012$, $\eta^2 = 0.123$, Figure 4F). This
519 suggests that while the ability of the cingulate cortex to transform visual into phonological
520 information declines just in the delta-band, the occipital cortex shows a decline over a broad
521 range of frequencies and therefore in general visual speech processing.

522
523



524 *Figure 4: Statistical values of the voxelwise correlation of the formant-brain index with age split*
525 *between delta-band and theta-band. A) Tracking of the intelligibility index in the delta-band (1-*
526 *3 Hz, $p < 0.05$, cluster-corrected) indicates a strong decrease of intelligibility tracking in*
527 *cingulate cortex and frontal areas. Black circle indicates lowest t-value extracted for C and F.*
528 *B) Tracking of the intelligibility index in the theta-band (4-7 Hz, $p < 0.05$, cluster-corrected)*
529 *indicates a strong decrease of intelligibility tracking in visual areas. Black circle indicates*
530 *lowest t-value extracted for D and E. C) Correlation of formant-brain intelligibility index in the*
531 *voxel with the lowest t-value extracted from A (cingulate cortex) showing a significant decrease*
532 *with age ($p = 9.885e-06$) in the delta-band. D) Correlation of formant-brain intelligibility index*
533 *in the voxel with the lowest t-value extracted from B (visual cortex) showing a significant*
534 *decrease with age ($p = 7.089e-05$) in the theta-band. E) Correlation of formant-brain*
535 *intelligibility index in the voxel with the lowest t-value extracted from A (cingulate cortex)*
536 *showing no significant decrease with age ($p = 0.428$) in the theta-band. F) Correlation of*
537 *formant-brain intelligibility index in the voxel with the lowest t-value extracted from B (occipital*
538 *cortex) showing a significant decrease with age ($p = 0.012$) also in the delta-band. Blue lines*
539 *depict regression lines, shaded areas depict standard error of mean (SE).*

540 **4 Discussion**

541 Our study illustrates that during lip reading, the visual cortex represents multiple features of
542 the speech signal in low frequency bands (1-7 Hz), importantly including the corresponding
543 (unheard) acoustic signal. It has previously been shown that the visual cortex is able to track
544 the intelligible global envelope (unheard acoustic speech envelope; Hauswald et al. 2018).
545 We demonstrate here that the visual cortex is also able to track the modulation of intelligible
546 spectral fine-details (unheard formants and pitch). Furthermore, we found that ageing is
547 associated with a deterioration of this ability not only in the visual cortex, but also in the
548 cingulate cortex. Disentangling delta and theta-band revealed that while the age-related
549 decline of formant tracking is independent of frequency bands in visual cortex, it is unique in
550 cingulate cortex for the delta-band. Our results suggest that visuo-phonological transformation
551 processes are sensitive to age-related decline, in particular with regards to the modulation of
552 unheard spectral fine-details.

553

554 ***Visuo-phonological transformation processes are observable for global amplitude*** 555 ***modulations and spectral-fine detail modulations***

556 As expected, the current study replicates the main finding from Hauswald et al. (2018) showing
557 a visuo-phonological transformation process in visual cortex for the unheard speech envelope
558 in an Italian speaking sample. Our study using a German speaking sample suggests that the
559 postulated visuo-phonological transformation process at the level of the visual cortex is
560 generalizable across languages. This is unsurprising as it is in line with studies on the speech
561 envelope spectrum which show robust amplitude peaks between 3.5 and 4.5 Hz regardless of
562 language (Poeppel & Assaneo, 2020), providing evidence that different languages carry
563 similar temporal regularities not only for auditory properties, but also for visual properties
564 (Chandrasekaran et al., 2009). We argue that this similarity is a key property for making the
565 postulated visuo-phonological transformation process transferable to other languages.

566 By investigating different properties of auditory speech (global modulations vs. fine-detailed
567 modulations) and how they are tracked by the human brain, our results are furthermore adding
568 an important part to the understanding of how visual speech contributes to speech processing
569 in general. As lip movements and amplitude modulations are highly correlated
570 (Chandrasekaran et al., 2009), it is highly probable that amplitude modulations can be inferred
571 by lip movements alone as a learned association. Here we can show that the brain is also able
572 to perform a more fine-coarsed tracking than initially thought by especially processing the
573 spectral fine-details that are modulated near the lips, another potentially learned association
574 between lip-near auditory cues (i.e. merged F2 and F3 formants) and lip movements (Plass et
575 al., 2020). Additionally, it is not only formants that are subject to visuo-phonological
576 transformation, but also the fundamental frequency, as seen in our results. This is in line with

577 a recent study which shows that closing the lips is correlated with the tone falling (Garg et al.,
578 2019). How those modulations are influenced by behavioural measures still needs to be
579 discussed. Some studies suggest that enhanced lip reading abilities go in line with higher
580 activation in visual areas in persons with a cochlear implant (e.g. Giraud et al., 2001). Our
581 present results do not suggest that strong visuo-phonological transformation processes are
582 sufficient for improved lip reading abilities. Yet, they may be most useful in disambiguating
583 auditory signals in difficult listening situations.

584

585 ***Tracking of unheard formants accompanying lip movements is mostly affected in***
586 ***ageing***

587 With regards to the ageing effect, we could show that various neural tracking mechanisms are
588 differentially affected. Our study presents that tracking of unheard formants, especially the
589 combined F2 and F3 formants, is declining with age, while there is still a preserved tracking of
590 purely visual information (as seen in the lip-brain index, Figures 3C and 3F). Meanwhile, the
591 tracking of the unheard speech envelope and pitch signify an inconclusive picture: While
592 tracking of those properties seem to be preserved to some extent, both are showing a
593 tendency to diminish with age.

594 Especially the formants and the pitch are part of the temporal fine-structure (TFS) of speech
595 and are crucial for speech segregation or perceptual grouping for optimal speech processing
596 in complex situations (Alain et al., 2017; Bregman et al., 1990). The TFS is different from the
597 acoustic envelope in a sense that it does not display “coarse” amplitude modulations of the
598 audio signal but rather fluctuations that are close to the centre frequency of certain frequency
599 bands (Lorenzi et al., 2006). Hearing-impaired older participants show a relative deficit of the
600 representation of the TFS compared to the acoustic envelope (Anderson et al., 2013; Lorenzi
601 et al., 2006). The TFS also yields information when trying to interpret speech in fluctuating
602 background noise (Moore, 2008). Other studies also point to the fact that especially when
603 having cochlear hearing loss along with a normal audiometric threshold, the interpretation of
604 the TFS is reduced, resulting in diminished speech perception under noisy conditions (Lorenzi
605 et al., 2009). This suggests that hearing-impaired subjects mainly seem to rely on the temporal
606 envelope to interpret auditory information (Moore & Moore, 2003), while normal hearing
607 subjects can also use the presented temporal fine-structure. Interestingly, we found that even
608 when the TFS is inferred from lip movements, there is a decline in the processing of spectral
609 fine-details with age independent of hearing loss. Our results suggest that the visuo-
610 phonological transformation of certain spectral fine-details like the formants are impacted the
611 most in ageing, whereas the transformation of the pitch (or fundamental frequency) reveals a
612 more complex picture: We find preserved tracking of the unheard pitch contour in occipital
613 cortex, but a decline with age in the cingulate cortex. Interestingly, the cingulate cortex has

614 been found to show higher activation as response to processing of degraded speech (Erb &
615 Obleser, 2013), pointing to a possible compensatory mechanism when processing distorted
616 speech. How this altered processing of the unheard pitch (or fundamental frequency)
617 accompanying lip movements in cingulate cortex has an impact on speech understanding
618 needs to be discussed in further studies.

619 Further investigating the effects shown in our correlational analysis revealed that older
620 participants seem to be less able to distinguish between forward and backward unheard
621 speech (unheard formants) and that younger individuals show enhanced tracking of intelligible
622 speech (Figure 3D). This could point to the fact that the older population is losing the gain of
623 differentiating intelligible from unintelligible speech, obviously resulting in a less successful
624 visuo-phonological transformation process. Other studies suggest that the older population
625 seems to inefficiently use their cognitive resources, showing less deterioration of cortical
626 responses (measured by the envelope reconstruction accuracy) to a foreign language
627 compared to younger individuals (Presacco et al., 2016b) and also an association between
628 cognitive decline and increased cortical envelope tracking or even higher synchronization of
629 theta (Goossens et al., 2016). Auditory processing is also affected both in midbrain and cortex
630 in age, exhibiting a large reduction of speech envelope encoding when presented with a
631 competing talker, but at the same time a cortical overrepresentation of speech regardless of
632 the presented noise, suggesting an imbalance between inhibition and excitation in the human
633 brain (Presacco et al., 2016a) when processing speech. Other studies add to this hypothesis
634 by showing decreasing alpha modulation in the ageing population (Henry et al., 2017; Vaden
635 et al., 2012), strengthening the assumption that there is an altered interaction between age
636 and cortical tracking even in the visual modality that needs to be investigated further.

637 Considering all acoustic details accompanying lip movements we still see a tendency of the
638 speech envelope tracking to decline with age, suggesting that the transformation of the global
639 speech dynamics could also be deteriorating. Overall, our results provide evidence that the
640 transformation of fine-grained acoustic details seem to decline more reliably with age, while
641 the transformation of global information (in our case the speech envelope) seems to be less
642 impaired.

643

644 ***Possible implications for speech processing in challenging situations***

645 Our findings raise the question of how the decline in processing of unheard spectral fine-
646 details negatively influences other relevant aspects of hearing. In light of aforementioned
647 studies from the auditory domain of speech processing, we propose some thoughts on the
648 multi-sensory nature of speech and how different sensory modalities can contribute to speech
649 processing abilities under disadvantageous conditions (both intrapersonal and
650 environmental).

651 As mentioned in the previous section, optimal hearing requires processing of both the temporal
652 fine structure and the global acoustic envelope. However, especially under noisy conditions,
653 processing the TFS becomes increasingly important for understanding speech. Ageing in
654 general goes along with reduced processing of the TFS (Anderson & Karawani, 2020) and
655 this deteriorating effect seems to be even more detrimental when ageing is accompanied by
656 hearing loss (Anderson et al., 2013). Since listening in natural situations usually is a multi-
657 sensory (audiovisual) phenomenon, we argue that the impaired visuo-phonological
658 transformation process of the TFS adds to the difficulties of older (also audiometrically normal
659 hearing) individuals to follow speech in challenging situations. To follow up this idea, future
660 studies will need to quantify the benefit of audiovisual versus (unimodal) auditory processing,
661 depending on different visuo-phonological transformation abilities.

662 Our results also have implications for listening situations when relevant visual input from the
663 mouth area is obscured, a topic which has gained enormously in significance due to the wide
664 adoption of face masks to counteract the spread of SARS-CoV-2. In general, listening
665 becomes more difficult and performance declines when the mouth area is obscured (Brown et
666 al., 2021; Giovanelli et al., 2021). While face masks may diminish attentional focusing as well
667 as temporal cues, our work suggests that they also deprive the brain of deriving the acoustic
668 TFS from the lip movements especially in the formant frequency range which are modulated
669 near the lips (F2 and F3). This issue, which should become relevant particularly in noisy
670 situations, may be aggravated by the fact that face masks (especially highly protective ones)
671 impact sound propagation of frequencies between 1600-6000 Hz with a peak around 2000 Hz
672 (Caniato et al., 2021). Thus, face masks diminish relevant formant information in both sensory
673 modalities. This could disproportionately affect hearing impaired listeners, an urgent question
674 that should be followed up by future studies.

675 Overall, considering both the auditory and visual domain of speech properties, we suggest
676 that the underlying cause of speech processing difficulties in naturalistic settings
677 accompanying age or hearing impairment is more diverse than previously thought. The visual
678 system provides the proposed visuo-phonological transformation process as an important
679 mechanism for optimal speech understanding and crucially supports acoustic speech
680 processing.

681

682 ***Occipital cortex and cingulate cortex show different tracking properties dependent on***
683 ***the frequency-band***

684 With regards to different frequency bands, our results could yield important insights into
685 different brain regions showing distinct formant tracking properties: While we find a robust
686 decline of delta-band tracking with age in both occipital and cingulate cortex, theta-band
687 tracking is reliably declining only in occipital areas. In general, theta is corresponding to the

688 frequency of syllables and to the modulations in the amplitude envelope (Gross et al., 2013;
689 Keitel et al., 2018; Meyer, 2018; Poeppel & Assaneo, 2020), whereas delta seems to process
690 phrasal chunks based on acoustic cues (Ghitza, 2017; Keitel et al., 2018) and is therefore
691 responsible for a general perceptual chunking mechanism (Boucher et al., 2019). Our results
692 also show that the visual cortex extracts information provided by the perception of the lip
693 movements and connects them with phonological information that is already learned. This
694 points to a possible top-down influence of stored syntactic information provided by delta-band
695 tracking, which also seems to be deteriorating with increasing age both in occipital and
696 cingulate cortex. Interestingly, age-related hearing loss also leads to a volume reduction in
697 anterior cingulate cortex (Slade et al., 2020), which in turn also leads to more memory
698 impairments and cognitive deficits (Belkhiria et al., 2019). These and our current results
699 strengthen the notion that the cingulate cortex has an important function also in visual speech
700 processing, as this also goes in line with the mentioned compensatory mechanism in anterior
701 cingulate cortex (ACC) (Erb & Obleser, 2013). Together with the findings of the current study,
702 this involvement of the cingulate cortex in speech processing (or in general the cingulo-
703 opercular network; Peelle (2018)) underlines the fact that there seems to be a maladaptive
704 processing strategy in frontal areas. To fully understand the mechanisms behind this visuo-
705 phonological transformation process without the influence of ageing in distinct brain regions
706 and frequency bands, it would be advisable for future studies to focus on younger individuals,
707 especially since this study is the first to investigate the tracking of spectral fine-details
708 extracted from the spectrogram.

709 **5 Conclusion**

710 The current study demonstrates that the visual cortex is able to track intelligible unheard
711 spectral-fine detailed information just by observing lip movements. Crucially, we present a
712 differential pattern for the processing of global (i.e. envelope) and spectral fine-detailed
713 intelligible information, with ageing affecting in particular tracking of spectral speech
714 information (or the TFS), while showing partly preserved tracking of global modulations.
715 Furthermore, we see a distinct age-related decline of tracking dependent on the brain region
716 (i.e. visual and cingulate cortex) and on the frequency-band (i.e. delta and theta-band). The
717 results presented here may have important implications for hearing in the ageing population,
718 suggesting that hearing difficulties could also be exacerbated in natural audiovisual
719 environments as a result of reduced capacities of visual benefit. With respect to the current
720 pandemic situation, our results can provide a novel important insight on how missing visual
721 input (e.g. when carrying face masks) is critical for speech comprehension.

722 **6 Competing Interest Statement**

723 The authors have declared no competing interest.

724

725 **7 Acknowledgements**

726 This work is supported by the Austrian Science Fund, P31230 (“Audiovisual speech
727 entrainment in deafness”).

728

729 **8 Pre-registration**

730 The first part of the study analyses was pre-registered prior to the research being conducted
731 under <https://osf.io/ndvf6/>.

732

733 **9 Data availability**

734 The “mat” and “csv” files containing the data shown in the figures, along with the MATLAB
735 code and the R code to recreate the plots, are available under <https://osf.io/ndvf6/>. Readers
736 seeking access to the original, non-resampled data (~430 GB) should contact the lead
737 author (nina.suess@sbg.ac.at). Access will be granted in accordance with ethical
738 procedures governing the reuse of sensitive data.

739 **10 References**

- 740 Alain, C., Arsenault, J. S., Garami, L., Bidelman, G. M., & Snyder, J. S. (2017). Neural
741 Correlates of Speech Segregation Based on Formant Frequencies of Adjacent
742 Vowels. *Scientific Reports*, 7(1), 40790. <https://doi.org/10.1038/srep40790>
- 743 Anderson, S., & Karawani, H. (2020). Objective evidence of temporal processing deficits in
744 older adults. *Hearing Research*, 397, 108053.
745 <https://doi.org/10.1016/j.heares.2020.108053>
- 746 Anderson, S., Parbery-Clark, A., White-Schwoch, T., Drehobl, S., & Kraus, N. (2013). Effects
747 of hearing loss on the subcortical representation of speech cues. *The Journal of the*
748 *Acoustical Society of America*, 133(5), 3030–3038. <https://doi.org/10.1121/1.4799804>
- 749 Badin, P., Perrier, P., Boë, L., & Abry, C. (1990). Vocalic nomograms: Acoustic and
750 articulatory considerations upon formant convergences. *The Journal of the Acoustical*
751 *Society of America*, 87(3), 1290–1300. <https://doi.org/10.1121/1.398804>
- 752 Belkhiria, C., Vergara, R. C., San Martín, S., Leiva, A., Marcenaro, B., Martinez, M.,
753 Delgado, C., & Delano, P. H. (2019). Cingulate Cortex Atrophy Is Associated With
754 Hearing Loss in Presbycusis With Cochlear Amplifier Dysfunction. *Frontiers in Aging*
755 *Neuroscience*, 11. <https://doi.org/10.3389/fnagi.2019.00097>
- 756 Bernstein, L. E., & Liebenthal, E. (2014). Neural pathways for visual speech perception.
757 *Frontiers in Neuroscience*, 8. <https://doi.org/10.3389/fnins.2014.00386>
- 758 Boersma, P., & Weenink, D. (2019). *Praat: Doing phonetics by computer [Computer*
759 *program]* (6.0.48) [Computer software]. <http://www.praat.org/>
- 760 Boucher, V. J., Gilbert, A. C., & Jemel, B. (2019). The Role of Low-frequency Neural
761 Oscillations in Speech Processing: Revisiting Delta Entrainment. *Journal of Cognitive*
762 *Neuroscience*, 31(8), 1205–1215. https://doi.org/10.1162/jocn_a_01410
- 763 Bourguignon, M., Baart, M., Kapnola, E. C., & Molinaro, N. (2020). Lip-Reading Enables
764 the Brain to Synthesize Auditory Features of Unknown Silent Speech. *Journal of*
765 *Neuroscience*, 40(5), 1053–1065. <https://doi.org/10.1523/JNEUROSCI.1101-19.2019>
- 766 Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision*, 10(4), 433–436.

- 767 <https://doi.org/10.1163/156856897X00357>
- 768 Bregman, A. S., Liao, C., & Levitan, R. (1990). Auditory grouping based on fundamental
769 frequency and formant peak frequency. *Canadian Journal of Psychology*, *44*(3), 400–
770 413. <https://doi.org/10.1037/h0084255>
- 771 Brown, V. A., Engen, K. V., & Peelle, J. E. (2021). *Face mask type affects audiovisual*
772 *speech intelligibility and subjective listening effort in young and older adults*.
773 PsyArXiv. <https://doi.org/10.31234/osf.io/7waj3>
- 774 Caniato, M., Marzi, A., & Gasparella, A. (2021). How much COVID-19 face protections
775 influence speech intelligibility in classrooms? *Applied Acoustics*, *178*, 108051.
776 <https://doi.org/10.1016/j.apacoust.2021.108051>
- 777 Chandrasekaran, C., Trubanova, A., Stillitano, S., Caplier, A., & Ghazanfar, A. A. (2009).
778 The Natural Statistics of Audiovisual Speech. *PLoS Computational Biology*, *5*(7).
779 <https://doi.org/10.1371/journal.pcbi.1000436>
- 780 Crosse, M. J., Liberto, G. M. D., & Lalor, E. C. (2016). Eye Can Hear Clearly Now: Inverse
781 Effectiveness in Natural Audiovisual Speech Processing Relies on Long-Term
782 Crossmodal Temporal Integration. *Journal of Neuroscience*, *36*(38), 9888–9895.
783 <https://doi.org/10.1523/JNEUROSCI.1396-16.2016>
- 784 Erb, J., & Obleser, J. (2013). Upregulation of cognitive control networks in older adults’
785 speech comprehension. *Frontiers in Systems Neuroscience*, *7*.
786 <https://doi.org/10.3389/fnsys.2013.00116>
- 787 Erb, J., Schmitt, L.-M., & Obleser, J. (2020). Temporal selectivity declines in the aging
788 human auditory cortex. *eLife*, *9*, e55300. <https://doi.org/10.7554/eLife.55300>
- 789 Escoffier, N., Herrmann, C. S., & Schirmer, A. (2015). Auditory rhythms entrain visual
790 processes in the human brain: Evidence from evoked oscillations and event-related
791 potentials. *NeuroImage*, *111*, 267–276.
792 <https://doi.org/10.1016/j.neuroimage.2015.02.024>
- 793 Feld, J., & Sommers, M. (2009). Lipreading, Processing Speed, and Working Memory in
794 Younger and Older Adults. *Journal of Speech, Language, and Hearing Research*,

- 795 52(6), 1555–1565. [https://doi.org/10.1044/1092-4388\(2009/08-0137\)](https://doi.org/10.1044/1092-4388(2009/08-0137))
- 796 Garg, S., Hamarneh, G., Jongman, A., Sereno, J. A., & Wang, Y. (2019). Computer-vision
797 analysis reveals facial movements made during Mandarin tone production align with
798 pitch trajectories. *Speech Communication*, 113, 47–62.
799 <https://doi.org/10.1016/j.specom.2019.08.003>
- 800 Ghitza, O. (2017). Acoustic-driven delta rhythms as prosodic markers. *Language, Cognition*
801 *and Neuroscience*, 32(5), 545–561. <https://doi.org/10.1080/23273798.2016.1232419>
- 802 Giovanelli, E., Valzolgher, C., Gessa, E., Todeschini, M., & Pavani, F. (2021). Unmasking
803 the Difficulty of Listening to Talkers With Masks: Lessons from the COVID-19
804 pandemic. *I-Perception*, 12(2), 2041669521998393.
805 <https://doi.org/10.1177/2041669521998393>
- 806 Giraud, A.-L., & Poeppel, D. (2012). Cortical oscillations and speech processing: Emerging
807 computational principles and operations. *Nature Neuroscience*, 15(4), 511–517.
808 <https://doi.org/10.1038/nn.3063>
- 809 Giraud, A.-L., Price, C. J., Graham, J. M., Truy, E., & Frackowiak, R. S. J. (2001). Cross-
810 Modal Plasticity Underpins Language Recovery after Cochlear Implantation. *Neuron*,
811 30(3), 657–664. [https://doi.org/10.1016/S0896-6273\(01\)00318-X](https://doi.org/10.1016/S0896-6273(01)00318-X)
- 812 Goossens, T., Vercammen, C., Wouters, J., & Wieringen, A. van. (2016). Aging Affects
813 Neural Synchronization to Speech-Related Acoustic Modulations. *Frontiers in Aging*
814 *Neuroscience*, 8. <https://doi.org/10.3389/fnagi.2016.00133>
- 815 Gross, J., Hoogenboom, N., Thut, G., Schyns, P., Panzeri, S., Belin, P., & Garrod, S. (2013).
816 Speech Rhythms and Multiplexed Oscillatory Sensory Coding in the Human Brain.
817 *PLOS Biology*, 11(12), e1001752. <https://doi.org/10.1371/journal.pbio.1001752>
- 818 Hartmann, T., & Weisz, N. (2020). An Introduction to the Objective Psychophysics Toolbox.
819 *Frontiers in Psychology*, 11. <https://doi.org/10.3389/fpsyg.2020.585437>
- 820 Hauswald, A., Lithari, C., Collignon, O., Leonardelli, E., & Weisz, N. (2018). A Visual Cortical
821 Network for Deriving Phonological Information from Intelligible Lip Movements.
822 *Current Biology*, 28(9), 1453-1459.e3. <https://doi.org/10.1016/j.cub.2018.03.044>

- 823 Henry, M. J., Herrmann, B., Kunke, D., & Obleser, J. (2017). Aging affects the balance of
824 neural entrainment and top-down neural modulation in the listening brain. *Nature*
825 *Communications*, 8, ncomms15801. <https://doi.org/10.1038/ncomms15801>
- 826 Hopkins, K., Moore, B. C. J., & Stone, M. A. (2008). Effects of moderate cochlear hearing
827 loss on the ability to benefit from temporal fine structure information in speech. *The*
828 *Journal of the Acoustical Society of America*, 123(2), 1140–1153.
829 <https://doi.org/10.1121/1.2824018>
- 830 Keitel, A., Gross, J., & Kayser, C. (2018). Perceptually relevant speech tracking in auditory
831 and motor cortex reflects distinct linguistic features. *PLOS Biology*, 16(3), e2004473.
832 <https://doi.org/10.1371/journal.pbio.2004473>
- 833 Keitel, A., Gross, J., & Kayser, C. (2020). Shared and modality-specific brain regions that
834 mediate auditory and visual word comprehension. *ELife*, 9, e56972.
835 <https://doi.org/10.7554/eLife.56972>
- 836 Keitel, A., Ince, R. A. A., Gross, J., & Kayser, C. (2017). Auditory cortical delta-entrainment
837 interacts with oscillatory power in multiple fronto-parietal networks. *NeuroImage*, 147,
838 32–42. <https://doi.org/10.1016/j.neuroimage.2016.11.062>
- 839 Liberman, M. C. (2017). Noise-induced and age-related hearing loss: New perspectives and
840 potential therapies. *F1000Research*, 6.
841 <https://doi.org/10.12688/f1000research.11310.1>
- 842 Löhler, J., Akcicek, B., Kappe, T., Schlattmann, P., Wollenberg, B., & Schönweiler, R.
843 (2014). Entwicklung und Anwendung einer APHAB-Datenbank. *HNO*, 62(10), 735–
844 745. <https://doi.org/10.1007/s00106-014-2915-4>
- 845 Lorenzi, C., Debrulle, L., Garnier, S., Fleuriot, P., & Moore, B. C. J. (2009). Abnormal
846 processing of temporal fine structure in speech for frequencies where absolute
847 thresholds are normal. *The Journal of the Acoustical Society of America*, 125(1), 27–
848 30. <https://doi.org/10.1121/1.2939125>
- 849 Lorenzi, C., Gilbert, G., Carn, H., Garnier, S., & Moore, B. C. J. (2006). Speech perception
850 problems of the hearing impaired reflect inability to use temporal fine structure.

- 851 *Proceedings of the National Academy of Sciences*, 103(49), 18866–18869.
- 852 <https://doi.org/10.1073/pnas.0607364103>
- 853 Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data.
- 854 *Journal of Neuroscience Methods*, 164(1), 177–190.
- 855 <https://doi.org/10.1016/j.jneumeth.2007.03.024>
- 856 Mattout, J., Henson, R. N., & Friston, K. J. (2007, June 25). *Canonical Source*
- 857 *Reconstruction for MEG* [Research Article]. Computational Intelligence and
- 858 Neuroscience; Hindawi. <https://doi.org/10.1155/2007/67613>
- 859 Meyer, L. (2018). The neural oscillations of speech processing and language
- 860 comprehension: State of the art and emerging mechanisms. *European Journal of*
- 861 *Neuroscience*, 48(7), 2609–2621. <https://doi.org/10.1111/ejn.13748>
- 862 Moore, B. C. J. (2008). The Role of Temporal Fine Structure Processing in Pitch Perception,
- 863 Masking, and Speech Perception for Normal-Hearing and Hearing-Impaired People.
- 864 *Journal of the Association for Research in Otolaryngology*, 9(4), 399–406.
- 865 <https://doi.org/10.1007/s10162-008-0143-x>
- 866 Moore, B. C. J., & Moore, G. A. (2003). Discrimination of the fundamental frequency of
- 867 complex tones with fixed and shifting spectral envelopes by normally hearing and
- 868 hearing-impaired subjects. *Hearing Research*, 182(1), 153–163.
- 869 [https://doi.org/10.1016/S0378-5955\(03\)00191-6](https://doi.org/10.1016/S0378-5955(03)00191-6)
- 870 Nolte, G. (2003). The magnetic lead field theorem in the quasi-static approximation and its
- 871 use for magnetoencephalography forward calculation in realistic volume conductors.
- 872 *Physics in Medicine & Biology*, 48(22), 3637. [https://doi.org/10.1088/0031-](https://doi.org/10.1088/0031-9155/48/22/002)
- 873 [9155/48/22/002](https://doi.org/10.1088/0031-9155/48/22/002)
- 874 O'Brien, F., & Cousineau, D. (2014). Representing Error bars in within-subject designs in
- 875 typical software packages. *Tutorials in Quantitative Methods for Psychology*, 10(1),
- 876 56–67.
- 877 Oostenveld, R., Fries, P., Maris, E., & Schoffelen, J.-M. (2011). FieldTrip: Open source
- 878 software for advanced analysis of MEG, EEG, and invasive electrophysiological data.

- 879 *Intell. Neuroscience*, 2011, 1:1-1:9. <https://doi.org/10.1155/2011/156869>
- 880 O’Sullivan, A. E., Crosse, M. J., Di Liberto, G. M., & Lalor, E. C. (2017). Visual Cortical
881 Entrainment to Motion and Categorical Speech Features during Silent Lipreading.
882 *Frontiers in Human Neuroscience*, 10. <https://doi.org/10.3389/fnhum.2016.00679>
- 883 Park, H., Kayser, C., Thut, G., & Gross, J. (2016). Lip movements entrain the observers’ low-
884 frequency brain oscillations to facilitate speech intelligibility. *ELife*, 5, e14521.
885 <https://doi.org/10.7554/eLife.14521>
- 886 Peelle, J. E. (2018). Listening Effort: How the Cognitive Consequences of Acoustic
887 Challenge Are Reflected in Brain and Behavior. *Ear and Hearing*, 39(2), 204–214.
888 <https://doi.org/10.1097/AUD.0000000000000494>
- 889 Plass, J., Brang, D., Suzuki, S., & Grabowecy, M. (2020). Vision perceptually restores
890 auditory spectral dynamics in speech. *Proceedings of the National Academy of*
891 *Sciences*. <https://doi.org/10.1073/pnas.2002887117>
- 892 Poeppel, D., & Assaneo, M. F. (2020). Speech rhythms and their neural foundations. *Nature*
893 *Reviews Neuroscience*, 1–13. <https://doi.org/10.1038/s41583-020-0304-4>
- 894 Presacco, A., Simon, J. Z., & Anderson, S. (2016a). Evidence of degraded representation of
895 speech in noise, in the aging midbrain and cortex. *Journal of Neurophysiology*,
896 116(5), 2346–2355. <https://doi.org/10.1152/jn.00372.2016>
- 897 Presacco, A., Simon, J. Z., & Anderson, S. (2016b). Effect of informational content of noise
898 on speech representation in the aging midbrain and cortex. *Journal of*
899 *Neurophysiology*, 116(5), 2356–2367. <https://doi.org/10.1152/jn.00373.2016>
- 900 Slade, K., Plack, C. J., & Nuttall, H. E. (2020). The Effects of Age-Related Hearing Loss on
901 the Brain and Cognitive Function. *Trends in Neurosciences*, 43(10), 810–821.
902 <https://doi.org/10.1016/j.tins.2020.07.005>
- 903 Smith, Z. M., Delgutte, B., & Oxenham, A. J. (2002). Chimaeric sounds reveal dichotomies in
904 auditory perception. *Nature*, 416(6876), 87–90. <https://doi.org/10.1038/416087a>
- 905 Suess, N., Hauswald, A., Zehentner, V., Depireux, J., Herzog, G., Rösch, S., & Weisz, N.
906 (2021). *Influence of linguistic properties and hearing impairment on lip reading skills*

- 907 *in the German language*. PsyArXiv. <https://doi.org/10.31234/osf.io/rcfxv>
- 908 Sumbly, W. H., & Pollack, I. (1954). Visual Contribution to Speech Intelligibility in Noise. *The*
909 *Journal of the Acoustical Society of America*, 26(2), 212–215.
910 <https://doi.org/10.1121/1.1907309>
- 911 Taulu, S., Simola, J., & Kajola, M. (2005). Applications of the signal space separation
912 method. *IEEE Transactions on Signal Processing*, 53(9), 3359–3372.
913 <https://doi.org/10.1109/TSP.2005.853302>
- 914 Tun, P. A., & Wingfield, A. (1999). One Voice Too Many: Adult Age Differences in Language
915 Processing With Different Types of Distracting Sounds. *The Journals of Gerontology:*
916 *Series B*, 54B(5), P317–P327. <https://doi.org/10.1093/geronb/54B.5.P317>
- 917 Tye-Murray, N., Sommers, M. S., & Spehar, B. (2007a). Audiovisual Integration and
918 Lipreading Abilities of Older Adults with Normal and Impaired Hearing. *Ear and*
919 *Hearing*, 28(5), 656–668. <https://doi.org/10.1097/AUD.0b013e31812f7185>
- 920 Tye-Murray, N., Sommers, M., & Spehar, B. (2007b). The Effects of Age and Gender on
921 Lipreading Abilities. *Journal of the American Academy of Audiology*, 18(10), 883–
922 892.
- 923 Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N.,
924 Mazoyer, B., & Joliot, M. (2002). Automated Anatomical Labeling of Activations in
925 SPM Using a Macroscopic Anatomical Parcellation of the MNI MRI Single-Subject
926 Brain. *NeuroImage*, 15(1), 273–289. <https://doi.org/10.1006/nimg.2001.0978>
- 927 Vaden, R. J., Hutcheson, N. L., McCollum, L. A., Kentros, J., & Visscher, K. M. (2012). Older
928 adults, unlike younger adults, do not modulate alpha power to suppress irrelevant
929 information. *NeuroImage*, 63(3), 1127–1133.
930 <https://doi.org/10.1016/j.neuroimage.2012.07.050>
- 931 Veen, B. D. V., Drongelen, W. V., Yuchtman, M., & Suzuki, A. (1997). Localization of brain
932 electrical activity via linearly constrained minimum variance spatial filtering. *IEEE*
933 *Transactions on Biomedical Engineering*, 44(9), 867–880.
934 <https://doi.org/10.1109/10.623056>

935 Wong, P. C. M., Jin, J. X., Gunasekera, G. M., Abel, R., Lee, E. R., & Dhar, S. (2009). Aging
936 and cortical mechanisms of speech perception in noise. *Neuropsychologia*, *47*(3),
937 693–703. <https://doi.org/10.1016/j.neuropsychologia.2008.11.032>