Largely distinct networks mediate perceptually-relevant auditory and visual speech representations

Anne Keitel[*1,2], *Joachim Gross*[1,3], *Christoph Kayser*[4]

1) Institute of Neuroscience and Psychology, University of Glasgow, 62 Hillhead Street, Glasgow G12 8QB, UK
2) Psychology, University of Dundee, Scrymgeour Building, Dundee DD1 4HN, UK
3) Institute for Biomagnetism and Biosignalanalysis, University of Münster, Malmedyweg 15, 48149 Münster, Germany
4) Department for Cognitive Neuroscience & Cognitive Interaction Technology, Center of Excellence, Bielefeld University, 33615 Bielefeld, Germany

* *Corresponding author:* Anne Keitel, Psychology, School of Social Sciences, Scrymgeour Building, Dundee DD1 4HN, UK

Tel.: +44 (0)1382 386 754

E-mail: a.keitel@dundee.ac.uk

Acknowledgements: This research was supported by the UK Biotechnology and Biological Sciences Research Council (BBSRC, BB/L027534/1). CK is supported by the European Research Council (ERC-2014-CoG; grant No 646657); JG by the Wellcome Trust (Joint Senior Investigator Grant, No 098433). The authors declare no competing financial interests.

Abstract

Visual speech is an integral part of communication. Yet it remains unclear whether semantic information carried by movements of the lips or tongue is represented in the same brain regions that mediate acoustic speech representations. Behaviourally, our ability to understand acoustic speech seems independent from that to understand visual speech, but neuroimaging studies suggest that acoustic and visual speech representations largely overlap. To resolve this discrepancy, and to understand whether acoustic and lip-reading speech comprehension are mediated by the same cerebral representations, we systematically probed where the brain represents acoustically and visually conveyed word identities in a human MEG study. We designed a single-trial classification paradigm to dissociate where cerebral representations merely reflect the sensory stimulus and where they are predictive of the participant's percept. In general, those brain regions allowing for the highest word classification were distinct from those in which cerebral representations were predictive of participant's percept. Across the brain, word representations were largely modality-specific and auditory and visual comprehension were mediated by distinct left-lateralised ventral and dorsal fronto-temporal regions, respectively. Only within the inferior frontal gyrus and the anterior temporal lobe did auditory and visual representations converge. These results provide a neural explanation for why acoustic speech comprehension is a poor predictor of lip-reading skills and suggests that those cerebral speech representations that encode word identity may be more modality-specific than often upheld.

Words abstract: 226.

Keywords: visual speech, speech decoding, MEG, lip reading, speech reading, auditory pathways, audio-visual integration

Introduction

Acoustic and visual speech signals are both elemental for everyday communication. While acoustic speech consists of temporal and spectral modulations of sound pressure, visual speech consists of movements of the mouth, head, and hands. Movements of the lips, teeth and tongue in particular provide both redundant and complementary information to acoustic cues (Hall, Fussell, & Summerfield, 2005; Peelle & Sommers, 2015; Summerfield, 1992), and can help to enhance speech intelligibility in noisy environments or in a second language (Navarra & Soto-Faraco, 2007; Sumby & Pollack, 1954; Yi, Wong, & Eizenman, 2013). While a plethora of studies have investigated the cerebral mechanisms underlying speech in general, we still have a limited understanding of the networks specifically mediating visual speech perception, i.e. lip-reading (Bernstein & Liebenthal, 2014; Capek et al., 2008; Crosse, ElShafei, Foxe, & Lalor, 2015). In particular, it remains unclear whether visual speech signals are largely represented in specific and dedicated regions, or whether these visual signals are encoded by the same networks that mediate auditory speech perception.

Behaviourally, our ability to understand acoustic speech seems to be independent from our ability to understand visual speech. In the typical adult population, performance in auditory/verbal and visual speech comprehension tasks are uncorrelated (Conrad, 1977; Jeffers & Barley, 1980; Mohammed, Campbell, Macsweeney, Barry, & Coleman, 2006; Summerfield, 1991, 1992). In contrast to this behavioural dissociation, neuroimaging and neuroanatomical studies have suggested the convergence of acoustic and visual speech information in some brain regions (Calvert et al., 1997; Campbell, 2007; Ralph, Jefferies, Patterson, & Rogers, 2017; Simanova, Hagoort, Oostenveld, & Van Gerven, 2012). Prevalent models postulate a fronto-temporal network mediating acoustic speech representations, comprising a word-meaning pathway from auditory cortex to inferior frontal areas, and an articulatory pathway that extends from auditory to motor regions (Giordano et al., 2017; Giraud & Poeppel, 2012; Gross et al., 2013; Hickok, 2012; Huth, de Heer, Griffiths, Theunissen, & Gallant, 2016). Specifically, a number of anterior-temporal and frontal regions have been implied in implementing a-modal semantic representations (MacSweeney, Capek, Campbell, & Woll, 2008; Ralph, et al., 2017; Simanova, et al., 2012) and in enhancing speech perception in adverse environments, based on the combination of acoustic and visual signals (Giordano, et al., 2017).

Yet, when it comes to representing visual speech signals themselves, our understanding becomes much less clear. That is, we know relatively little about which brain regions mediate speech reading (or lip reading; terms used interchangeably). Previous studies have shown that visual speech activates ventral and dorsal visual pathways and bilateral fronto-temporal circuits (Bernstein & Liebenthal, 2014; Calvert, et al., 1997; Campbell, 2007; Capek, et al., 2008). Some studies have explicitly suggested that auditory regions are also involved in speech reading (Calvert, et al., 1997; Calvert & Campbell, 2003; Capek, et al., 2008; Lee & Noppeney, 2011; Pekkola et al., 2005). While these findings can be seen to suggest that largely the same brain regions represent acoustic and visual speech, neuroimaging studies have left the nature and the functional specificity of these visual speech representations

unclear (Bernstein & Liebenthal, 2014; Crosse, et al., 2015; Ozker, Yoshor, & Beauchamp, 2018). This is in part because most studies focused on mapping activations rather than specific semantic or lexical speech content. Indeed, alternative accounts have been proposed, which hold that visual and auditory speech representations are largely distinct (Bernstein & Liebenthal, 2014; for spoken vs sign language, Evans, Price, Diedrichsen, Gutierrez-Sigut, & MacSweeney, 2019).

When investigating how speech is encoded in the brain, it is important to distinguish purely stimulus driven neural activity (e.g. classic 'activation') from activity specifically representing a stimulus while also mediating the participant's percept, or behavioural choice, on an individual trial (Bouton et al., 2018; Grootswagers, Cichy, & Carlson, 2018; Keitel, Gross, & Kayser, 2018; Panzeri, Harvey, Piasini, Latham, & Fellin, 2017; Tsunada, Liu, Gold, & Cohen, 2016). Indeed, recent studies have suggested that those cerebral representations representing the physical speech may be distinct from those reflecting the actually perceived meaning. For example, syllable identity can be decoded from temporal, occipital and frontal areas, but only focal activity in the IFG and pSTG mediates perceptual categorisation (Bouton, et al., 2018). Similarly, the encoding of the acoustic speech envelope is seen widespread in the brain, but correct word comprehension correlates only with focal activity in temporal and motor regions (Keitel, et al., 2018). In general, activity in lower sensory pathways seems to correlate more with the actual physical stimulus, while activity in specific higher-tier regions correlates with the subjective percept (Crochet, Lee, & Petersen, 2018; Romo, Lemus, & de Lafuente, 2012). However, this differentiation poses a challenge for data analysis, and studies on sensory perception are only beginning to address this systematically (Grootswagers, et al., 2018; Panzeri, et al., 2017; Ritchie, Tovar, & Carlson, 2015).

We here capitalise on this functional differentiation of cerebral speech representations linked to the physical stimulus or the actual percept, to identify comprehension-relevant encoding of auditory and visual word identity in the human brain. That is, we ask where and to what degree comprehension-relevant representations of auditory and visual speech overlap. To this end, we exploit a paradigm in which participants performed a comprehension task based on individual sentences that were presented either acoustically or visually (lip reading), while brain activity was recorded using MEG (Keitel, et al., 2018). We then extract single trial word representations and, apply multivariate classification analysis geared to quantify i) where brain activity correctly encodes the actual stimulus, and ii) where the strength of the cerebral representation of word identity is predictive of the participant's comprehension.

Results

Behavioural performance

On each trial participants viewed or listened to visual or acoustically presented sentences (presented in blocks), and performed a comprehension task (4-alternative forced choice) on a specific target word. Acoustic sentences were presented mixed with background noise, to equalise performance between visual and auditory trials. On average, participants perceived the correct target word in approximately 70% of trials across auditory and visual conditions (chance level was 25%). The behavioural performance did not differ significantly between these conditions ($M_{auditory}$ = 69.7%, SD = 7.1%, M_{visual} = 71.7%, SD = 20.0%; $t(19) = -0.42$, $p = 0.68$; **Figure 1**), demonstrating that the addition of acoustic background noise indeed equalised performance between conditions. Still, the between-subject variability in performance was larger in the visual condition (between 31.7% and 98.3%), in line with the notion that lip reading abilities vary extremely across individuals (Bernstein & Liebenthal, 2014; Summerfield, 1992; Tye-Murray, Hale, Spehar, Myerson, & Sommers, 2014). An F-test confirmed that the variance between the auditory and visual condition differed significantly ($F(17,17) = 0.13$, $p < .00001$). Due to the near ceiling performance (above 95% correct), the data from three participants in the visual condition had to be excluded from the neuro-behavioural analysis. Participants also performed the task with auditory and visual stimuli presented at the same time (audiovisual condition), but as performance in this condition was near ceiling, we present the corresponding data only in the supplementary material (**Suppl. Figure 1**).

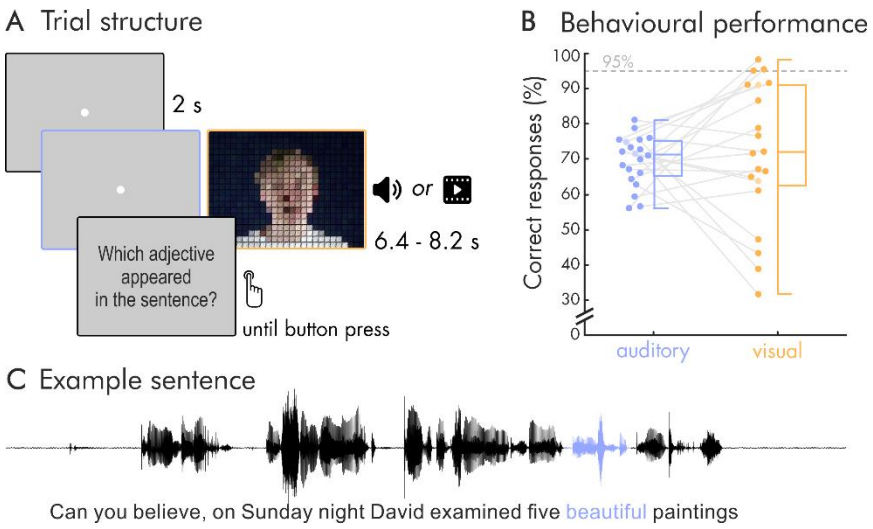


Figure 1. Trial structure and behavioural performance. **A)** Trial structure was identical in the auditory and visual conditions. Participants listened to sentences while a fixation dot was presented (auditory condition) or watched videos of a speaker saying sentences (visual condition). The face of the speaker is obscured for this figure only, it was clear to participants. After each trial, a prompt on the screen asked which adjective (or number) appeared in the sentence and participants chose one of four alternatives by pressing a corresponding button. **B)** Dots represent individual participants, boxes denote median and interquartile ranges, whiskers denote minima and maxima (no outliers present) for all 20 participants. MEG data of two participants (shaded in a lighter colour) were not included in neural analyses due to excessive artifacts. Subjects exceeding a performance of 95% correct (grey line) were excluded from the neuro-behavioural analysis (for the visual condition, three participants had a performance above 95% correct). **C)** Example sentence with target adjective marked in blue.

Decoding word identity from MEG source activity

Using multivariate classification, we quantified how well the single-trial word identity could be correctly predicted from source-localised brain activity. Classification was computed in source space at the single-subject level and converted to *z*-scores for group-level analysis. Importantly, for each trial we computed classification performance within the subset of the four presented alternative words in each trial, based on which participants performed their behavioural judgement. We did this to be able to directly link neural representations of word identity with perception in a later analysis. We first quantified how well brain activity encoded the word identity regardless of behaviour ('stimulus-classification'; c.f. **Materials and Methods**). The group-level analysis (*t*-test, two-sided, FDR-corrected) revealed significant stimulus classification performance in both conditions within a widespread network of temporal, occipital and frontal regions (**Figure 2**).

Auditory speech was represented bilaterally in fronto-temporal areas, extending into intra-parietal regions within the left hemisphere (**Figure 2A**), with classification performance ranging from 25.3% to 29.2% (with a chance level of 25%). Visual speech was represented bilaterally in occipital areas, as well as in left parietal and frontal areas (**Figure 2B**), with classification performance between 25.1% and 34.3%. Interestingly, the regions representing word identity in visual and auditory conditions overlapped only little (mostly in left intraparietal regions; **Figure 2C**; overlap in green). This suggests that largely distinct regions represent visual and acoustic speech, in line with the notion that auditory and visual speech signals are reflected most strongly within the respective sensory cortices (Hauswald, Keitel, Roesch, & Weisz, 2019; Keitel, et al., 2018). Results for the audiovisual condition essentially mirror these unimodal findings and exhibit significant stimulus classification in bilateral temporal and occipital regions (**Suppl. Figure 1B**).

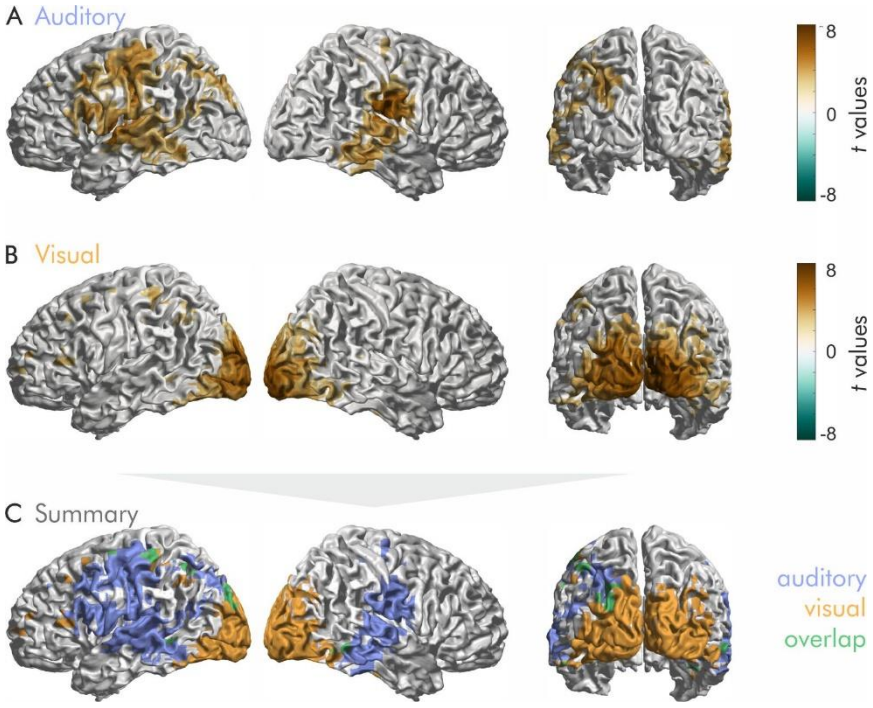


Figure 2. Word classification performance regardless of behavioural performance ('stimulus classification'). Surface projections show areas with significant classification performance at the group level (surface projection of the *t*-statistics, $p < 0.05$, two-sided, FDR corrected). Results show strongest classification performance in temporal regions for the auditory condition **(A)** and occipital areas for the visual condition **(B)**. Panel **(C)** overlays the significant effects from both conditions, with the overlaps shown in green.

Cerebral speech representations that are predictive of comprehension

The above analysis leaves it unclear which of the neural representations are perceptually relevant and shape single-trial word comprehension. To directly address this, we computed an index of how strongly the evidence for a specific word identity in the neural single-trial word representations is predictive of the participant's response. That is, we regressed the evidence in the cerebral classifier for word identity against the participants' behaviour (see **Materials and Methods**). The resulting neuro-behavioural weights (regression *betas*) were converted into *t*-values for group-level analysis. The results in **Figure 3** (two-sided cluster-based permutation statistics, corrected at $p = 0.05$ FWE) reveal largely distinct regions in which neural representations of word identity are predictive of behaviour. In the auditory condition, we found a large left-lateralised cluster covering ventral portions of occipital, temporal, and inferior frontal areas ($T_{sum} = 868.32$, $p < .001$), and a cluster in the right inferior parietal cortex ($T_{sum} = 157.46$, $p < .001$; **Figure 3A**). In the visual condition, we found three dorsal clusters in the left superior frontal gyrus ($T_{sum} = 201.54$, $p < .001$), the inferior frontal gyrus ($T_{sum} = 379.18$, $p < .001$), and premotor cortex ($T_{sum} = 23.55$, $p < .001$), and one cluster in the right supramarginal cortex ($T_{sum} = 167.93$, $p < .001$; **Figure 3B**). MNI coordinates of local maxima and the corresponding *beta* and *t*-values are given in **Table 1**. The corresponding results for the audiovisual condition are presented in **Suppl. Figure 1C**.

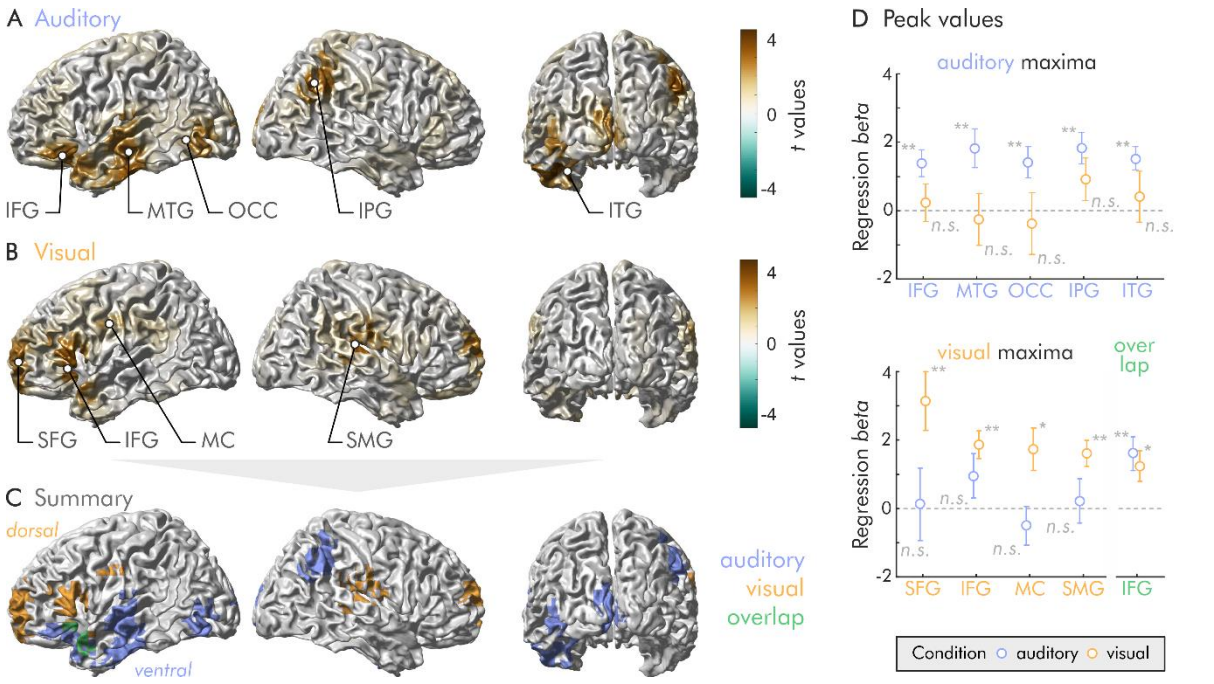


Figure 3. Cortical areas in which neural word representations predict participants' percept. Coloured areas denote significant group-level effects (surface projection of the cluster-based permutation statistics, corrected at p<0.05 FWE). In the auditory condition **(A)**, we found a large left-lateralised ventral cluster (a global peak in ITG and three local maxima marked with dots), as well as a smaller cluster in inferior parietal cortex (peak marked with dot). In the visual condition **(B)**, we found three clusters in left frontal and somato-motor cortex, as well as one cluster in right supramarginal cortex (all peaks are marked with dots). Panel **(C)** overlays the significant effects from both conditions, with the overlap shown in green. The overlap comprises regions in the left inferior frontal gyrus and temporal pole. **D)** Neuro-behavioural effect (at local and global maxima, and maximum of overlap). Regions that predict auditory word perception do not predict visual word perception, and vice versa. Asterisks indicate results of statistical t-test against zero (**: p < .01, *: p < .05, n.s.: p > .05; all *p*-values FDR corrected).
IFG – inferior frontal gyrus; MTG – middle temporal gyrus; OCC – occipital gyrus; IPG – inferior parietal gyrus; ITG – inferior temporal gyrus; SFG – superior frontal gyrus; MC – motor cortex; SMG – supramarginal gyrus.

Table 1. Global and local maxima of neuro-behavioural analysis in both conditions. Labels are taken from the AAL atlas (Tzourio-Mazoyer et al., 2002). For each peak, MNI coordinates, regression *beta* (SEM) and corresponding *t*-value are presented. Abbreviations as used in **Figure 3** are given in parentheses. Global maxima noted in italics. For the peak within the significant overlap of auditory and visual conditions, averaged (across both conditions) *beta* and *t*-values are given.

Atlas label	MNI coordinates	*Beta* (SEM)	*t*-value
Auditory			
Frontal Inf Orb L (IFG)	-33 30 -14	1.38 (0.39)	3.53
Temporal Mid L (MTG)	-64 -26 -14	1.82 (0.57)	3.12
Occipital Mid L, Occipital Inf L, Temporal Mid L (OCC)	-40 -67 -2	1.41 (0.45)	3.16
Parietal Inf R, Angular R (IPG)	51 -60 40	1.82 (0.47)	3.92
Temporal Inf L (ITG)	-35 -25 -29	1.52 (0.34)	4.47
Visual			
Frontal Sup Medial L (SFG)	-9 53 10	2.63 (0.76)	3.69
Frontal Inf Tri L (IFG)	-56 23 -1	1.68 (0.41)	4.71
Postcentral L (MC)	-62 - 8 35	1.44 (0.54)	2.80
Rolandic Oper R, Heschl R (SMG)	43 -26 21	1.41 (0.33)	4.18
Overlap			
Frontal Inf Orb L, Temporal Pole Sup L	-46 18 -15	1.44 (0.34)	2.99

Collectively, these results highlight that in the left hemisphere, perception-relevant representations of acoustic speech reside mainly in ventral regions, whereas those for visual speech are found mostly in dorsal frontal areas (**Figure 3C**). In the right hemisphere, auditory speech representations in parietal regions and visual speech representations in auditory (supramarginal) regions are also predictive of perception. In large, these auditory and visual representations seem distinct, but overlap within higher-order language areas, such as the left inferior frontal and anterior superior temporal gyri.

Given that individual effects where sometimes only significant in one hemisphere, we performed a direct statistical test on whether these effects are indeed lateralised (c.f. **Materials and Methods**). We only found evidence for a statistically significant lateralisation for the large ventral cluster in the auditory condition ($t(17) = 2.88$, $p_{FDR} = .02$, corresponding to local maxima in IFG, MTG, OCC and ITG). In the other clusters, corresponding *betas* in the contralateral hemisphere were systematically smaller, but did not differ significantly from original effects (all $p_{FDR} \geq .15$).

To substantiate that perception-relevant auditory and visual representations are largely distinct, we performed two control analyses. First, we tested whether the representations identified as relevant for visual (auditory) speech are also predictive of perception in the respective other condition. That is, we directly compared the perceptual-relevance for visual and auditory speech representations for those significant clusters shown in **Figure 3A,B**. The result, **Figure 3D**, shows that each region predicts perception only within one modality, with the exception of the overlap in the left IFG.

Second, we implemented a cross-decoding analysis, in which we directly quantified whether the activity patterns of local speech representations are the same across modalities. At the whole-brain level, we found no evidence for significant cross-classification (at $p = 0.05$, FDR corrected, **Figure 4A**), although statistically significant cross-classification is in principle possible from the data, as shown by the audiovisual condition (**Suppl. Figure 1D**).

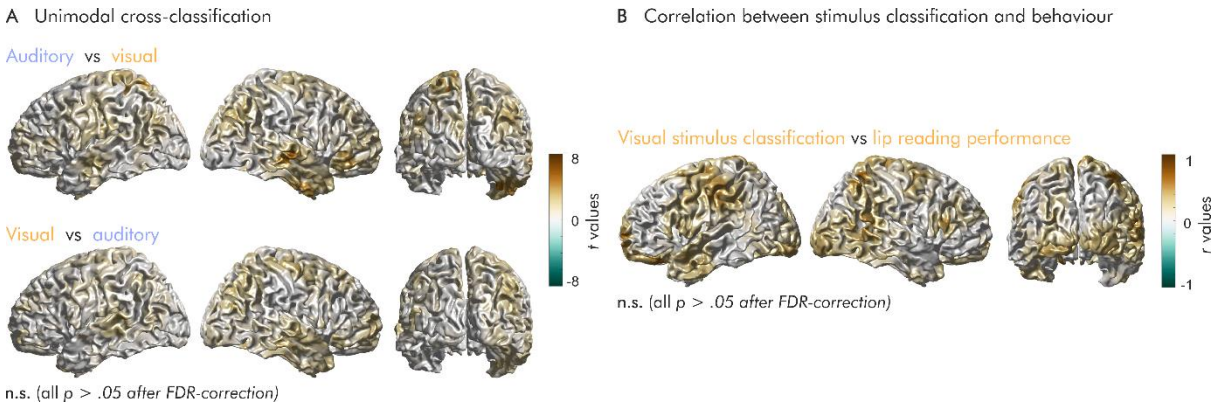


Figure 4. Control analyses. **A)** Cross-classification between auditory and visual conditions. Areas where word identity in the auditory trial can be predicted based on the word representations obtained from the visual condition (upper panel), and vice versa (lower panel). Classification performance did not survive correction for multiple comparison at an alpha-level of 5%, supporting the result that auditory and visual word identities are largely represented in different networks. Colour scale is adapted from Figure 2, to allow a comparison of results. **B)** Correlation between visual word classification performance and behavioural lip reading performance. Surface projection of resulting rho-values. None of the results survived correction for multiple comparisons at an alpha-level of 5%, supporting the finding that stimulus classification alone does not predict behaviour.

Strong sensory representations do not necessarily predict behaviour

The above results suggest that the brain regions in which sensory representations shape speech comprehension are distinct from those allowing the best prediction of the actual stimulus. In other words, the accuracy by which local activity reflects the physical stimulus is not predictive of its' perceptual impact. To test this formally, we performed within-participant regression analyses between the overall stimulus classification performance and the perceptual weight of each local representation across all grid points. Group-level statistics of the participant-specific *beta* values provided no support for a consistent relationship between these (auditory condition: $b = 1.44 \pm 1.62$ [M ± SEM], $t(17) = 0.90$, $p_{FDR} = .58$; visual condition: $b = 1.59 \pm 2.57$ [M ± SEM], $t(14) = 0.56$, $p_{FDR} = .58$).

Still, this leaves it unclear whether variations in the strength of neural speech representations can explain variations in the behavioural performance differences *between* participants. Such an analysis was feasible only for the visual condition, as participants' performance here reflects their individual lipreading skills, whereas performance in the auditory condition was manipulated to yield around 70% correct responses. We correlated the stimulus classification performance for all grid points with participants' visual performance. Stimulus classification performance was not significantly correlated with lip reading performance across participants (all $p_{FDR} > .94$, **Figure 4B**).

Discussion

Acoustic and visual speech are represented in largely distinct brain regions

The principal finding of this study is that the cerebral representations of unimodal auditory and visual speech signals are spatially dissociated and each dominates within distinct brain regions. This is the case for overall strength of word representations, which are mostly related to the physical stimuli themselves, and it is also the case for those word representations that are directly predictive of the individual's single-trial percept. The inability to cross-classify auditory and visual speech from local brain activity further supports the conclusion that acoustic and visual speech representations are largely distinct. These results provide an explanation for the generally observed finding that auditory or verbal skills and visual lip reading are uncorrelated in normal-hearing adults (Jeffers & Barley, 1980; Mohammed, et al., 2006; Summerfield, 1992). Indeed, it has been suggested that individual differences in lipreading represent something other than normal variation in speech perceptual abilities (Summerfield, 1992). For example, lip reading skills are unrelated to reading abilities in the typical adult population (Arnold & Köpsel, 1996; Mohammed, et al., 2006), although a relationship is sometimes found in deaf or dyslexic children (Arnold & Köpsel, 1996; de Gelder & Vroomen, 1998; Kyle, Campbell, & MacSweeney, 2016). The only language ability that can accurately predict speech reading skills in the typical population seems to be guessing strategies (Lyxell & Ronnberg, 1989; Van Tasell & Hawkins, 1981).

We found that perceptually relevant representations of acoustic and visual speech converge only within small regions in the left temporal pole and inferior frontal cortex. These two regions coincide with the higher-order part of the ventral speech pathway. Thus, our results confirm that these regions represent the a-modal and perceived meaning of words, based on a direct assessment of the cerebral speech representations predictive of single trial comprehension (Ralph, et al., 2017; Simanova, et al., 2012).

Previous imaging studies suggested that silent lipreading engages similar regions of the auditory cortex as acoustic speech (Calvert, et al., 1997; Calvert & Campbell, 2003; Capek, et al., 2008; MacSweeney et al., 2000; Paulesu et al., 2003; Pekkola, et al., 2005), implying a direct route for visual speech into the auditory pathways and an overlap of acoustic and visual speech representations in these regions (Bernstein & Liebenthal, 2014). Studies comparing semantic representations of categories from different modalities (e.g. pictures and words) also found large networks with modality-independent activations (Fairhall & Caramazza, 2013; Shinkareva, Malave, Mason, Mitchell, & Just, 2011; Simanova, et al., 2012). Yet, most studies have focused on mapping activation strength rather than the *word identity* of cerebral speech representations. Hence, it could be that visual speech may activate a large language network in an unspecific manner, without engaging specific semantic or lexical representations, maybe as a result of attentional engagement or feed-back (Balk et al., 2013; Ozker, et al., 2018). Support for this interpretation comes from lip reading studies showing that auditory cortical areas are equally activated by visual words and pseudo-words (Calvert, et al., 1997; Paulesu, et al., 2003). While our results suggest that visual speech is largely represented in occipital and frontal regions, we found that the cerebral encoding of visual speech in right auditory

regions (supramarginal and superior temporal gyrus) is also predictive of participants' percept. We therefore support the notion that auditory temporal regions can also contribute to lip-reading. Importantly though, these regions differ from the ones that contribute to auditory speech comprehension.

Another specific region mediating lip-reading comprehension was the IFG, which we have previously also shown to participate in the visual facilitation of auditory speech-in-noise perception (Giordano, et al., 2017). Behavioural studies have shown that lip-reading drives the improvement of speech perception in noise (MacLeod & Summerfield, 1987), hence suggesting that the representations of visual speech in the IFG revealed here are indeed central for hearing in noisy environments, as suggested previously (Giordano, et al., 2017).. Interestingly, these regions resemble the left-lateralised dorsal pathway activated in deaf signers when seeing signed verbs (Emmorey, McCullough, Mehta, Ponto, & Grabowski, 2011). Still, our study cannot directly address whether these auditory and visual speech representations are the same as those that mediate the multisensory facilitation of speech comprehension in adverse environments (Bishop & Miller, 2009; Giordano, et al., 2017). The analysis of the audiovisual condition suggested that stimulus-related representations can be found in auditory and visual sensory areas, similar to unimodal conditions. The preliminary results from a small sample of participants suggest that right precentral and inferior frontal areas drive speech perception in multisensory conditions, in agreement with our previous work (Giordano, et al., 2017).

Sub-optimally encoding brain areas contribute critically to behaviour

To understand which cerebral representations of sensory information guide behaviour, it is important to dissociate those that mainly correlate with the indicated percept from those that encode sensory information and guide behavioural choice (Grootswagers, et al., 2018; Panzeri, et al., 2017; Pica et al., 2017). Single neuron studies have proposed that only those neurons encoding the specific stimulus optimally are readout and used to drive behaviour by downstream areas (Britten, Newsome, Shadlen, Celebrini, & Movshon, 1996; Pitkow, Liu, Angelaki, DeAngelis, & Pouget, 2015; Purushothaman & Bradley, 2005). However, other studies suggest that "plain" sensory information, and sensory information predictive of choice can be decoupled across neurons (Runyan, Piasini, Panzeri, & Harvey, 2017). On a larger scale, the proportions of neurons correlating with the physical stimulus and those correlating with the subjective percept are also de-correlated, with perceptually-relevant neurons dominating within high-level sensory and frontal regions (Leopold & Logothetis, 1999; Romo, et al., 2012). In general, such a dissociation of sensory and choice-related neural representations necessarily emerges in any paradigm where performance is below ceiling, as those regions most predictive of the participants' choice will not be those best representing the stimulus (de-Wit, Alexander, Ekroll, & Wagemans, 2016; Panzeri, et al., 2017). Theoretically, these different types of neural representations can be dissected by considering the intersection of brain activity predictive of stimulus and choice (Panzeri, et al., 2017). In practice, however, it remains a challenge to elucidate these distinct representations, as stimulus and response may correlate for multiple reasons, including confounding factors (Panzeri, et al., 2017).

We here capitalised on the use of a stimulus-classifier to first pinpoint brain activity carrying relevant word-level information and to then test where the quality of the single trial word representation is predictive of participants' comprehension (Cichy, Kriegeskorte, Jozwik, van den Bosch, & Charest, 2017; Grootswagers, et al., 2018; Ritchie, et al., 2015). This revealed that brain regions allowing for a sub-optimal readout of the actual stimulus are predictive of the perceptual outcome, whereas those areas allowing the best read-out not necessarily predict behaviour, a dissociation emerging in several recent studies on the neural basis underlying perception (Bouton, et al., 2018; Grootswagers, et al., 2018; Hasson, Skipper, Nusbaum, & Small, 2007; Keitel, et al., 2018).

One factor that may shape the behavioural relevance of local sensory representations is the specific task imposed (Hickok & Poeppel, 2007). In studies showing the perceptual relevance of optimally encoding neurons, the tasks were mostly dependent on low-level features (Pitkow, et al., 2015; Tsunada, et al., 2016), while studies pointing to a behavioural relevance of high level regions were relying on high-level information such as semantics or visual object categories (Grootswagers, et al., 2018; Keitel, et al., 2018). One prediction from our results is therefore that if the nature of the task was changed from speech comprehension to an acoustic task, the perceptual relevance of word representations would shift from left anterior regions to strongly word encoding regions in the temporal and supramarginal regions. Similarly, if the task would concern detecting basic kinematic features of the visual lip trajectory, activity within early visual cortices tracking the stimulus dynamics should be more predictive of behavioural performance (Di Russo et al., 2007; Keitel et al., 2019; Keitel, Thut, & Gross, 2017).

Conclusion

Overall, our results suggest that cerebral representations of acoustic and visual speech might be more modality-specific than often assumed, and provide a neural explanation for why acoustic speech comprehension is a poor predictor of lip-reading skills. Our results also suggest that those cerebral speech representations that directly drive comprehension are largely distinct from those best representing the physical stimulus, strengthening the notion that neuroimaging studies need to more specifically quantify the cerebral mechanisms driving single trial behaviour.

Materials & Methods

Part of the dataset analysed in the present study has been used in a previous publication (Keitel, et al., 2018). The data analysis performed here is entirely different from the previous work and includes unpublished data.

Participants and data acquisition

Twenty healthy, native volunteers participated in this study (9 female, age 23.6 ± 5.8 y [*M* ± *SD*]). The sample size was set based on previous recommendations (Bieniek, Bennett, Sekuler, & Rousselet, 2016; Poldrack et al., 2017; Simmons, Nelson, & Simonsohn, 2011). MEG data of two participants had to be excluded due to excessive artefacts. Analysis of MEG data therefore included 18 participants (7 female), whereas the analysis of behavioural data included 20 participants. All participants were right-handed (Edinburgh Handedness Inventory; Oldfield, 1971), had normal hearing (Quick Hearing Check; Koike, Hurst, & Wetmore, 1994), and normal or corrected-to-normal vision. Participants had no self-reported history of neurological or language disorders. All participants provided written informed consent prior to testing and received monetary compensation of £10/h. The experiment was approved by the ethics committee of the College of Science and Engineering, University of Glasgow (approval number 300140078), and conducted in compliance with the Declaration of Helsinki.

MEG was recorded with a 248-magnetometers, whole-head MEG system (MAGNES 3600 WH, 4-D Neuroimaging) at a sampling rate of 1 KHz. Head positions were measured at the beginning and end of each run, using five coils placed on the participants' head. Coil positions were co-digitised with the head-shape (FASTRAK®, Polhemus Inc., VT, USA). Participants sat upright and fixated a fixation point projected centrally on a screen. Visual stimuli were displayed with a DLP projector at 25 frames/second, a resolution of 1280 × 720 pixels, and covered a visual field of 25 × 19 degrees. Sounds were transmitted binaurally through plastic earpieces and 370-cm long plastic tubes connected to a sound pressure transducer and were presented stereophonically at a sampling rate of 22,050 Hz. Stimulus presentation was controlled with Psychophysics toolbox (Brainard, 1997) for MATLAB (The MathWorks, Inc.) on a Linux PC.

Stimuli

Data of two conditions across two experimental sessions were used for the current analysis: an auditory only (A) and visual only (V) condition. Participants also completed a third condition in which the same stimulus material was presented audiovisually. This condition could not be used for the present analysis as participants performed near ceiling level in the behavioural task (correct trials: *M* = 96.5%, SD = 3.4%; see **suppl. Figure 1A** for results). The stimulus material consisted of two equivalent sets of 90 sentences (180 in total) that were spoken by a trained, male, native British actor. Sentences were recorded with a high-performance camcorder (Sony PMW-EX1) and external microphone. The speaker was instructed to speak clearly and naturally. Each sentence had the same linguistic structure (Keitel, et al., 2018). An example is: "*Did you notice* (filler phase), *on Sunday night* (time phrase) *Graham* (name)

offered (verb) *ten* (number) *fantastic* (adjective) *books* (noun)". In total, 18 possible names, verbs, numbers, adjectives, and nouns were each repeated ten times. Sentence elements were re-combined within a set of 90 sentences. As a result, sentences made sense, but no element could be semantically predicted from the previous material. To measure comprehension performance, a target word was selected that was either the adjective in one set of sentences ('fantastic' in the above example) or a three-syllable number in the other set (for example, 'thirty-two'). The duration of sentences ranged from 4.2 s to 6.5 s (5.4 ± 0.4 s [*M* ± *SD*]). Noise/video onset and offset was approximately 1 second before and after the speech, resulting in stimulus lengths of 6.4 s to 8.2 s (**Figure 1**).

The acoustic speech was embedded in noise to match performance between auditory and visual conditions. The noise consisted of ecologically valid, environmental sounds (traffic, car horns, talking), combined into a uniform mixture of 50 different background noises. The individual noise level for each participant was determined with a one-up-three-down staircase procedure that was designed to yield a performance of 70% correct. For the staircase procedure, only the 18 possible target words (i.e. adjectives and numbers) were used instead of whole sentences. Participants were presented with a single target word embedded in noise and had to choose between two alternatives. The average signal-to-noise ratio across participants was approximately -6 dB.

Experimental Design

The 180 sentences were presented in two conditions (A, V), each consisting of four blocks with 45 sentences each. In each block, participants either reported the comprehended adjective or number, resulting in two 'adjective blocks' and two 'number blocks'. The order of sentences and blocks was randomised for each participant. The first trial of each block was a 'dummy' trial that was discarded for subsequent analysis; this trial was repeated at the end of the block.

During the presentation of the sentence, participants fixated either a dot (auditory condition) or a small cross on the speaker's mouth (visual condition; see **Figure 1** for depiction of trial structure). After each sentence, participants were presented with four target words (either adjectives or written numbers) on the screen and had to indicate which one they perceived by pressing one of four buttons on a button box. After 2 seconds, the next trial started automatically. Each block lasted approximately 10 minutes. The two separate sessions were completed within one week.

MEG pre-processing

Pre-processing of MEG data was carried out in MATLAB (The MathWorks, Inc.) using the Fieldtrip toolbox (Oostenveld, Fries, Maris, & Schoffelen, 2011). All experimental blocks were pre-processed separately. Single trials were extracted from continuous data starting 2 sec before sound/video onset and until 10 sec after onset. MEG data were denoised using a reference signal. Known faulty channels (N = 7) were removed before further pre-processing. Trials with SQUID jumps (on average 3.86% of trials) were detected and removed using Fieldtrip procedures with a cutoff z-value of 30. Before further artifact rejection, data were

filtered between 0.2 and 150 Hz (fourth order Butterworth filters, forward and reverse) and down-sampled to 300 Hz. Data were visually inspected to find noisy channels (4.95 ± 5.74 on average across blocks and participants) and trials (0.60 ± 1.24 on average across blocks and participants). There was no indication for a statistical difference between the number of rejected channels or trials between conditions ($p > .48$ for channels, $p > .40$ for trials). Finally, heart and eye movement artifacts were removed by performing an independent component analysis with 30 principal components (2.5 components removed on average). Data were further down-sampled to 150 Hz and bandpass-filtered between 0.8 and 30 Hz (fourth order Butterworth filters, forward and reverse).

Source reconstruction

Source reconstruction was performed using Fieldtrip, SPM8, and the Freesurfer toolbox. We acquired T1-weighted structural magnetic resonance images (MRIs) for each participant. These were co-registered to the MEG coordinate system using a semi-automatic procedure (Gross, et al., 2013; Keitel, Ince, Gross, & Kayser, 2017). MRIs were then segmented and linearly normalised to a template brain (MNI space). A forward solution was computed using a single-shell model (Nolte, 2003). We projected sensor-level timeseries into source space using a frequency-specific linear constraint minimum variance (LCMV) beamformer (Van Veen, van Drongelen, Yuchtman, & Suzuki, 1997) with a regularisation parameter of 7% and optimal dipole orientation (singular value decomposition method). Covariance matrices for source were based on the whole length of trials to make use of the longer signal (Brookes et al., 2008). Grid points had a spacing of 6 mm, resulting in 12,337 points covering the whole brain. For subsequent analyses, we selected grid points that corresponded to cortical regions only (parcellated using the AAL atlas; Tzourio-Mazoyer, et al., 2002). This resulted in 5,131 grid points in total.

Neural timeseries were spatially smoothed (Gross, et al., 2013) and normalised in source space. For this, the bandpass-filtered timeseries for the whole trial (i.e. the whole sentence) were projected into source space and smoothed using SPM8 routines with a Full-Width Half Max value of 3. The timeseries for each cortical grid point and trial was then normalised by computing the *z*-score.

Decoding analysis

We used multi-variate single trial classification to localise cerebral representations of the target word in source activity (Grootswagers, Wardle, & Carlson, 2017; Guggenmos, Sterzer, & Cichy, 2018). Each target word was presented in ten different trials. We extracted the 500 ms of activity following the onset of each target word and re-binned the source activity at 20 ms resolution. Classification was performed on spatial searchlights of 1.5 cm radius. We initially tested a number of different classifiers, including linear-discriminant and diagonal-linear classifiers, and then selected a correlation-based nearest-neighbour classifier as this performed slightly better than the others. This (leave-one-trial-out) classifier computed, for a given trial, the Pearson correlation of the spatio-temporal searchlight activity in this test-trial with the activities for the same words in all nine other trials (within-target distances), and with

the activities of the ten repeats of the three other words offered as alternative words on this test trial to the participant (between-word distances). That is, each trial was classified within the sub-set of words that was available to the participant as potential behavioural choices. We then averaged correlations within the four candidate words and decoded the target trial as the word identity with the strongest average correlation (that is, smallest classifier distance). This classification measure is comparable to previous studies probing how well speech can be discriminated based on patterns of dynamic brain activity (Luo & Poeppel, 2007; Rimmele, Zion Golumbic, Schroger, & Poeppel, 2015).

To quantify the degree to which the evidence of local speech representations in favour of a specific word identity is predictive of comprehension, we extracted an index of how well the classifier separated the correct word identity from the three alternatives (Cichy, et al., 2017; Grootswagers, et al., 2018; Ritchie, et al., 2015). This representational distance was defined as the average correlation with trials of the same (correct) word identity and the mean of the correlation with the three alternatives. If a local cerebral representation allows a clear and robust classification of a specific word identity, this representational distance would be large, while if a representation allows only for poor classification, or mis-classifies a trial, this distance will be small or negative. For cross-condition classification (**Figure 4A**), we classified the single trial activity from the auditory (visual) condition against all trials with the same word alternatives from the other condition, or from the audiovisual condition.

Quantifying the behavioural relevance of speech representations

To determine the degree to which local speech representations are predictive of the individual percept, that is the participant's choice on each trial, we quantified the statistical relation between subjects performance (accuracy) and the single trial representational distances (Cichy, et al., 2017; Grootswagers, et al., 2018; Panzeri, et al., 2017; Pica, et al., 2017; Ritchie, et al., 2015). This analysis was based on a regularised logistic regression (Parra, Spence, Gerson, & Sajda, 2005), which was computed across all trials per participant. To avoid biasing, the regression model was computed across randomly selected subsets of trials with equal numbers of correct and wrong responses, averaging betas across 50 randomly selected trials. The resulting *beta* values were then entered into a group-level analysis.

Statistical analyses

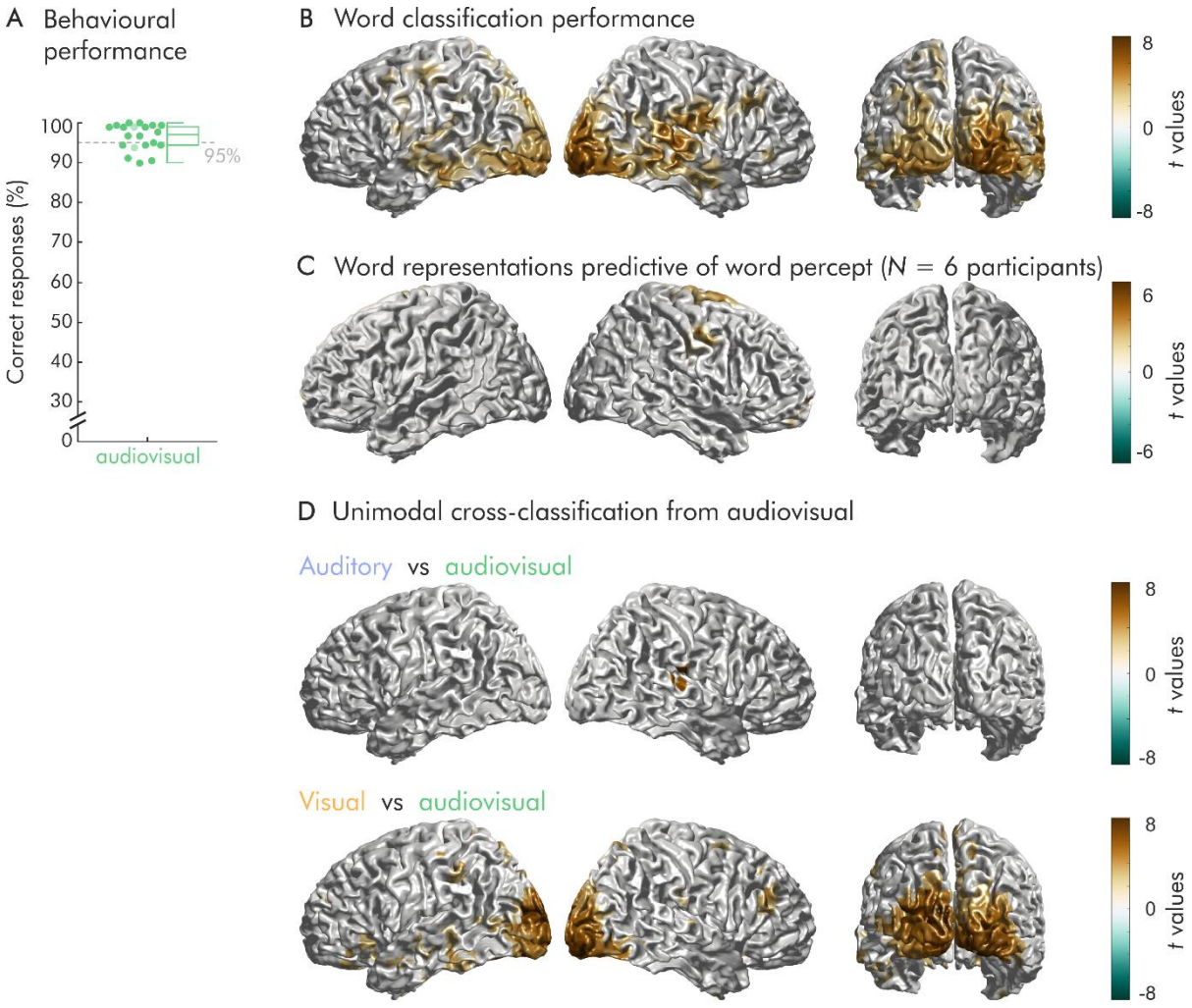
To test the overall stimulus classification performance, we transformed the performance per grid point into *z*-values relative to a surrogate distribution obtained from 2000 within-subject permutations trial labels (i.e. mean and standard deviation of this normally distributed variable were used for the *z*-transformation). These *z*-values were tested against zero, using a two-sided, dependent *t*-test. Resulting p-values were corrected for multiple comparisons by controlling the false discovery rate (FDR) at $p \leq 0.05$, using the Benjamini-Hochberg procedure (Benjamini & Hochberg, 1995).

For the neuro-behavioural analyses, the regression *betas* obtained from the logistic regression were transformed into group-level *t*-values. These were compared with a surrogate distribution

of *t*-values obtained from 1000 within-subject permutations using shuffled trial labels. Results of the two-sided, dependent *t* test were corrected for multiple comparisons with cluster-based permutations (Maris & Oostenveld, 2007), corrected at $p = 0.05$ family-wise error (FWE). Significant clusters were identified based on a first-level significance two-tailed critical *t*-value of $t = 2.1$ for the 18 participants in the auditory condition and $t = 2.2$ for 15 participants in the visual condition. Clusters were selected based on a minimal cluster size of 10. We report the summed *t*-values (T_{sum}) as measure of effect size.

Resulting clusters of the neuro-behavioural analysis were tested for lateralisation (Liegeois et al., 2002). For this, we extracted the participant-specific regression *betas* for each cluster and for the corresponding contralateral grid points. Betas were averaged within each cluster and the between-hemispheric difference was computed using a group-level, two-sided *t*-test. Resulting *p*-values were corrected for multiple comparisons by controlling the FDR at $p \leq 0.05$ (Benjamini & Hochberg, 1995). We only use the term “lateralised” if the between-hemispheric difference is statistically significant.

Supplementary Figure



Suppl. Figure 1. Results of the audiovisual condition. **A)** Behavioural performance of 20 participants. Scaling of the figure is identical to the auditory and visual results for better comparability. Dots represent individual participants, boxes denote median and interquartile ranges, whiskers denote minimum and maximum (no outliers present). MEG data of two participants (shaded in a lighter colour) were not included in neural analyses due to excessive artifacts. Subjects exceeding a performance of 95% correct (grey line) were excluded from the neuro-behavioural analysis (for the audiovisual condition, twelve participants had a performance above 95% correct). **B)** Word classification performance in the audiovisual condition. Surface projections show areas with significant classification performance at the group level (surface projection of the *t*-statistics, $p < 0.05$, two-sided, FDR corrected). Results show strongest classification performance in right auditory and bilateral visual sensory areas, and a classification performance ranging from 25.03% to 33.3% (with a chance level of 25%). **C)** Cortical areas in which neural word representations predict participants' audiovisual percept. Coloured areas denote significant group-level effects (surface projection of the cluster-based permutation statistics, corrected at p=0.05 FWE). Three positive right-lateralised clusters emerged: two in fronto-central regions (superior cluster: $T_{sum} = 260.13$, $p < .001$; inferior cluster: $T_{sum} = 59.15$, $p < .001$), and one in the orbito-frontal region ($T_{sum} = 63.42$, $p < .001$). **D)** Areas where word identity in the auditory (upper panel) or visual (lower panel) conditions can be predicted significantly based on word representations obtained from the audiovisual condition. Auditory word identities can be significantly classified from audiovisual word representations in a small region in right temporal and supramarginal gyrus. Visual word identities can be classified from audiovisual word presentations mainly in bilateral occipital cortex.

References

Arnold, P., & Köpsel, A. (1996). Lipreading, reading and memory of hearing and hearing-impaired children. *Scandinavian Audiology, 25*(1), 13-20.

Balk, M. H., Kari, H., Kauramäki, J., Ahveninen, J., Sams, M., Autti, T., & Jääskeläinen, I. P. (2013). Silent lipreading and covert speech production suppress processing of non-linguistic sounds in auditory cortex. *Open journal of neuroscience, 3*.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society Series B-Methodological, 57*(1), 289-300.

Bernstein, L. E., & Liebenthal, E. (2014). Neural pathways for visual speech perception. *Frontiers in Neuroscience, 8*. doi: 10.3389/Fnins.2014.00386

Bieniek, M. M., Bennett, P. J., Sekuler, A. B., & Rousselet, G. A. (2016). A robust and representative lower bound on object processing speed in humans. *European Journal of Neuroscience, 44*(2), 1804-1814.

Bishop, C. W., & Miller, L. M. (2009). A multisensory cortical network for understanding speech in noise. *Journal of cognitive neuroscience, 21*(9), 1790-1804.

Bouton, S., Chambon, V., Tyrand, R., Guggisberg, A. G., Seeck, M., Karkar, S., . . . Giraud, A. L. (2018). Focal versus distributed temporal cortex activity for speech sound category assignment. *Proc Natl Acad Sci U S A*. doi: 10.1073/pnas.1714279115

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spat Vis, 10*(4), 433-436.

Britten, K. H., Newsome, W. T., Shadlen, M. N., Celebrini, S., & Movshon, J. A. (1996). A relationship between behavioral choice and the visual responses of neurons in macaque MT. *Visual neuroscience, 13*(1), 87-100.

Brookes, M. J., Vrba, J., Robinson, S. E., Stevenson, C. M., Peters, A. M., Barnes, G. R., . . . Morris, P. G. (2008). Optimising experimental design for MEG beamformer imaging. *Neuroimage, 39*(4), 1788-1802.

Calvert, G. A., Bullmore, E. T., Brammer, M. J., Campbell, R., Williams, S. C., McGuire, P. K., . . . David, A. S. (1997). Activation of auditory cortex during silent lipreading. *science, 276*(5312), 593-596.

Calvert, G. A., & Campbell, R. (2003). Reading speech from still and moving faces: the neural substrates of visible speech. *Journal of cognitive neuroscience, 15*(1), 57-70.

Campbell, R. (2007). The processing of audio-visual speech: empirical and neural bases. *Philosophical Transactions of the Royal Society B: Biological Sciences, 363*(1493), 1001-1010.

Capek, C. M., MacSweeney, M., Woll, B., Waters, D., McGuire, P. K., David, A. S., . . . Campbell, R. (2008). Cortical circuits for silent speechreading in deaf and hearing people. *Neuropsychologia, 46*(5), 1233-1241. doi: 10.1016/j.neuropsychologia.2007.11.026

Cichy, R. M., Kriegeskorte, N., Jozwik, K. M., van den Bosch, J. J. F., & Charest, I. (2017). Neural dynamics of real-world object vision that guide behaviour. *bioRxiv*, 147298. doi: 10.1101/147298

Conrad, R. (1977). Lip-reading by deaf and hearing children. *British Journal of Educational Psychology, 47*(1), 60-65.

Crochet, S., Lee, S.-H., & Petersen, C. C. (2018). Neural Circuits for Goal-Directed Sensorimotor Transformations. *Trends in neurosciences*.

Crosse, M. J., ElShafei, H. A., Foxe, J. J., & Lalor, E. C. (2015). *Investigating the temporal dynamics of auditory cortical activation to silent lipreading.* Paper presented at the 2015 7th International IEEE/EMBS Conference on Neural Engineering (NER).

de-Wit, L., Alexander, D., Ekroll, V., & Wagemans, J. (2016). Is neuroimaging measuring information in the brain? *Psychonomic bulletin & review, 23*(5), 1415-1428.

de Gelder, B., & Vroomen, J. (1998). Impaired speech perception in poor readers: Evidence from hearing and speech reading. *Brain and Language, 64*(3), 269-281.

Di Russo, F., Pitzalis, S., Aprile, T., Spitoni, G., Patria, F., Stella, A., . . . Hillyard, S. A. (2007). Spatiotemporal analysis of the cortical sources of the steady-state visual evoked potential. *Human brain mapping, 28*(4), 323-334.

Emmorey, K., McCullough, S., Mehta, S., Ponto, L. L., & Grabowski, T. J. (2011). Sign language and pantomime production differentially engage frontal and parietal cortices. *Language and cognitive processes, 26*(7), 878-901.

Evans, S., Price, C. J., Diedrichsen, J., Gutierrez-Sigut, E., & MacSweeney, M. (2019). Evidence for shared conceptual representations for sign and speech. *bioRxiv*, 623645.

Fairhall, S. L., & Caramazza, A. (2013). Brain regions that represent amodal conceptual knowledge. *Journal of Neuroscience, 33*(25), 10552-10558.

Giordano, B. L., Ince, R. A. A., Gross, J., Schyns, P. G., Panzeri, S., & Kayser, C. (2017). Contributions of local speech encoding and functional connectivity to audio-visual speech perception. *Elife, 6*. doi: 10.7554/eLife.24763

Giraud, A. L., & Poeppel, D. (2012). Cortical oscillations and speech processing: emerging computational principles and operations. *Nat Neurosci, 15*(4), 511-517. doi: 10.1038/nn.3063

Grootswagers, T., Cichy, R. M., & Carlson, T. A. (2018). Finding decodable information that is read out in behaviour. *bioRxiv*, 248583. doi: 10.1101/248583

Grootswagers, T., Wardle, S. G., & Carlson, T. A. (2017). Decoding dynamic brain patterns from evoked responses: A tutorial on multivariate pattern analysis applied to time series neuroimaging data. *Journal of cognitive neuroscience, 29*(4), 677-697.

Gross, J., Hoogenboom, N., Thut, G., Schyns, P., Panzeri, S., Belin, P., & Garrod, S. (2013). Speech rhythms and multiplexed oscillatory sensory coding in the human brain. *PLoS Biol, 11*(12), e1001752. doi: 10.1371/journal.pbio.1001752

Guggenmos, M., Sterzer, P., & Cichy, R. M. (2018). Multivariate pattern analysis for MEG: a comparison of dissimilarity measures. *Neuroimage, 173*, 434-447.

Hall, D. A., Fussell, C., & Summerfield, A. Q. (2005). Reading fluent speech from talking faces: typical brain networks and individual differences. *Journal of cognitive neuroscience, 17*(6), 939-953.

Hasson, U., Skipper, J. I., Nusbaum, H. C., & Small, S. L. (2007). Abstract coding of audiovisual speech: beyond sensory representation. *Neuron, 56*(6), 1116-1126.

Hauswald, A., Keitel, A., Roesch, S., & Weisz, N. (2019). Degraded auditory and visual speech affects theta synchronization and alpha power differently. *bioRxiv*, 615302. doi: 10.1101/615302

Hickok, G. (2012). The cortical organization of speech processing: Feedback control and predictive coding the context of a dual-stream model. *Journal of Communication Disorders, 45*(6), 393-402. doi: 10.1016/j.jcomdis.2012.06.004

Hickok, G., & Poeppel, D. (2007). Opinion - The cortical organization of speech processing. *Nature Reviews Neuroscience, 8*(5), 393-402. doi: 10.1038/nrn2113

Huth, A. G., de Heer, W. A., Griffiths, T. L., Theunissen, F. E., & Gallant, J. L. (2016). Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature, 532*(7600), 453.

Jeffers, J., & Barley, M. (1980). *Speechreading (lipreading)*: Charles C. Thomas Publisher.

Keitel, A., Gross, J., & Kayser, C. (2018). Perceptually relevant speech tracking in auditory and motor cortex reflects distinct linguistic features. *PLoS Biol, 16*(3), e2004473.

Keitel, A., Ince, R. A., Gross, J., & Kayser, C. (2017). Auditory cortical delta-entrainment interacts with oscillatory power in multiple fronto-parietal networks. *Neuroimage, 147*, 32-42. doi: 10.1016/j.neuroimage.2016.11.062

Keitel, C., Keitel, A., Benwell, C. S. Y., Daube, C., Thut, G., & Gross, J. (2019). Stimulus-driven brain rhythms within the alpha band: The attentional-modulation conundrum. *Journal of Neuroscience*, 1633-1618.

Keitel, C., Thut, G., & Gross, J. (2017). Visual cortex responses reflect temporal structure of continuous quasi-rhythmic sensory stimulation. *Neuroimage, 146*, 58-70.

Koike, K. J., Hurst, M. K., & Wetmore, S. J. (1994). Correlation between the American-Academy-of-Otolaryngology-Head-and-Neck-Surgery 5-minute hearing test and standard audiological data. *Otolaryngology-Head and Neck Surgery, 111*(5), 625-632.

Kyle, F. E., Campbell, R., & MacSweeney, M. (2016). The relative contributions of speechreading and vocabulary to deaf and hearing children's reading ability. *Research in developmental disabilities, 48*, 13-24.

Lee, H., & Noppeney, U. (2011). Physical and Perceptual Factors Shape the Neural Mechanisms That Integrate Audiovisual Signals in Speech Comprehension. *The Journal of Neuroscience, 31*(31), 11338-11350. doi: 10.1523/jneurosci.6510-10.2011

Leopold, D. A., & Logothetis, N. K. (1999). Multistable phenomena: changing views in perception. *Trends in cognitive sciences, 3*(7), 254-264.

Liegeois, F., Connelly, A., Salmond, C., Gadian, D., Vargha-Khadem, F., & Baldeweg, T. (2002). A direct test for lateralization of language activation using fMRI: comparison with invasive assessments in children with epilepsy. *Neuroimage, 17*(4), 1861-1867.

Luo, H., & Poeppel, D. (2007). Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex. *Neuron, 54*(6), 1001-1010. doi: 10.1016/j.neuron.2007.06.004

Lyxell, B., & Ronnberg, J. (1989). Information-processing skill and speech-reading. *Br J Audiol, 23*(4), 339-347.

MacLeod, A., & Summerfield, Q. (1987). Quantifying the contribution of vision to speech perception in noise. *British journal of audiology, 21*(2), 131-141.

MacSweeney, M., Amaro, E., Calvert, G. A., Campbell, R., David, A. S., McGuire, P., . . . Brammer, M. J. (2000). Silent speechreading in the absence of scanner noise: an event-related fMRI study. *Neuroreport, 11*(8), 1729-1733.

MacSweeney, M., Capek, C. M., Campbell, R., & Woll, B. (2008). The signing brain: the neurobiology of sign language. *Trends in cognitive sciences, 12*(11), 432-440.

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *J Neurosci Methods, 164*(1), 177-190. doi: 10.1016/j.jneumeth.2007.03.024

Mohammed, T., Campbell, R., Macsweeney, M., Barry, F., & Coleman, M. (2006). Speechreading and its association with reading among deaf, hearing and dyslexic individuals. *Clin Linguist Phon, 20*(7-8), 621-630. doi: 10.1080/02699200500266745

Navarra, J., & Soto-Faraco, S. (2007). Hearing lips in a second language: visual articulatory information enables the perception of second language sounds. *Psychological research, 71*(1), 4-12.

Nolte, G. (2003). The magnetic lead field theorem in the quasi-static approximation and its use for magnetoencephalography forward calculation in realistic volume conductors. *Phys Med Biol, 48*(22), 3637-3652.

Oldfield, R. C. (1971). The assessment and analysis of handedness: the Edinburgh inventory. *Neuropsychologia, 9*(1), 97-113.

Oostenveld, R., Fries, P., Maris, E., & Schoffelen, J. M. (2011). FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Comput Intell Neurosci, 2011*, 156869. doi: 10.1155/2011/156869

Ozker, M., Yoshor, D., & Beauchamp, M. S. (2018). Frontal cortex selects representations of the talker's mouth to aid in speech perception. *eLife, 7*, e30387.

Panzeri, S., Harvey, C. D., Piasini, E., Latham, P. E., & Fellin, T. (2017). Cracking the neural code for sensory perception by combining statistics, intervention, and behavior. *Neuron, 93*(3), 491-507.

Parra, L. C., Spence, C. D., Gerson, A. D., & Sajda, P. (2005). Recipes for the linear analysis of EEG. *Neuroimage, 28*(2), 326-341.

Paulesu, E., Perani, D., Blasi, V., Silani, G., Borghese, N., De Giovanni, U., . . . Fazio, F. (2003). A functional-anatomical model for lipreading. *Journal of neurophysiology, 90*(3), 2005-2013.

Peelle, J. E., & Sommers, M. S. (2015). Prediction and constraint in audiovisual speech perception. *Cortex, 68*, 169-181. doi: 10.1016/j.cortex.2015.03.006

Pekkola, J., Ojanen, V., Autti, T., Jääskeläinen, I. P., Möttönen, R., Tarkiainen, A., & Sams, M. (2005). Primary auditory cortex activation by visual speech: an fMRI study at 3 T. *Neuroreport, 16*(2), 125-128.

Pica, G., Piasini, E., Safaai, H., Runyan, C., Harvey, C., Diamond, M., . . . Panzeri, S. (2017). *Quantifying how much sensory information in a neural code is relevant for behavior.* Paper presented at the Advances in Neural Information Processing Systems.

Pitkow, X., Liu, S., Angelaki, D. E., DeAngelis, G. C., & Pouget, A. (2015). How can single sensory neurons predict behavior? *Neuron, 87*(2), 411-423.

Poldrack, R. A., Baker, C. I., Durnez, J., Gorgolewski, K. J., Matthews, P. M., Munafò, M. R., . . . Yarkoni, T. (2017). Scanning the horizon: towards transparent and reproducible neuroimaging research. *Nature Reviews Neuroscience, 18*(2), 115-126.

Purushothaman, G., & Bradley, D. C. (2005). Neural population code for fine perceptual decisions in area MT. *Nature neuroscience, 8*(1), 99.

Ralph, M. A. L., Jefferies, E., Patterson, K., & Rogers, T. T. (2017). The neural and computational bases of semantic cognition. *Nature Reviews Neuroscience, 18*(1), 42.

Rimmele, J. M., Zion Golumbic, E., Schroger, E., & Poeppel, D. (2015). The effects of selective attention and speech acoustics on neural speech-tracking in a multi-talker scene. *Cortex, 68*, 144-154. doi: 10.1016/j.cortex.2014.12.014

Ritchie, J. B., Tovar, D. A., & Carlson, T. A. (2015). Emerging object representations in the visual system predict reaction times for categorization. *PLoS computational biology, 11*(6), e1004316.

Romo, R., Lemus, L., & de Lafuente, V. (2012). Sense, memory, and decision-making in the somatosensory cortical network. *Current opinion in neurobiology, 22*(6), 914-919.

Runyan, C. A., Piasini, E., Panzeri, S., & Harvey, C. D. (2017). Distinct timescales of population coding across cortex. *Nature, 548*, 92. doi: 10.1038/nature23020

Shinkareva, S. V., Malave, V. L., Mason, R. A., Mitchell, T. M., & Just, M. A. (2011). Commonality of neural representations of words and pictures. *Neuroimage, 54*(3), 2418-2425.

Simanova, I., Hagoort, P., Oostenveld, R., & Van Gerven, M. A. (2012). Modality-independent decoding of semantic information from the human brain. *Cerebral Cortex, 24*(2), 426-434.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological science, 22*(11), 1359-1366.

Sumby, W. H., & Pollack, I. (1954). Visual Contribution to Speech Intelligibility in Noise. *Journal of the Acoustical Society of America, 26*(2), 212-215. doi: Doi 10.1121/1.1907309

Summerfield, Q. (1991). *Visual perception of phonetic gestures.* Paper presented at the Modularity and the motor theory of speech perception: Proceedings of a conference to honor Alvin M. Liberman.

Summerfield, Q. (1992). Lipreading and Audiovisual Speech-Perception. *Philosophical Transactions of the Royal Society of London Series B-Biological Sciences, 335*(1273), 71-78. doi: DOI 10.1098/rstb.1992.0009

Tsunada, J., Liu, A. S., Gold, J. I., & Cohen, Y. E. (2016). Causal contribution of primate auditory cortex to auditory perceptual decision-making. *Nat Neurosci, 19*(1), 135-142. doi: 10.1038/nn.4195

Tye-Murray, N., Hale, S., Spehar, B., Myerson, J., & Sommers, M. S. (2014). Lipreading in school-age children: the roles of age, hearing status, and cognitive ability. *Journal of Speech, Language, and Hearing Research, 57*(2), 556-565.

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., . . . Joliot, M. (2002). Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *Neuroimage, 15*(1), 273-289. doi: 10.1006/nimg.2001.0978

Van Tasell, D. J., & Hawkins, D. B. (1981). Effects of guessing strategy on speechreading test scores. *Am Ann Deaf, 126*(7), 840-844.

771 Van Veen, B. D., van Drongelen, W., Yuchtman, M., & Suzuki, A. (1997). Localization of brain
772 electrical activity via linearly constrained minimum variance spatial filtering. *IEEE Trans*
773 *Biomed Eng, 44*(9), 867-880. doi: 10.1109/10.623056
774 Yi, A., Wong, W., & Eizenman, M. (2013). Gaze patterns and audiovisual speech
775 enhancement. *Journal of Speech, Language, and Hearing Research*.