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

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

1) Psychology, University of Dundee, Scrymgeour Building, Dundee DD1 4HN, UK
2) Institute of Neuroscience and Psychology, University of Glasgow, 62 Hillhead Street, Glasgow G12 8QB, UK
3) Institute for Biomagnetism and Biosignalanalysis, University of Münster, Malmedyweg 15, 48149 Münster, Germany
4) Department for Cognitive Neuroscience, Faculty of Biology, 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. We are grateful to Lea-Maria Schmitt for guiding the semantic distance analysis.

Abstract

Visual speech is an integral part of communication, but it remains unclear whether information carried by lip movements is represented in the same brain regions that mediate acoustic speech comprehension. Our ability to understand acoustic speech seems independent from that to understand visual speech, yet neuroimaging studies suggest that the neural representations largely overlap. Addressing this discrepancy, we tested where the brain represents acoustically and visually conveyed word identities in a full-brain MEG study. Our analyses dissociate cerebral representations that merely reflect the physical stimulus from those that also predict comprehension, and suggest that these overlap only in specific temporal and frontal regions. Moreover, representations predictive of auditory and visual comprehension converge only in angular and inferior frontal regions. These results provide a neural explanation for the behavioural dissociation of acoustic and visual speech comprehension and suggest that cerebral representations encoding word identities may be more modality-specific than often upheld.

Words abstract: 150.

Keywords: speech decoding, visual speech, 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 mouth, lips and tongue in particular provide both redundant and complementary information to acoustic cues (Hall, Fussell, & Summerfield, 2005; Peelle & Sommers, 2015; Plass, Brang, Suzuki, & Grabowecky, 2019; Summerfield, 1992), and can help to enhance speech intelligibility in noisy environments and 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 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). Moreover, large inter-individual differences in lip reading skills contrast with the low variability seen in auditory speech tests (Summerfield, 1992). In contrast to this behavioural dissociation, neuroimaging and neuroanatomical studies have suggested the convergence of acoustic and visual speech information in specific 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; Morillon, Arnal, Schroeder, & Keitel, 2019). 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 lip reading. 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 lip reading (Calvert et al., 1997; Calvert & Campbell, 2003; Capek et al., 2008; Lee & Noppeney, 2011; Pekkola et al., 2005), for example by receiving

signals from visual cortices that can be exploited to establish coarse-grained acoustic representations (Bourguignon, Baart, Kapnoula, & Molinaro, 2020). 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; 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; A. Keitel, J. Gross, & C. Kayser, 2018; Panzeri, Harvey, Piasini, Latham, & Fellin, 2017; Tsunada, Liu, Gold, & Cohen, 2016). Recent neuroimaging studies have suggested that those cerebral representations representing the physical speech are partly 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 inferior frontal gyrus (IFG) and posterior superior temporal gyrus (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 (A. Keitel et al., 2018; S. K. Scott, 2019). 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 exploited 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 (A. Keitel et al., 2018). We then extracted single-trial word representations and applied multivariate classification analysis geared to quantify i) where brain activity correctly encodes the actual word identity, and ii) where the strength of the cerebral representation of word identity is predictive of the participant's comprehension. Note that the term "word identity" in the present study refers to the semantic, as well as the phonological form of a word (see **suppl. Fig 1** for semantic and phonological analysis).

Results

Behavioural performance

On each trial, participants viewed or listened to visually or acoustically presented sentences (presented in blocks), and performed a comprehension task on a specific target word (4-alternative forced-choice identification of word identity). 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 considerably across individuals (Bernstein & Liebenthal, 2014; Summerfield, 1992; Tye-Murray, Hale, Spehar, Myerson, & Sommers, 2014). 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 ($M_{audiovisual}$ = 96.4%, SD = 3.3%) we present the corresponding data only in the supplementary material (**Suppl. Figure 2**).

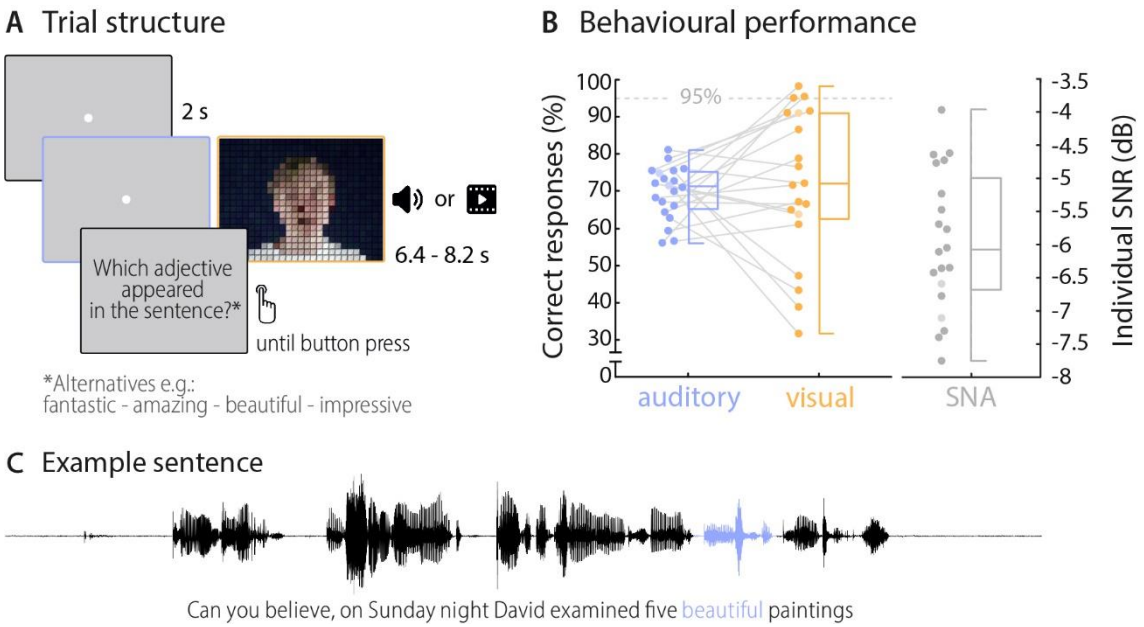


Figure 1. Trial structure and behavioural performance. **A)** Trial structure was identical in the auditory and visual conditions. Participants listened to sentences while a fixation dot was presented (auditory condition) or watched videos of a speaker saying sentences (visual condition). The face of the speaker is obscured for visualisation here only. 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)** Participants' behavioural performance in auditory (blue) and visual (orange) conditions, and their individual SNR values (grey) used for the

auditory condition. 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 artefacts. Participants exceeding a performance of 95% correct (grey line) were excluded from the neuro-behavioural analysis (which was the case for three participants in the visual condition). **C)** Example sentence with target adjective marked in blue.

A representational similarity analysis (RSA) (Evans & Davis, 2015; Kriegeskorte, Mur, & Bandettini, 2008) indicated that participants' behavioural responses were influenced by both semantic and phonological features in both conditions (see **Materials and Methods**, and **suppl. Figure 1**). A repeated- measurements ANOVA yielded a main effect of condition ($F(1,19) = 12.72$, $p = .002$; mean correlations: $M_{auditory} = .35$, SEM = .01; $M_{visual} = .43$, SEM = .02) and a significant interaction ($F(1,19) = 5.96$, $p = .025$). A post-hoc comparison revealed that in the visual condition phonological features influenced behaviour stronger than semantic features (Wilcoxon Signed-ranks test; $Z = 166$, p = .023, no difference in auditory condition, $Z = 91$). While the small number of distinct word identities used here (n = 9) precludes a clear link between these features and the underlying brain activity, these results suggest that participants' lip-reading performance in particular was also driven by non-semantic information.

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. Importantly, for each trial we computed classification performance within the subset of the four presented alternative words in each trial, 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 (two-sided cluster-based permutation statistics, corrected at $p = 0.001$ FWE) revealed significant stimulus classification performance in both conditions within a widespread network of temporal, occipital and frontal regions (**Figure 2**).

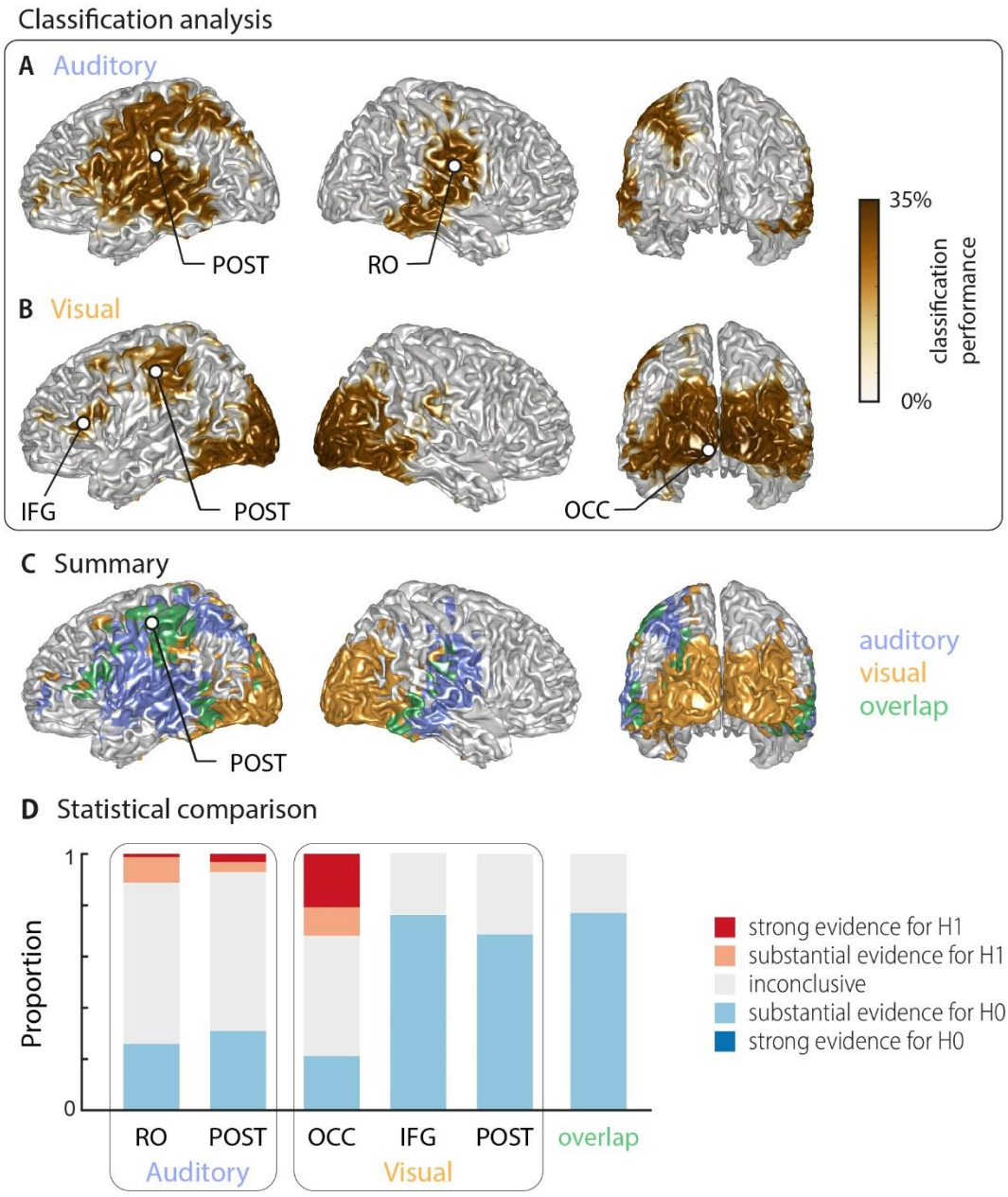


Figure 2. Word classification based on MEG activity regardless of behavioural performance ('stimulus classification'). Surface projections show areas with significant classification performance at the group level (surface projection of the cluster-based permutation statistics, corrected at $p < 0.001$ FWE). Results show strongest classification in temporal regions for the auditory condition **(A)** and occipital areas for the visual condition **(B)**. Cluster peaks are marked with dots. Panel **(C)** overlays the significant effects from both conditions, with the overlap shown in green. The overlap contains bilateral regions in middle and inferior temporal gyrus, the inferior frontal cortex and dorsal regions of the postcentral and supramarginal gyrus (SMG). The peak of the overlap is in the postcentral gyrus. Panel **(D)** shows the distribution of grid-point-wise Bayes factors for a difference between auditory and visual word classification performance in each cluster and the overlap (red: evidence for a difference between conditions, i.e. in favour of H1 [alternative hypothesis]; blue: evidence for no difference between conditions, i.e. in favour of H0 [null hypothesis]).
RO – Rolandic Operculum; POST – postcentral gyrus; IFG – inferior frontal gyrus; OCC – occipital gyrus.

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 25.9% to 29.0% (chance level 25%). Cluster-based permutation statistics yielded two

large clusters: a left-lateralised cluster peaking in inferior postcentral gyrus (left POST; T_{sum} = 230.42, p < .001), and a right-lateralised cluster peaking in the Rolandic operculum (right RO; T_{sum} = 111.17, p < .001). Visual speech was represented bilaterally in occipital areas, as well as in left parietal and frontal areas (**Figure 2B**), with classification performance between 25.9% and 33.9%. There were three clusters: a large bilateral posterior cluster that peaked in the left calcarine gyrus (left OCC; T_{sum} = 321.78, p < .001), a left-hemispheric cluster that peaked in the inferior frontal gyrus (left IFG; T_{sum} = 10.98, p < .001), and a left-hemispheric cluster that peaked in the postcentral gyrus (left POST; T_{sum} = 35.83, p < .001). The regions representing word identity in both visual and auditory conditions overlapped in the middle and inferior temporal gyrus, the postcentral and supramarginal gyri, and the left inferior frontal gyrus (**Figure 2C**; overlap in green). MNI coordinates of cluster peaks and the corresponding classification values are given in **Table 1**. Results for the audio-visual condition essentially mirror the unimodal findings and exhibit significant stimulus classification in bilateral temporal and occipital regions (**Suppl. Figure 2B**).

Table 1. Peak effects of stimulus classification performance based on MEG activity. Labels are taken from the AAL atlas (Tzourio-Mazoyer et al., 2002). For each peak, MNI coordinates, and classification performance (mean and SEM) are presented. Abbreviations as used in **Figure 2** are given in parentheses.

Atlas label	**MNI coordinates**	***Classification %* (SEM)**
Auditory peaks		
Rolandic Oper R *(RO)*	41 -14 20	28.89 (0.78)
Postcentral L *(POST)*	-48 -21 25	29.04 (1.00)
Visual peaks		
Calcarine L *(OCC)*	-5 -101 -7	33.92 (1.53)
Frontal Inf Tri L *(IFG)*	-48 23 1	26.70 (0.83)
Postcentral L *(POST)*	-51 -24 47	26.85 (1.02)
Peak of overlap		
Postcentral L *(POST)*	-47 -15 52	26.50 (0.67)

To directly investigate whether regions differed in their classification performance between visual and auditory conditions, we investigated the Bayesian factor for a direct contrast between conditions (group-level t-test, **Figure 2D**). This revealed no conclusive evidence for many grid points within these clusters ($1/3 < bf_{10} < 3$). However, both auditory clusters and the occipital visual cluster contained grid points with substantial or strong ($bf_{10} > 3$ and $bf_{10} > 10$, respectively) evidence for a significant modality difference. In contrast, the visual postcentral region (POST), the IFG and the overlap region contained many grid points with substantial evidence for no difference between modalities ($1/10 < bf_{10} < 1/3$).

Given that most clusters were found in only one hemisphere, we performed a direct test on whether these effects are indeed lateralised in a statistical sense (c.f. **Materials and Methods**). We found evidence for a statistically significant lateralisation for both auditory clusters (left cluster peaking in POST: $t(17) = 5.15$, pFDR < .001; right cluster peaking in RO: $t(17) = 4.26$, pFDR < .01). In the visual condition, the test for the two left clusters reached marginal significance only in terms of a statistical lateralisation (left cluster peaking in IFG: $t(17) = 2.19$, pFDR = .058; left cluster peaking in POST: $t(17) = 1.87$, pFDR = .078). Note that

the large occipital cluster in the visual condition is bilateral and we therefore did not test this for a lateralisation effect. This suggests that distinct frontal, occipital and temporal regions represent visual and acoustic speech, while inferior temporal and frontal regions, the SMG and dorsal post-central regions reflect word identities in both modalities.

Cerebral speech representations that are predictive of comprehension

The above analysis leaves it unclear which of the neural representations are actually perceptually relevant and drive 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 several distinct regions in which neural representations of word identity are predictive of behaviour. In the auditory condition, we found five distinct clusters. Three were in the left hemisphere, peaking in the left inferior temporal gyrus (left ITG; $T_{sum} = 469.55$, $p < .001$), the inferior frontal gyrus (left IFG; $T_{sum} = 138.70$, $p < .001$), and the middle occipital gyrus (left MOG; $T_{sum} = 58.44$, $p < .001$). In the right hemisphere, the two found clusters were in the supplementary motor area (right SMA; $T_{sum} = 312.48$, $p < .001$) and in the angular gyrus (right AG; $T_{sum} = 68.59$, $p < .001$; **Figure 3A**). In the visual condition, we found four clusters: A left-hemispheric cluster in the inferior frontal gyrus (left IFG; $T_{sum} = 144.15$, $p < .001$) and three clusters with right-hemispheric peaks, in the superior temporal gyrus (right STG; $T_{sum} = 168.68$, $p < .001$), the superior frontal gyrus (right SFG; $T_{sum} = 158.39$, $p < .001$) and the angular gyrus (right AG; $T_{sum} = 37.42$, $p < .001$; **Figure 3B**). MNI coordinates of cluster peaks and the corresponding *beta* and *t*-values are given in **Table 2**. Importantly, these perception-relevant (i.e. predictive) auditory and visual representations did not overlap (**Figure 3C**), although some of them occurred in adjacent regions in the IFG and AG.

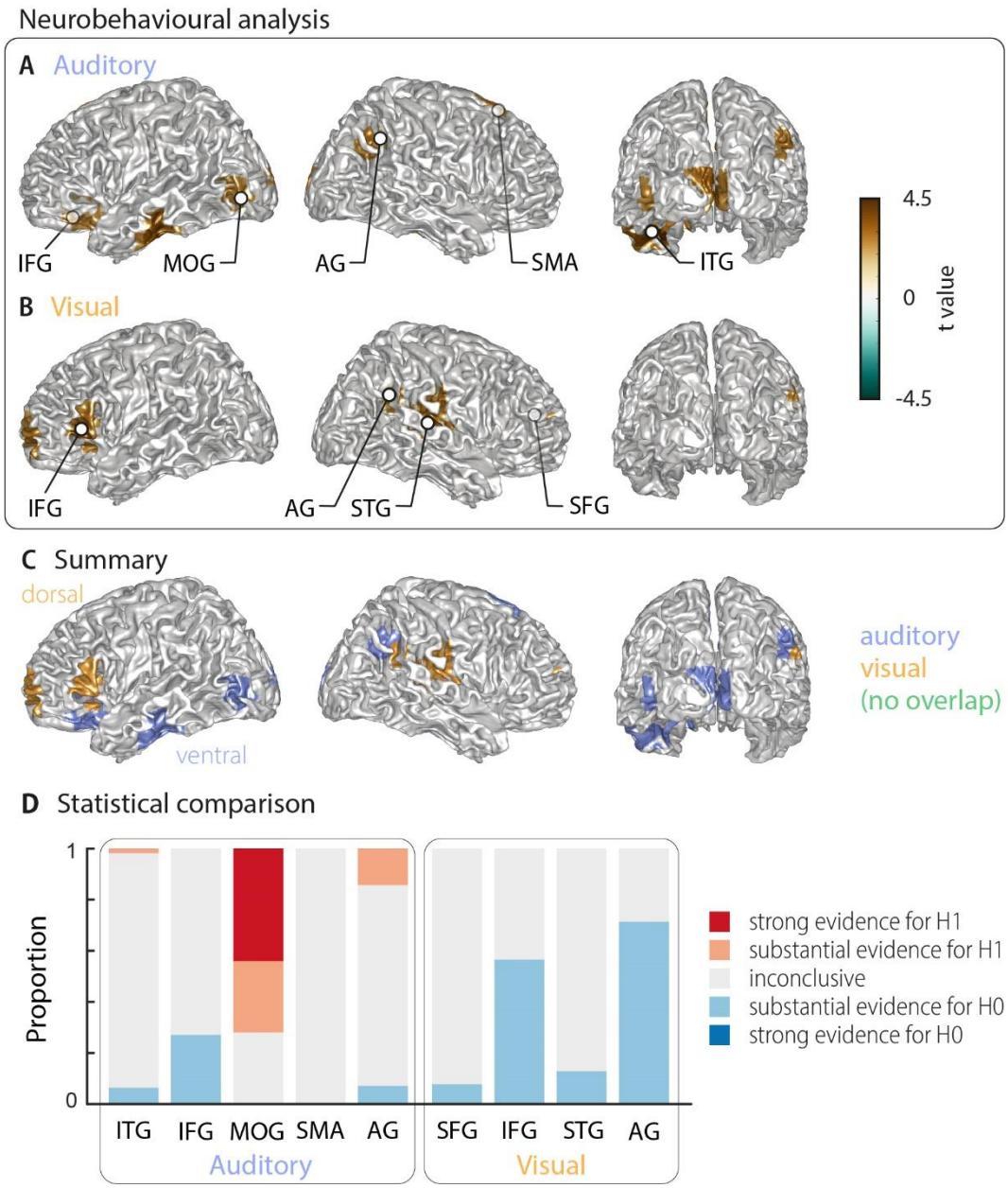


Figure 3. Cortical areas in which neural word representations predict participants' response. Coloured areas denote significant group-level effects (surface projection of the cluster-based permutation statistics, corrected at p<0.05 FWE). In the auditory condition **(A)**, we found five clusters in total (cluster peaks are marked with dots). Three were in left ventral regions, in the inferior frontal gyrus, the inferior temporal gyrus, and the occipital gyrus, the other two were in the right hemisphere, in the angular gyrus and the supplementary motor area. In the visual condition **(B)**, we found four clusters in total: In the left (dorsal) inferior frontal gyrus, the right anterior cingulum stretching to left dorsal frontal regions, in the right angular gyrus and the right superior temporal gyrus (all peaks are marked with dots). Panel **(C)** overlays the significant effects from both conditions. There was no overlap. However, both auditory and visual effects were found in adjacent regions within the left IFG and the right AG. Panel **(D)** shows distributions of grid-point-wise Bayes factors for a difference between auditory and visual conditions for these clusters (red: evidence for differences between conditions, i.e. in favour of H1 [alternative hypothesis]; blue: evidence for no difference between conditions, i.e. in favour of H0 [null hypothesis]).
IFG – inferior frontal gyrus; MOG – middle occipital gyrus; AG – angular gyrus; SMA – supplementary motor area; ITG – inferior temporal gyrus; IFG – inferior frontal gyrus; STG – superior temporal gyrus; SFG – superior frontal gyrus.

Table 2. Peak effects for the neuro-behavioural analysis. Labels are taken from the AAL atlas (Tzourio-Mazoyer et al., 2002). For each local peak, MNI coordinates, regression *beta* (mean and SEM across participants) and corresponding *t*-value are presented. Abbreviations as used in **Figure 3** are given in parentheses.

Atlas label	MNI coordinates			*Beta* (SEM)	*t*-value
Auditory					
Temporal Inf L *(ITG)*	-41	-23	-26	0.106 (0.024)	4.40
Frontal Inf Orb L *(IFG)*	-28	25	-9	0.082 (0.031)	2.66
Occipital Mid L, Occipital Inf L *(MOG)*	-46	-83	-4	0.079 (0.029)	2.75
Supp Motor Area R *(SMA)*	3	11	52	0.089 (0.027)	3.33
Angular R *(AG)*	49	-67	40	0.079 (0.027)	2.87
Visual					
Frontal Inf Tri L *(IFG)*	-57	30	4	0.075 (0.017)	4.34
Frontal Sup Medial R, Cingulum Ant R *(SFG)*	9	47	15	0.080 (0.028)	2.86
Temporal Sup R *(STG)*	38	-30	10	0.086 (0.023)	3.77
Angular R *(AG)*	60	-55	34	0.073 (0.020)	3.55

Again, we asked whether the behavioural relevance of these regions exhibit a significance bias towards either modality by investigating the between-condition contrast (Bayesian factor derived from the group-level *t*-test; **Figure 3D**). Three auditory clusters contained grid points that differed substantially or strongly ($bf_{10} > 3$ and $bf_{10} > 10$, respectively) between modalities (left ITG, left MOG, and right AG). In addition, in two regions the majority of voxels provided substantial evidence for no difference between modalities (IFG and AG from the visual condition). Regarding the lateralisation of these clusters, we found that corresponding *betas* in the contralateral hemisphere were systematically smaller in all clusters but did not differ significantly (all $p_{FDR} \geq .15$), hence providing no evidence for a strict lateralisation.

To further investigate whether perception-relevant auditory and visual representations are largely distinct, we performed 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 (two-sided cluster-based permutation statistics, neither at a corrected $p = 0.001$ nor a more lenient $p = 0.05$; **Figure S3A, left panel**). That significant cross-classification is possible in principle from the data is shown by the significant results for the audio-visual condition (**Figure S3A, right panel**). An analysis of the Bayes factors for this cross-classification test confirmed that most grid points contained substantial evidence for no cross-classification between the auditory and visual conditions (**Figure S3B, left panel**). On the other hand, there was strong evidence for significant cross-classification between the uni- and multisensory conditions in temporal and occipital regions (**Figure S3B, right panel**).

Strong sensory representations do not necessarily predict behaviour

The above results suggest that the brain regions in which sensory representations shape speech comprehension are mostly distinct from those allowing the best prediction of the actual stimulus (see **Figure 4** for on direct visualisation of the results from both analyses). Only parts of the left inferior temporal gyrus (auditory modality), the right superior temporal gyrus (visual

condition) and the left inferior frontal gyrus (both modalities) feature high stimulus classification and a driving role for comprehension. In other words, the accuracy by which local activity reflects the physical stimulus is generally not predictive of the impact of this local word representation on behaviour. To test this formally, we performed within-participant robust 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: $b = 0.05 \pm 0.01$ [M ± SEM], $t(17) = 0.50$, $p_{FDR} = .81$; visual condition: $b = 0.03 \pm 0.01$ [M ± SEM], $t(14) = 0.25$, $p_{FDR} = .81$). A Bayes factor analysis also provided substantial evidence for the H_0, i.e. no consistent relationship ($bf_{10} = 0.27$ and $bf_{10} = 0.27$, for auditory and visual conditions, respectively).

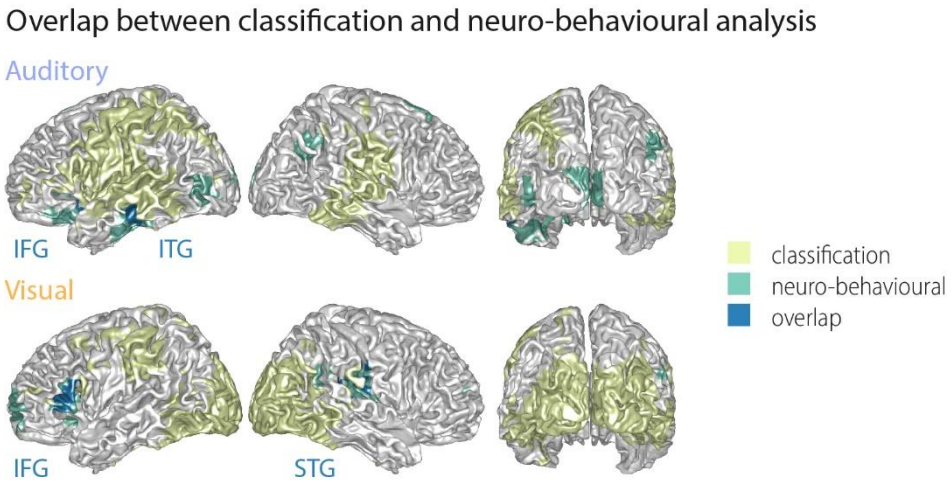


Figure 4. Largely distinct regions provide strong stimulus classification and mediate behavioural relevance. Areas with significant stimulus classification are shown in yellow, those with significant neuro-behavioural results in green, and the overlap in blue. The overlap in the auditory condition ($N = 14$ grid points) comprised parts of the left inferior and middle temporal gyrus (ITG), and the orbital part of the left inferior frontal gyrus (IFG). The overlap in the visual condition ($N = 27$ grid points) comprised the triangular part of the inferior frontal gyrus (IFG), and parts of the superior temporal gyrus (STG), extending dorsally to the Rolandic operculum.

Still, this leaves it unclear whether variations in the strength of neural speech representations (i.e. the 'stimulus classification') can explain variations in the behavioural differences *between* participants. We therefore correlated the stimulus classification performance for all grid points with participants' behavioural data, such as auditory and lip-reading performance, and the individual SNR value. We found no significant clusters (two-sided cluster-based permutation statistics, corrected at $p = 0.05$ FWE), indicating that stimulus classification performance was not significantly correlated with behavioural performance across participants (**suppl. Figure 5**).

Discussion

Acoustic and visual speech are represented in largely distinct brain regions

Our results show that the cerebral representations of auditory and visual speech are mediated by both modality-specific and overlapping (potentially amodal) representations. While several parietal, temporal and frontal regions were engaged in the encoding of both acoustically and visually conveyed word identities ('stimulus classification'), comprehension in both sensory modalities was largely driven by distinct networks. Only the inferior frontal and angular gyrus contained regions that contributed similarly to both auditory and visual comprehension.

This multi-level organisation of auditory and visual speech is supported by several of our findings. First, we found a partial intersection of the sensory information, where significant stimulus classification performance overlapped in bilateral postcentral regions, inferior temporal and frontal regions and the SMG. Second, it is also supported by the observation that angular and inferior frontal regions facilitate both auditory and visual comprehension, while distinct regions additionally support modality-specific aspects of comprehension: middle occipital and parietal representations drive auditory comprehension while superior frontal and temporal regions facilitate lip reading. None of these comprehension-relevant regions was strictly lateralised, in line with the notion that speech comprehension is largely a bilateral process (Kennedy-Higgins, Devlin, Nuttall, & Adank, 2020).

The inability to cross-classify auditory and visual speech from local activity further supports the conclusion that the nature of local representations of acoustic and visual speech is largely distinct. It is important to note that cross-classification probes not only the spatial overlap of two representations but also asks whether the local spatio-temporal activity patterns encoding word identity in the two sensory modalities are the same. It could be that local activity encodes a given word based on acoustic or visual evidence but using distinct activity patterns. Representations could therefore spatially overlap without using the same 'neural code'. Our results hence provide evidence that the activity patterns by which auditory and visual speech are encoded may be partly distinct, even within regions that represent both acoustically and visually mediated word information, such as the inferior frontal and angular gyrus.

The encoding of visual speech

The segregation of comprehension-relevant auditory and visual representations provides a possible explanation for the 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 lip reading 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).

Previous imaging studies suggested that silent lip reading engages similar auditory regions as engaged by acoustic speech (Bourguignon et al., 2020; 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 from different modalities also supported large modality-independent networks (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 encoding of *word identity* by cerebral speech representations. Hence, it could be that visual speech may activate many regions 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), and studies demonstrating cross-modal activations in early sensory regions also for simplistic stimuli (Ferraro et al., 2020; Ibrahim et al., 2016; Petro, Paton, & Muckli, 2017).

Our results suggest that visual speech comprehension is mediated by parietal and inferior frontal regions that likely contribute to both auditory and visual speech comprehension, but also engage superior temporal and superior frontal activity that seems to be modality specific. Thereby our results support a route of visual speech into auditory cortical and temporal regions but provide no evidence for an overlap of speech representations in the temporal lobe that would facilitate both lip-reading and acoustic speech comprehension.

Two specific regions mediating lip-reading comprehension were the IFG and the angular gyrus. Our results suggest that these facilitate both auditory and visual speech comprehension, in line with previous suggestions (Simanova et al., 2012). Previous work has also implicated these regions in the visual facilitation of auditory speech-in-noise perception (Giordano et al., 2017) and lip-reading itself (Bourguignon et al., 2020). 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 these regions may be central for hearing in noisy environments. Interestingly, these regions resemble the left-lateralised dorsal pathway activated in deaf signers when seeing signed verbs (Emmorey, McCullough, Mehta, Ponto, & Grabowski, 2011). Our results 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). Future work needs to directly contrast the degree of which multisensory speech representations overlap locally to the ability of these regions to directly fuse this information.

Cross-modal activations in visual cortex

We also found that acoustic comprehension was related to occipital brain activity (c.f. **Figure 3**). Previous work has shown that salient sounds activate visual cortices (Feng, Störmer,

Martinez, McDonald, & Hillyard, 2014; McDonald, Störmer, Martinez, Feng, & Hillyard, 2013), with top-down projections providing visual regions with semantic information, for example about object categories (Petro et al., 2017; Revina, Petro, & Muckli, 2018). The acoustic speech in the present study was presented in noise, and performing the task hence required attentional effort. Attention may therefore have automatically facilitated the entrance of top-down semantic information into occipital regions that in a multisensory context would encode the lip-movement trajectory, in order to maximise task performance (McDonald et al., 2013). The lack of significant cross-classification performance suggests that the nature of this top-down induced representation differs from that induced by direct lip-movement information.

Sub-optimal sensory representations 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 stimulus from those that encode sensory information and guide behavioural choice. At the single neuron level some studies have proposed that only those neurons encoding the specific stimulus optimally are driving behaviour (Britten, Newsome, Shadlen, Celebrini, & Movshon, 1996; Pitkow, Liu, Angelaki, DeAngelis, & Pouget, 2015; Purushothaman & Bradley, 2005; Tsunada et al., 2016), while others suggest that “plain” sensory information and sensory information predictive of choice can be decoupled across neurons (Runyan, Piasini, Panzeri, & Harvey, 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), that is, the neural representations that are informative about the sensory environment and are used to guide behaviour. While theoretically attractive, this intersection is difficult to quantify for high-dimensional data, in part as direct estimates of this intersection, e.g. based on information-theoretic approaches, are computationally costly (Pica et al., 2017). Hence, in the past most studies, also on speech, have focused on either studying sensory encoding, or behaviourally predictive activity only. However, the former type of cerebral representation may not guide behaviour at all, while the latter may also capture brain activity that drives perceptual errors due to intrinsic fluctuations in sensory pathways, the decision process, or even noise in the motor system (Grootswagers et al., 2018).

To directly quantify where auditory or visual speech is represented and this representation is used to guide comprehension we 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. This dissociation is emerging in several recent studies on the neural basis underlying perception (Bouton et al., 2018; Grootswagers et al., 2018; Hasson, Skipper, Nusbaum, & Small, 2007; A. Keitel et al., 2018). Importantly, it suggests that networks mediating speech comprehension can neither be understood by mapping speech

representations during passive perception nor during task performance, if the analysis itself is not geared towards directly revealing the perception-relevant representations.

On a technical level, it is important to keep in mind that the insights derived from any classification analysis are limited by the quality of the overall classification performance. Classification performance was highly significant and reached about 10% above the respective chance level, a number that is in accordance with other neuroimaging studies on auditory pathways (Bednar, Boland, & Lalor, 2017; Correia, Jansma, Hausfeld, Kikkert, & Bonte, 2015). Yet, more refined classification techniques, or data obtained using significantly larger stimulus sets and more repetitions of individual target words may be able to provide even more refined insights.

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; A. 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; C. Keitel et al., 2019; Tabarelli, Keitel, Gross, & Baldauf, 2020). This suggests that a discussion of the relevant networks underlying speech perception should always be task focused.

Conclusion

These 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 (A Keitel, J Gross, & C Kayser, 2018; A. 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

The experiment featured three conditions: auditory only (A), visual only (V), and a third condition in which the same stimulus material was presented audiovisually (AV). This condition could not be used for the main analyses as participants performed near ceiling level in the behavioural task (correct trials: *M* = 96.5%, SD = 3.4%; see **suppl. Figure 2A** for results). The stimulus material consisted of 180 sentences, based on a set of 9 target words derived from two categories (9 numbers and 9 adjectives), each repeated 10 times in a different sentence. Sentences 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 (A. 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 the number in the other set (for example, 'thirty-two'). All adjective target words had a positive valence (G. G. Scott, Keitel, Becirspahic, Yao, & Sereno, 2019); see **Suppl. Table 1** for all possible target words). 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 (3D1U) staircase procedure that targets the 79.4% probability correct level (Karmali, Chaudhuri, Yi, & Merfeld, 2016). 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. Note that due to the necessary differences between staircase procedure (single words and two-alternative-forced-choice) and behavioural experiment (sentences and four-alternative forced-choice), the performance in the behavioural task was lower than 79.4%. The signal-to-noise ratio across participants ranged from -7.75 dB to -3.97 dB (-5.96 ± 1.06 dB [*M* ± *SD*]; see **Figure 1B**).

Experimental Design

The 180 sentences were presented in three conditions (A, V, AV), 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 (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 cut-off 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 (two-sided t-tests; $p > .48$ 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; A. 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 6,490 grid points in total.

Neural time series were spatially smoothed (Gross et al., 2013) and normalised in source space. For this, the band-pass filtered time series 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 (FWHM) value of 3 mm. The time series for each grid point and trial was then z-scored.

Decoding analysis

We used multi-variate single-trial classification to localise cerebral representations of the target words in source activity (Grootswagers, Wardle, & Carlson, 2017; Guggenmos, Sterzer, &

Cichy, 2018). Each target word was presented in ten different trials per condition. 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 1.2-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 (although we note that the difference in peak classification performance was only on the range of 2-3% between different classifiers). We also probed classification performance based on a number of spatio-temporal searchlights, including temporal binning of the data at 3.3, 20, 40 and 60ms, and including each neighbouring spatial grid point into the searchlight or averaging across grid points. Comparing classification performance revealed that in particular the auditory condition was sensitive to the choice of searchlight. Classification performance dropped when ignoring the spatial configuration or sampling the data at a resolution lower than 20 ms (median performance of the 10% grid points with highest performance based on each searchlight: 27.0%, 26.9% 26.5% and 25.8% for 3.3, 20, 40 and 60-ms bins and the full spatial searchlight; and 26.4% at 20ms and ignoring the spatial pattern). We hence opted for a classifier based on the source data represented as spatial searchlight (1.2-cm radius) and sampled at 20-ms resolution. The typical searchlight contained 31 neighbours (median value), with 95% of searchlights containing 17 to 33 grid points. The (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 word in all other trials (within-target distances), and with the activities of the three alternative words in all trials (between-target distances). That is, each trial was classified within the sub-set of words that was available to the participant as potential behavioural choices (see **suppl. Figure 5** for illustration). 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). Classification performance was averaged across blocks with numbers and adjectives as task-relevant words. For cross-condition classification (suppl. **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 audio-visual condition.

Quantifying the behavioural relevance of speech representations

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 false 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 (within-target distances) word identity and the mean of the correlation with the three alternatives (between-target distances; see **suppl. Figure 5**). 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. We then quantified the statistical relation between participants performance (accuracy) and these 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 averaged across blocks with numbers and adjectives as targets and were entered into a group-level analysis.

Quantifying the role of phonological and semantic features to perception

For each pair of words we computed their phonological distance using the Phonological Corpus Tools (V1.4.0) based on the phonetic string similarity ('phonological edit distance') derived from the transcription tier, using the Irvine Phonotactic Online Dictionary (Vaden, Halpin, & Hickok, 2009). We also computed pairwise semantic distances using the fastTExt vector representation of English words trained on *Common Crawl* and *Wikipedia* obtained online (file cc.en.300.vec) (Grave, Bojanowski, Gupta, Joulin, & Mikolov, 2018). The individual word vectors (300 dimensions) were length-normalised and cosine distances were computed. For each participant, we obtained a behavioural representational dissimilarity matrix (RDM) as the pair-wise behavioural confusion matrix from their behavioural data. We then implemented a representational similarity analysis (RSA) (Kriegeskorte et al., 2008) between phonological (semantic) representations and participants' performance. Specifically, behavioural and semantic (phonetic) RDMs were compared using Spearman's rank correlation. The resulting correlations were *z*-scored and averaged across adjectives and numbers (see **suppl. Figure 1**).

Statistical analyses

To test the overall stimulus classification performance, we averaged the performance per grid point across participants and compared this group-averaged value to a group-average permutation distribution obtained from 2000 within-subject permutations derived with random trial labels. Cluster-based permutation was used to correct for multiple comparisons (Maris & Oostenveld, 2007). Significant clusters were identified based on a first-level significance derived from the 99.95th percentile of the permuted distribution (family-wise error [FWE] of $p = .001$), using the summed statistics (T_{sum}) across grid points within a cluster, and by requiring a minimal cluster size of 10 grid points. The resulting clusters were considered if they reached a *p*-value smaller than 0.05.

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 2000 within-subject permutations using shuffled trial labels and using cluster-based permutations as above. The first-level significance threshold (at $p<0.05$) was determined per condition based on the included sample size (*t*-value of $t = 2.1$ for the 18

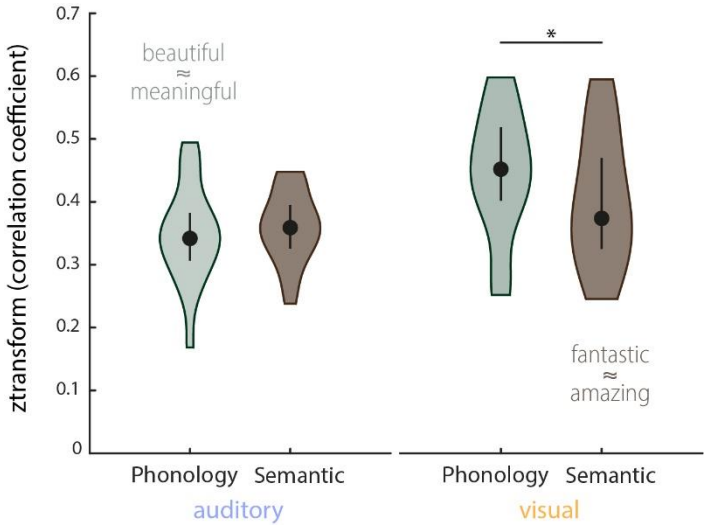
participants in the auditory condition and t = 2.2 for 15 participants in the visual condition), and the resulting clusters were considered significant if they reached a p-value smaller than 0.05.

Resulting clusters were tested for lateralisation (Liegeois et al., 2002; Park & Kayser, 2019). For this, we extracted the participant-specific classification performance (or regression *betas*, respectively) for each cluster and for the corresponding contralateral grid points. These values were averaged within each hemisphere 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.

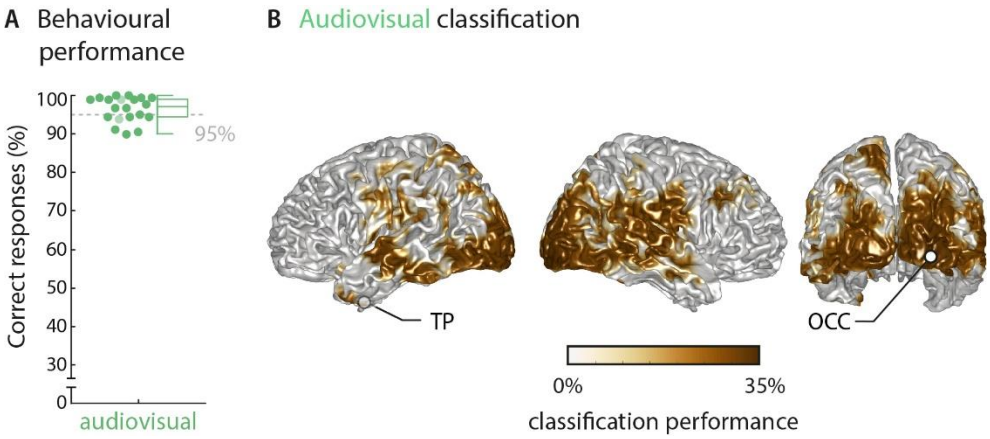
To determine whether individual local effects (e.g. stimulus classification or behavioural prediction) were specific to either condition, we implemented a direct contrast between conditions. For each grid point, we computed a group-level t-test and converted these to a JZS Bayes factor using a default scale factor of 0.707 (Rouder, Speckman, Sun, Morey, & Iverson, 2009). We then quantified the number of grid points per region of interest that exhibited a specific level of evidence in favour of the null hypothesis of no effect versus the alternative hypothesis (H_0 vs H_1) (Jeffreys, 1998). Using previous conventions (Wagenmakers, Wetzels, Borsboom, & Van Der Maas, 2011), the Bayes factor showed evidence for H_1 if it exceeded a value of 3, and evidence for H_0 if it was below ⅓, with the intermediate range yielding inconclusive results. We also calculated Bayes factors from Pearson correlation coefficients (for a control analysis between classification performance and behavioural data), using the same conventions (Wetzels & Wagenmakers, 2012).

To investigate the relationship between stimulus classification and neurobehavioral results, we performed a robust linear regression within each participant for all grid points. The participant-specific *beta* values were then tested against zero using a two-sided t-test (A. Keitel et al., 2017).

Supplementary Figures

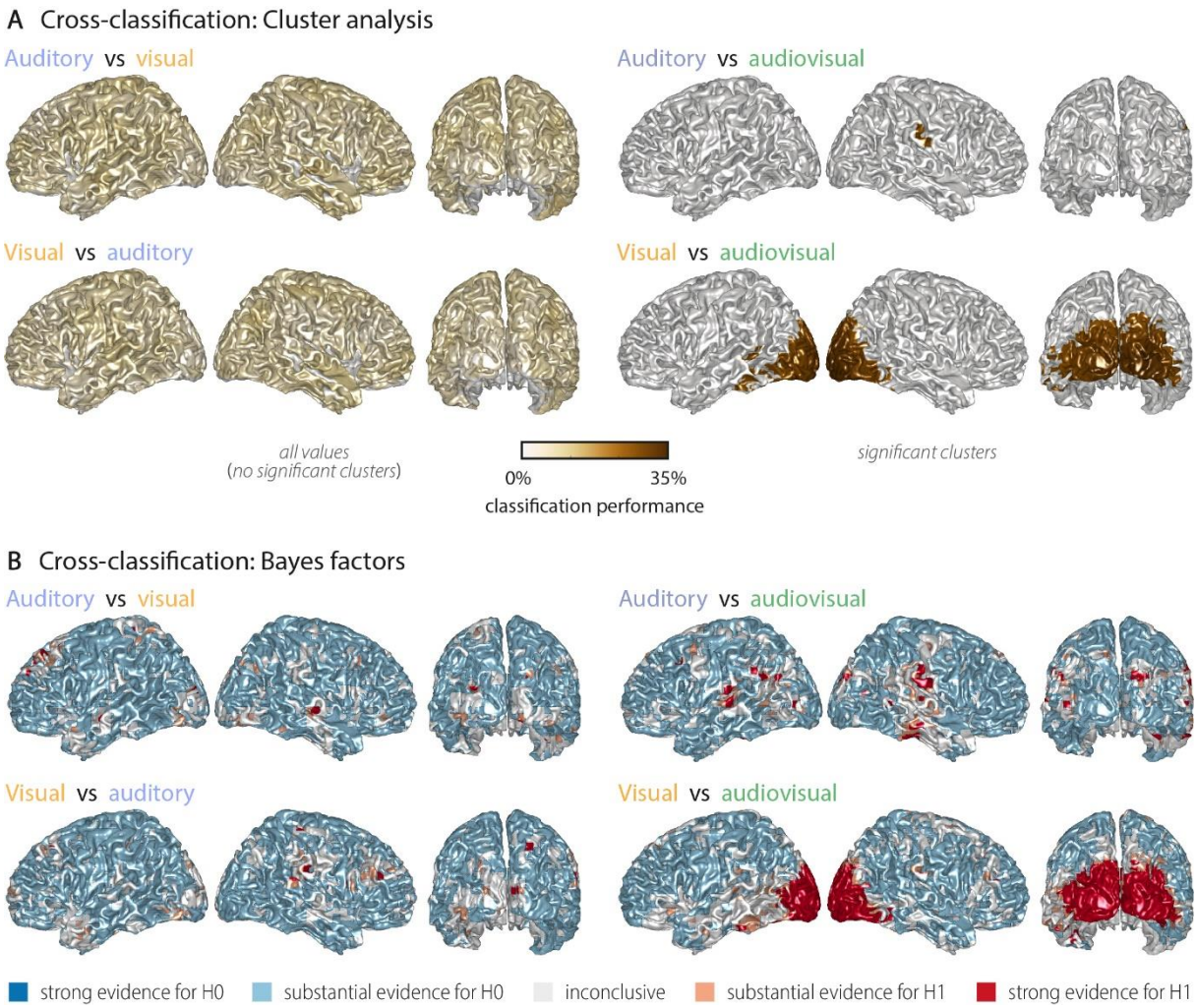


Suppl. Figure 1. Representational similarity analysis (RSA) of the behavioural data. Shown is the distribution of within-participant correlations between behavioural representational dissimilarity matrices (RDMs) and RDMs obtained from phonological and semantic representations of the target words (medians, interquartile ranges and densities). Behavioural RDMs were obtained by computing the confusion matrix from participants' behavioural responses. Pairwise phonological distances for the stimulus material were computed using Phonological Corpus Tools (V1.4.0), and semantic distances were computed using fastText vector representations (see **Materials and Methods** for details). Representational similarity was computed using Spearman rank correlations. Overall, phonological and semantic features both influenced participants' responses. A repeated measurements ANOVA (2 (conditions) x 2 (features)) yielded a main effect of condition (F(1,19) = 12.72, p = .002; mean correlations: M_{auditory} = .35, SEM = .01; M_{visual} = .43, SEM = .02) and a significant interaction (F(1,19) = 5.96, p = .025). A post-hoc comparison between phonological and semantic features yielded that, in the visual condition, phonological features influenced behaviour stronger than semantic features (Wilcoxon Signed-ranks test;,Z = 166, p = .023, no difference in auditory condition, Z = 91). Examples for close phonological and semantic relationships between two words are given in image.



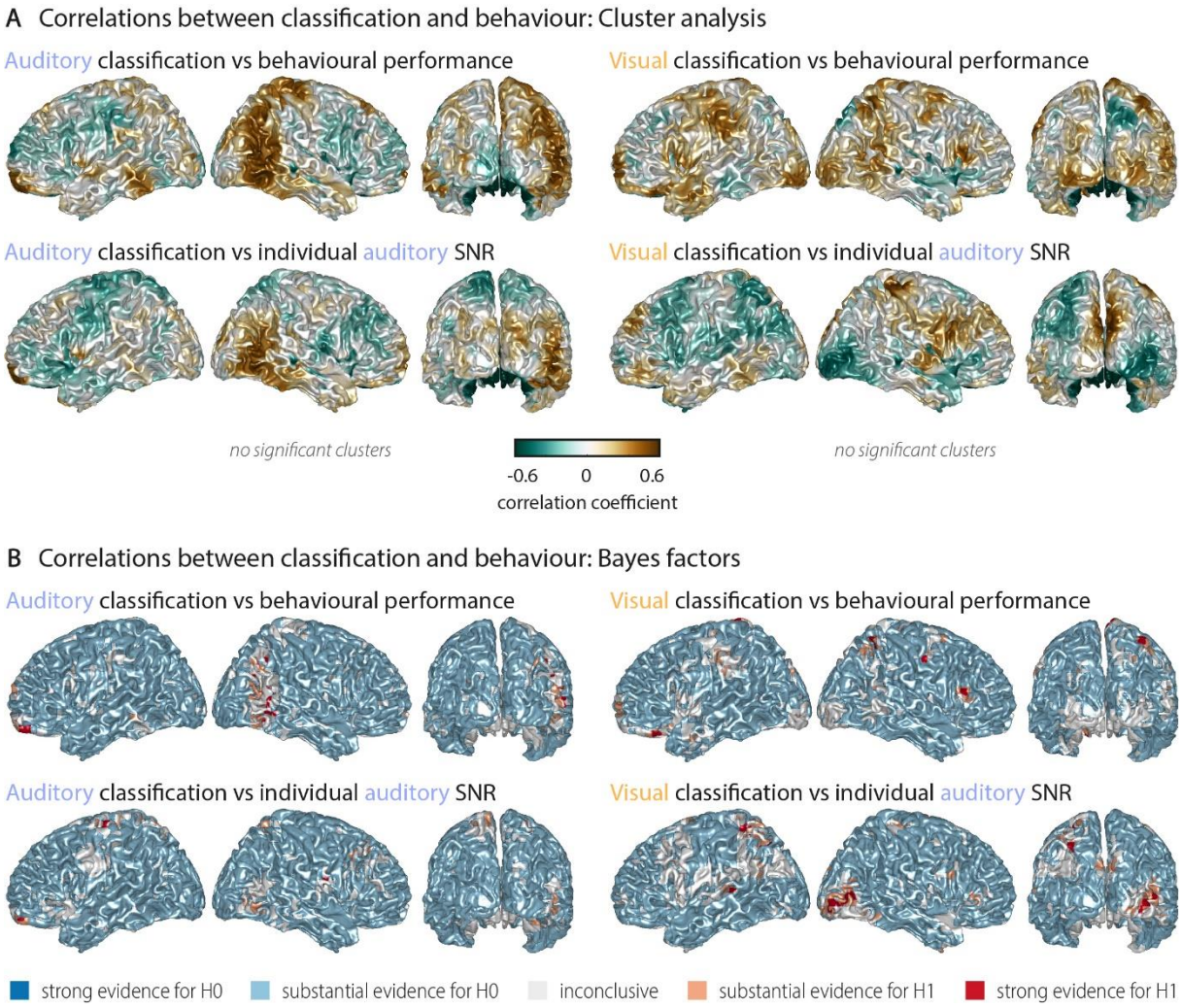
Suppl. Figure 2. Results of the audiovisual condition. **A)** Behavioural performance of 20 participants. Scaling of the figure is identical to the auditory and visual results for better comparability. Dots represent individual participants, boxes denote median and interquartile ranges, whiskers denote 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. **B)** Word classification regardless of behavioural performance ('stimulus classification'). Surface

projections show areas with significant classification performance at the group level (surface projection of the cluster-based permutation statistics, corrected at p<0.001 FWE). Strongest classification performance was observed in right auditory and bilateral visual sensory areas, with performance ranging from 25.63% to 33.43% (with a chance level of 25%). Statistical analysis yielded two clusters: a large bilateral cluster covering occipital and temporal regions that peaked in the right inferior occipital gyrus (right OCC; T_{sum} = 461.36, p < .001) and a left-hemispheric cluster that peaked in the middle temporal pole (left TP; Tsum = 12.62, p < .001). Cluster peaks are marked with dots TP – temporal pole; OCC – occipital gyrus.

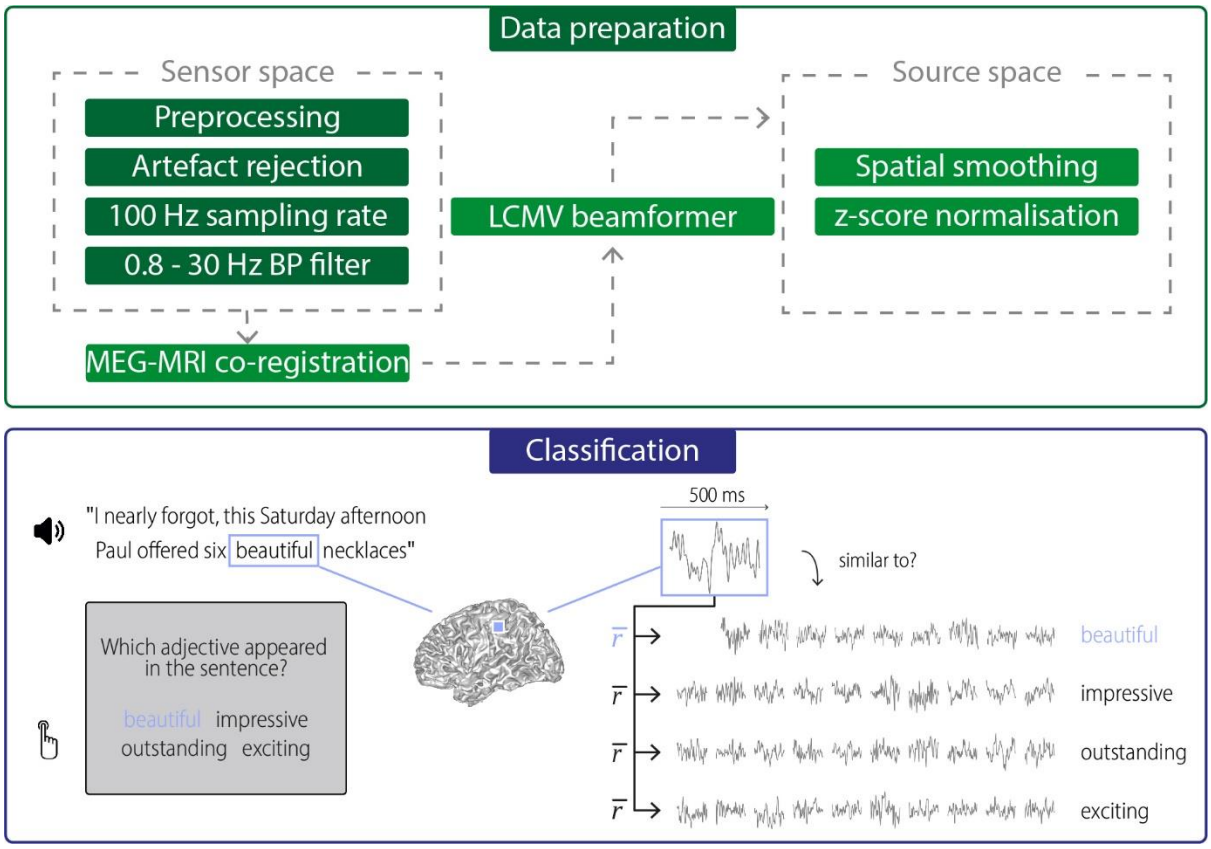


Suppl. Figure 3. Cross-classification between auditory, visual conditions and audiovisual conditions. **A)** Left panel: No significant cross-classification performance between the auditory and visual conditions was found (neither at p < 0.001, nor at a more lenient p < 0.05), supporting the notion that auditory and visual word identities are largely represented in different networks. Right panel: 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 (T_{sum} = 2.61, p < .001). Visual word identities can be classified from audiovisual word presentations in a large cluster in bilateral occipital cortex (T_{sum} = 224.62, p < .001) and a small left-hemispheric cluster in the left inferior temporal gyrus (T_{sum} = 7.05, p < .001). Colour scale is adapted from Figure 2, to allow a comparison of results. **B)** Results of a Bayes factor analysis derived from a comparison of the actual cross-classification performance to a distribution of performance values after data randomisation (red: evidence for significant cross-decoding, i.e. in favour of H1 [alternative hypothesis]; blue: evidence for no significant cross-decoding, i.e. in favour of H0 [null hypothesis]). The resulting Bayes factors show that for most grid points

there was substantial evidence for no cross-classification between the auditory and visual conditions, while there was substantial or strong evidence for cross-classification between the auditory (visual) and the audio-visual condition.



Suppl. Figure 4. Correlations between word classification and behavioural indices. **A)** Surface projection of resulting rho-values. None of the correlations yielded any significant clusters at an alpha-level of 0.05, supporting that stimulus classification performance alone does not predict behaviour. Upper panel: Correlation between auditory (visual) word classification and participants' performance in the auditory (visual) conditions. Lower panel: Correlation between auditory (visual) word classification and participant-specific SNR values used for the auditory condition. **B)** Results of a Bayes factor analysis for correlation coefficients (red: evidence for significant cross-decoding, i.e. in favour of H1 [alternative hypothesis]; blue: evidence for no significant cross-decoding, i.e. in favour of H0 [null hypothesis]). For the large majority of brain regions, there is substantial evidence for no correlation. Exceptions are scattered grid points, most notably in right occipital regions for a (negative) correlation between visual word classification and individual auditory SNR values across participants.



Suppl. Figure 5. Data preparation and classification procedures. **Data preparation**: Raw data were first de-noised and SQUID jumps were removed (preprocessing). Eye and heart artefacts and noisy channels were removed via visual inspection and ICA. Clean data were then down-sampled to 100 Hz and bandpass-filtered for further analysis. Data were projected into source space using an LCMV beamformer after a semi-automatic co-registration of MEG and individual anatomical MRIs. In source space, data were spatially smoothed (FWHM = 3 mm) and the time series for each grid point and trial was *z*-scored. **Classification**: Neural activity during hearing/seeing the target words was extracted for each grid point (500-ms windows) For each trial, this activity (during hearing "beautiful" in the above example) was compared with activity to the same target word in other trials (within-target distance), and with activity to the three alternative words in other trials (between-target distances). Distances were quantified using Pearson correlations of the spatio-temporal searchlight activity and were averaged across words ($\bar{r}$). If the within-target correlation was larger than the between-target correlations, the stimulus was classified correctly for this trial and grid point.

Supplementary Table

Suppl. Table 1. Target words used in this study were adjectives and numbers. Note that adjectives were comparable with regard to their positive valence (G. G. Scott et al., 2019).

Adjectives	Numbers
Beautiful	Eleven
Amazing	Twenty-one
Exciting	Thirty-two
Fantastic	Forty-three
Outstanding	Fifty-four
Colourful	Sixty-five
Impressive	Seventy-six
Glorious	Eighty-seven
Meaningful	Ninety-eight

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*.

Bednar, A., Boland, F. M., & Lalor, E. C. (2017). Different spatio-temporal electroencephalography features drive the successful decoding of binaural and monaural cues for sound localization. *European Journal of Neuroscience, 45*(5), 679-689.

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. Retrieved from <Go to ISI>://WOS:A1995QE45300017.

Bernstein, L. E., & Liebenthal, E. (2014). Neural pathways for visual speech perception. *Frontiers in Neuroscience, 8*. Retrieved from <Go to ISI>://000346554900001. 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.

Bourguignon, M., Baart, M., Kapnoula, E. C., & Molinaro, N. (2020). Lip-reading enables the brain to synthesize auditory features of unknown silent speech. *Journal of Neuroscience, 40*(5), 1053-1065.

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*. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/29363598. doi:10.1073/pnas.1714279115

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spat Vis, 10*(4), 433-436. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9176952.

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. Retrieved from <Go to ISI>://000256789500005. 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. Retrieved from https://www.biorxiv.org/content/biorxiv/early/2017/06/08/147298.full.pdf. doi:10.1101/147298

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

Correia, J. M., Jansma, B., Hausfeld, L., Kikkert, S., & Bonte, M. (2015). EEG decoding of spoken words in bilingual listeners: from words to language invariant semantic-conceptual representations. *Frontiers in psychology, 6*, 71.

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 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., & Davis, M. H. (2015). Hierarchical Organization of Auditory and Motor Representations in Speech Perception: Evidence from Searchlight Similarity Analysis. *Cereb Cortex, 25*(12), 4772-4788. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/26157026. doi:10.1093/cercor/bhv136

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.

Feng, W., Störmer, V. S., Martinez, A., McDonald, J. J., & Hillyard, S. A. (2014). Sounds activate visual cortex and improve visual discrimination. *Journal of Neuroscience, 34*(29), 9817-9824.

Ferraro, S., Van Ackeren, M. J., Mai, R., Tassi, L., Cardinale, F., Nigri, A., . . . Weisz, N. (2020). Stereotactic electroencephalography in humans reveals multisensory signal in early visual and auditory cortices. *cortex*.

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*. Retrieved from <Go to ISI>://000402875800001. 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. Retrieved from <Go to ISI>://WOS:000302114500008. doi:10.1038/nn.3063

Grave, E., Bojanowski, P., Gupta, P., Joulin, A., & Mikolov, T. (2018). Learning word vectors for 157 languages. *arXiv preprint arXiv:1802.06893*.

Grootswagers, T., Cichy, R. M., & Carlson, T. A. (2018). Finding decodable information that is read out in behaviour. *bioRxiv*, 248583. Retrieved from https://www.biorxiv.org/content/biorxiv/early/2018/01/16/248583.full.pdf. 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. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/24391472. 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.

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. Retrieved from <Go to ISI>://WOS:000311176000002. 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. Retrieved from <Go to ISI>://WOS:000246395200017. 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.

Ibrahim, L. A., Mesik, L., Ji, X.-y., Fang, Q., Li, H.-f., Li, Y.-t., . . . Tao, H. W. (2016). Cross-modality sharpening of visual cortical processing through layer-1-mediated inhibition and disinhibition. *Neuron, 89*(5), 1031-1045.

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

Jeffreys, H. (1998). *The theory of probability*: OUP Oxford.

Karmali, F., Chaudhuri, S. E., Yi, Y., & Merfeld, D. M. (2016). Determining thresholds using adaptive procedures and psychometric fits: evaluating efficiency using theory, simulations, and human experiments. *Experimental brain research, 234*(3), 773-789.

Keitel, A., Gross, J., & Kayser, C. (2018). Data from: Perceptually relevant speech tracking in auditory and motor cortex reflects distinct linguistic features. In.

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. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/27903440. 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.

Kennedy-Higgins, D., Devlin, J. T., Nuttall, H. E., & Adank, P. (2020). The Causal Role of Left and Right Superior Temporal Gyri in Speech Perception in Noise: A Transcranial Magnetic Stimulation Study. *Journal of cognitive neuroscience*, 1-11.

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. Retrieved from <Go to ISI>://A1994PR79400014.

Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008). Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience, 2*, 4.

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. Retrieved from http://www.jneurosci.org/content/jneuro/31/31/11338.full.pdf. doi:10.1523/jneurosci.6510-10.2011

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. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/17582338. doi:10.1016/j.neuron.2007.06.004

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. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/17517438. doi:10.1016/j.jneumeth.2007.03.024

McDonald, J. J., Störmer, V. S., Martinez, A., Feng, W., & Hillyard, S. A. (2013). Salient Sounds Activate Human Visual Cortex Automatically. *The Journal of Neuroscience, 33*(21), 9194-9201. Retrieved from https://www.jneurosci.org/content/jneuro/33/21/9194.full.pdf. doi:10.1523/jneurosci.5902-12.2013

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. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/17056494. doi:10.1080/02699200500266745

Morillon, B., Arnal, L. H., Schroeder, C. E., & Keitel, A. (2019). Prominence of delta oscillatory rhythms in the motor cortex and their relevance for auditory and speech perception. *Neuroscience & Biobehavioral Reviews*. doi:https://doi.org/10.1016/j.neubiorev.2019.09.012

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. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/14680264.

Oldfield, R. C. (1971). The assessment and analysis of handedness: the Edinburgh inventory. *Neuropsychologia, 9*(1), 97-113. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/5146491.

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. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/21253357. 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.

Park, H., & Kayser, C. (2019). Shared neural underpinnings of multisensory integration and trial-by-trial perceptual recalibration in humans. *eLife, 8*.

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. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/25890390. 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.

Petro, L., Paton, A., & Muckli, L. (2017). Contextual modulation of primary visual cortex by auditory signals. *Philosophical Transactions of the Royal Society B: Biological Sciences, 372*(1714), 20160104.

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.

Plass, J., Brang, D., Suzuki, S., & Grabowecky, M. (2019). Vision Perceptually Restores Auditory Spectral Dynamics in Speech.

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.

Revina, Y., Petro, L. S., & Muckli, L. (2018). Cortical feedback signals generalise across different spatial frequencies of feedforward inputs. *Neuroimage, 180*, 280-290.

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. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/25650107. 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.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic bulletin & review, 16*(2), 225-237.

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

Scott, G. G., Keitel, A., Becirspahic, M., Yao, B., & Sereno, S. C. (2019). The Glasgow Norms: Ratings of 5,500 words on nine scales. *Behavior research methods, 51*(3), 1258-1270.

Scott, S. K. (2019). From speech and talkers to the social world: The neural processing of human spoken language. *science, 366*(6461), 58-62.

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. Retrieved from <Go to ISI>://A1954UD61600011. 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. Retrieved from <Go to ISI>://A1992HC17000009. doi:DOI 10.1098/rstb.1992.0009

Tabarelli, D., Keitel, C., Gross, J., & Baldauf, D. (2020). Spatial attention enhances cortical tracking of quasi-rhythmic visual stimuli. *Neuroimage, 208*, 116444.

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. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/26656644. 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. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11771995. doi:10.1006/nimg.2001.0978

Vaden, K. I., Halpin, H., & Hickok, G. S. (2009). Irvine phonotactic online dictionary, Version 2.0.[Data file]. In.

Van Veen, B. D., van Drongelen, W., Yuchtman, M., & Suzuki, A. (1997). Localization of brain electrical activity via linearly constrained minimum variance spatial filtering. *IEEE Trans Biomed Eng, 44*(9), 867-880. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9282479. doi:10.1109/10.623056

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., & Van Der Maas, H. L. (2011). Why psychologists must change the way they analyze their data: the case of psi: comment on Bem (2011).

Wetzels, R., & Wagenmakers, E.-J. (2012). A default Bayesian hypothesis test for correlations and partial correlations. *Psychonomic bulletin & review, 19*(6), 1057-1064.

Yi, A., Wong, W., & Eizenman, M. (2013). Gaze patterns and audiovisual speech enhancement. *Journal of Speech, Language, and Hearing Research*.