NeuroImage 224 (2021) 117365

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/neuroimage

Breaking down the cocktail party: Attentional modulation of cerebral audiovisual speech processing

Patrik Wikman[a,*], Elisa Sahari[a], Viljami Salmela[a,e], Alina Leminen[a,b], Miika Leminen[a,c], Matti Laine[d], Kimmo Alho[a,e]

[a] Department of Psychology and Logopedics, University of Helsinki, Helsinki, Finland
[b] Department of Digital Humanities, University of Helsinki, Helsinki, Finland
[c] Department of Phoniatrics, Helsinki University Hospital, Helsinki, Finland
[d] Department of Psychology, Åbo Akademi University, Turku, Finland
[e] Advanced Magnetic Imaging Centre, Aalto NeuroImaging, Aalto University, Espoo, Finland

ARTICLE INFO

Keywords:
Selective attention
Audiovisual speech
Cocktail party
fMRI
Semantics
Multivariate pattern analysis (MVPA)

ABSTRACT

Recent studies utilizing electrophysiological speech envelope reconstruction have sparked renewed interest in the cocktail party effect by showing that auditory neurons entrain to selectively attended speech. Yet, the neural networks of attention to speech in naturalistic audiovisual settings with multiple sound sources remain poorly understood. We collected functional brain imaging data while participants viewed audiovisual video clips of life-like dialogues with concurrent distracting speech in the background. Dialogues were presented in a full-factorial design, comprising task (listen to the dialogues vs. ignore them), audiovisual quality and semantic predictability. We used univariate analyses in combination with multivariate pattern analysis (MVPA) to study modulations of brain activity related to attentive processing of audiovisual speech. We found attentive speech processing to cause distinct spatiotemporal modulation profiles in distributed cortical areas including sensory and frontal-control networks. Semantic coherence modulated attention-related activation patterns in the earliest stages of auditory cortical processing, suggesting that the auditory cortex is involved in high-level speech processing. Our results corroborate views that emphasize the dynamic nature of attention, with task-specificity and context as cornerstones of the underlying neuro-cognitive mechanisms.

1. Introduction

Listening and comprehending speech in noisy environments is so effortless for humans that we often ignore its computational demands. Although artificial intelligence (AI) platforms can transcribe noise-free speech with ever increasing accuracy, the AI transcription of speech in noisy conversational settings has proven much less satisfactory (for recent advances, see Ephrat et al., 2018). The superior performance of humans compared to computers relates to attentional selection that enables speech comprehension in noise. We still know little about the neural processes underlying such attentional selection. One reason is that research on speech comprehension (e.g., Friederici, 2011; Friederici and Gierhan, 2013) has advanced largely in isolation from research on selective attention. Indeed, attention-related modulation of neural responses to speech sounds has classically been presumed to comprise simple mechanisms that non-specifically increase the gain and fidelity of neuronal responses (Briggs et al., 2013). This view has, however, changed during the last decade thanks to methodological advances that enable studying selective attention in ecologically valid settings with multiple sound sources. These studies, using electrocorticography (ECog; Golumbic et al., 2013; Mesgarani and Chang, 2012; O'Sullivan et al., 2019), electroencephalography (EEG; O'Sullivan et al., 2015; Riecke et al., 2019) and magnetoencephalography (MEG; Ahissar et al., 2001; Zion Golumbic et al., 2013), have shown that when listeners attend to one speech stream in the presence of competing speech streams, auditory neurons differentially track changes in the speech-sound envelope of the attended stream. Furthermore, functional magnetic resonance imaging (fMRI) studies suggest that attention modulates processing in auditory cortex (here core, belt and parabelt, Moerel et al., 2014) and surrounding regions when listening to speech in the presence of noise (e.g., Alho et al., 2006; Alho et al., 2003).

Since the pioneering studies on selective listening (Cherry, 1953), it has been debated whether attentional selection occurs based on simple physical features of the attended voice (e.g., location or pitch) or whether it reflects higher-level (e.g., semantic) processing

* Corresponding author.
E-mail address: patrik.wikman@helsinki.fi (P. Wikman).

https://doi.org/10.1016/j.neuroimage.2020.117365
Received 15 May 2020; Received in revised form 19 August 2020; Accepted 7 September 2020
Available online 14 September 2020

of speech sounds. Early psychophysical studies suggested that attentional selection of speech operates on low-level auditory features (Broadbent, 1954). Later studies showed that unattended words may slow down processing of coinciding attended words when the attended and unattended words are semantically related, supporting late selection models (Deutsch and Deutsch, 1963; Lewis, 1970). The results of Treisman and colleagues (Treisman et al., 1974) partly reconciled these views by showing that the distracting effect of semantic similarity between unattended and attended words decreases with time, probably because attention becomes more effectively entrained to attended voice (Hansen and Hillyard, 1988; Näätänen et al., 1992). To explain such findings, Näätänen (1990) suggested that during selective listening there is a gradual buildup of a so-called 'attentional trace' in the auditory cortex. This trace represents the physical features (e.g., location or pitch) that distinguish the relevant sounds from the irrelevant ones, and all incoming sounds are compared with this trace and only the matching ones are selected for further processing. Thus, the assumption that semantics do not influence early speech processing has remained and has received some support from electrophysiological studies. For example, semantic manipulations in attended sentences do not modulate event-related potentials (ERP) until 250–500 ms after stimulus presentation (see, e.g., Connolly et al., 2001; Kutas and Hillyard, 1980; Lau et al., 2008), suggesting that these modulations are generated by higher-level speech processing, beyond the primary and adjacent auditory cortex. Yet, some fMRI (Tuennerhoff and Noppeney, 2016; Wild et al., 2012) and EEG/MEG studies (Luo and Poeppel, 2007; O'Sullivan et al., 2019; Peelle et al., 2013; Sohoglu et al., 2012) have suggested that speech intelligibility modulates low-level speech processing in the auditory cortex, presumably by enhancing the processing of the acoustic and phonetic features of speech while concurrently improving semantic processing. However, some recent studies suggest that semantic features alone may modulate processing in auditory cortex: A recent EEG study, utilizing continuous natural speech and speech-amplitude-envelope-decoding analyses, showed that the semantic association strength of a word to its sentential context modulated cortical tracking of the speech-amplitude envelope (Broderick et al., 2019). Importantly, this semantic modulation was evident as early as 50–100 ms from word onset, suggesting modulation of early cortical processing by semantics. Thus, there is increasing evidence that higher-order information such as phonology and semantics intervenes in low-level auditory processing (see also, Rutten et al., 2019).

In the present study, we used fMRI to systematically investigate attentional modulation of speech processing during a simulated audiovisual cocktail-party-like setting. In contrast to the recent advances in speech envelope analysis using electrophysiological measurements, there are currently much less fMRI studies on attention-related processes of speech in settings with multiple sound sources. Furthermore, unlike previous studies that have mostly used auditory speech only and a single speaker, here we instead utilized audiovisual (AV) speech with two speakers interacting with each other. In our novel paradigm, participants attended to video clips of dialogues between a male and a female speaker with a distracting speech stream played in the background (Fig. 1A). To increase attentional demands, we modulated the auditory quality by noise-vocoding (Boersma and Weenink, 2001) and visual quality in the videos by masking (Sumby and Pollack, 1954). We also modulated the semantic context: Half of the dialogues had a coherent discourse while in the other dialogues the discourse was incoherent and consisted of mixed lines from different dialogues (see Fig. 1D). We employed a fully factorial design where participants performed two different tasks: 1) AV speech task, where the participants attended to the dialogue while ignoring distracting speech, and 2) visual control task, where the participants ignored the dialogue and the distracting speech, and instead counted rotations of a cross presented near the mouth of either speaker. Comparing these two tasks enabled us to separate modulations of brain activity due to attentive processing of AV speech from stimulus-dependent activations and the switching of attention between the left and right hemifields.

We used a combination of univariate analyses and multivariate decoding/classification to study attentive processing of AV speech. We expected that attentive processing of AV speech would modulate activations in a widespread network consisting of the auditory cortex and adjacent STG/STS regions involved in sensory processing of speech. The AV quality of the videos was thought to interact with attention-related brain activity in two alternative and even opposing ways: Either good AV quality might cause stronger entrainment of attention to speech (Ding et al., 2014; Evans et al., 2016; Golumbic et al., 2013; Kong et al., 2015; Leminen et al., 2020), or poor AV quality might lead to enhanced attention-related modulations of brain activity due to increased effort. Moreover, improved semantic clarity might modulate activity either only in the higher-order auditory regions such as the STG/STS (Näätänen et al., 1992; Tylen et al., 2015), or also in the low-level processing regions of the auditory cortex (Broderick et al., 2019).

2. Methods and materials

2.1. Participants

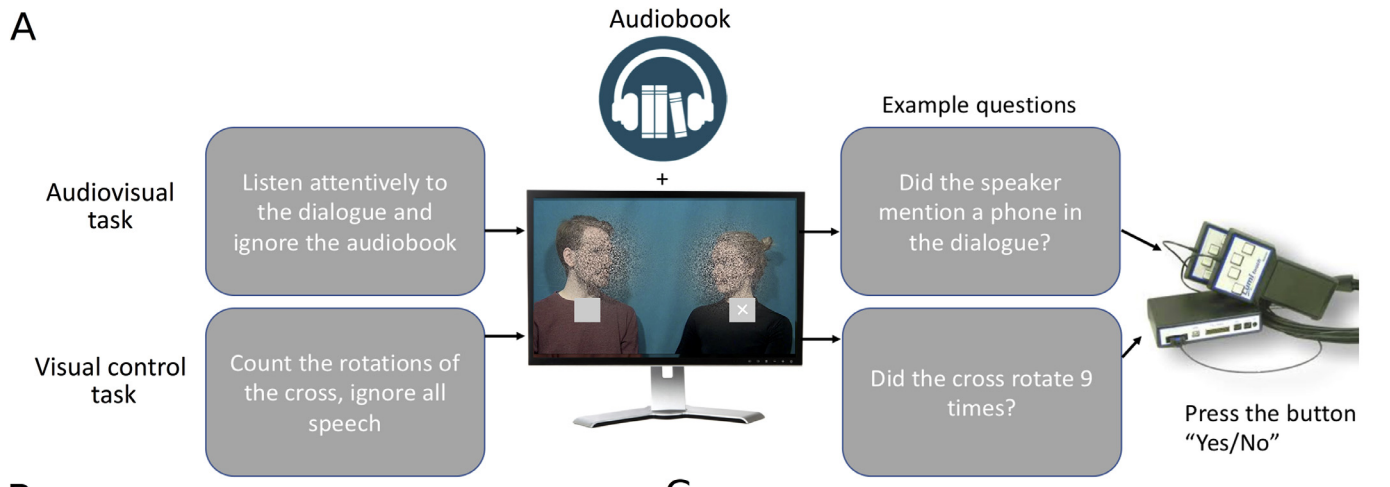
Functional magnetic resonance imaging (fMRI) data were collected from 23 adult participants (14 females, age range 19–30 years, mean 24.3 years). fMRI data were excluded based on pre-established criteria. Two participants were excluded due to excessive head motion (> 5 mm) and two participants due to anatomical anomalies that effected coregistration. Thus, data form 19 participants (12 females) was used in the analyses. Participants were right-handed native Finnish speakers with normal hearing, normal or corrected vision, and had no history of psychiatric or neurological illnesses. Handedness was verified by the Edinburgh Handedness Inventory (Oldfield, 1971). Before the experiment, a written consent was obtained from each participant. The experiment was approved by the Ethics Committee of the Hospital District of Helsinki and Uusimaa, Finland.

2.2. Preparation of stimulus materials

The stimuli consisted of dialogues between two (female and male) native Finnish speakers. Consent has been obtained from the two individuals (see, Leminen et al., 2020). The dialogues were related to neutral everyday subjects such as the weather. The dialogues consisted of seven lines (ca. 5.4 s of duration) followed by a ca. 3.5 s break (2.9–4.3 s). Thus, the total length of each dialogue was 55–65 s (mean 59.2 s). Talkers spoke their lines in an alternating fashion. In half of the video clips, the female talker started the conversation.

The original 36 dialogues (Leminen et al., 2020) were recorded so that the talkers sat next to one another with their faces slightly tilted towards each other (see Fig. 1A). This enabled us to keep the setting as natural as possible with the talkers' faces visible. For more details on the recordings, (see, Leminen et al., 2020).

In the present study, we used 24 of the original dialogues for the coherent context conditions. The rest of the dialogues were used to construct 24 new dialogues with an incoherent plot for the incoherent context conditions. This was achieved by shuffling lines from different dialogues. First, the audio stream was removed from the video. Thereafter, the video image was edited with Adobe Premiere Pro CC -software (Adobe Inc, San Jose, California, USA). To prevent participants from noticing the changes between the dialogues, the transition from one dialogue to another always occurred on the side where the talker was silent. In other words, if the female was talking, half-way into the line the video image of the male talker would change into that of another dialogue (see Supplementary video material 1–8; https://osf.io/agxth/). Transitions between the two dialogues were made as smooth as possible using Adobe Premiere (Adobe Inc, San Jose, California, USA) morph-cut



B

Example lines from a coherent dialogue	Example lines from a incoherent dialogue
A: I bought a new phone and it has so many features that I am completely lost.	B: He first graduated from commercial institute, but then decided to continue to school of economics.
B: Oh yes, you used to have that ancient phone, which was not even a smart phone.	A: But it is not classical music, I am going to a rock concert.
A: Yes, and it was a perfectly good phone until my cat dropped it on the floor from a table, and that was it.	B: I would gladly try something new and exciting, let´s go together to the course.

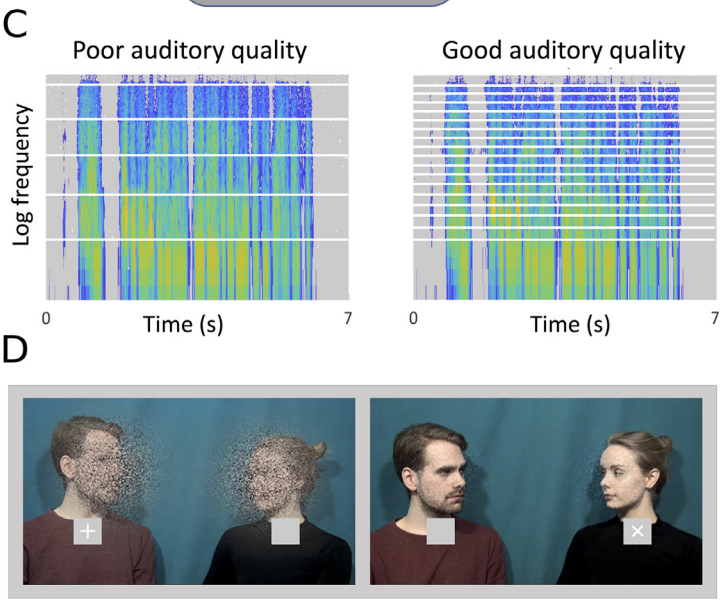


Fig. 1. The audiovisual (AV) cocktail-party paradigm used in the current study. **(A)** Participants were presented with video clips (ca. 1 min in duration) of male and female speakers discussing neutral topics, such as the weather, while a continuous audiobook was played in the background. Speech from the two talkers alternated with a short break between talkers. Participants performed two tasks: 1) an AV speech task, where they attended to the dialogue while ignoring the audiobook and answered questions about each line of the dialogue immediately after the video-clip finished, and 2) visual control task, where the participants ignored the dialogue and audiobook, and instead counted rotations of a cross presented below the neck of the talker who was speaking at the moment (see 2.2–2.4 details). **(B)** The semantic context of the dialogue was either coherent (left), that is, each line logically followed the previous one, or incoherent (right), where successive lines were unrelated. **(C)** Videos were presented at two levels of auditory quality: Poor auditory quality, where the audio stream of the dialogue was noise-vocoded (Boersma and Weenink, 2001) with four logarithmically equidistant frequency bands above 0.3 kHz (i.e., the fundamental frequency was untouched), and good auditory quality, where it was noise-vocoded using 16 bands above 0.3 kHz (white horizontal lines on the spectrograms denote the frequency band borders). **(D)** Visual quality of the faces was modulated by masking the speakers' faces with different amounts of dynamic white noise.

function. The lighting of the videos was edited to fade small differences between the different clips.

The original dialogues used for creating the shuffled dialogues were chosen based on the location and posture of the speakers so that there would be minimal visual transition between each line of the shuffled dialogues. However, because slight differences in lightning and posture of the speakers, we chose to divide the videos into pools of six videos that were maximally similar. Note that each new incoherent dialogue had seven lines. Thus, each of the five line were from a separate dialogue, and the remaining two from one dialogue. To secure that all lines were equally unpredictable, we made sure that the two lines from the same original dialogue were separated by at least 4 other lines.

Two small grey squares (size 1.5° × 1.5 °) were added to each video below the faces of the speakers. A white cross (height 0.5 °) was placed in the middle of the square below the face of the talker who was speaking. This cross faded out immediately as the talker ended his/her line and reappeared 1.5 s later. Thus, most of the time there were two crosses present (see Suppl. Video material 1–8; unlike in our experiments, these videos have English subtitles). In the visual control task, the disappearance of the cross indicated that the participant should turn their attention to the other side of the frame. The cross changed from a multiplication sign (×) to a plus sign (+), or vice versa, randomly 9–15 times during each dialogue. The cross rotated only on the side where the talker was speaking. During each of the seven lines, the cross rotated 1–4 times (every 1.25–2.5 s).

Before adding the audio streams back to the videos, they were noise-vocoded (see, Leminen et al., 2020). We divided the audio streams into 4 and 16 logarithmically spaced frequency bands between 0.3 and 5 kHz using the Praat software (version 6.0.27, Boersma and Weenink, 2001). The talkers' F0 (frequencies 0–0.3 kHz) was unchanged to maintain a clear male-female difference. According to the behavioural ratings (N = 5) done in our previous publication (Leminen et al., 2020), the

speech with 4 noise-vocoded frequency bands was considered almost unintelligible (poor auditory quality), and the speech with 16 noise-vocoded frequency bands was considered intelligible (good auditory quality).

To manipulate the amount of visual speech seen by the participants, we added dynamic white noise onto the speakers' faces (two levels of magnitude: good and poor visual quality; for details see Leminen et al., 2020). This was done by using custom-made Matlab scripts (Matlab R2016, Mathworks Inc., Natick, Massachusetts, USA). According to behavioural ratings (N = 5, Leminen et al., 2020), the visual noise rendered the mouth movements and facial features poorly visible in the poor visual quality conditions.

Finally, the poor and good quality audio files were recombined with the poor and good quality videos with a custom Matlab script.

As the last step to render the videos into a cocktail-party-like setting, we added a continuous background speech stream to the dialogues. We chose a freely available audiobook about cultural history (a Finnish translation of *The Autumn of the Middle Ages* by Johan Huizinga, distributed online by YLE, the Finnish Broadcasting company). The book was read by a female native Finnish professional actor. The F0 of the female voice was lowered to 0.16 kHz. Additionally, the audiobook was low-pass filtered at 5.0 kHz (for more information see, Leminen et al., 2020).

2.3. Procedure

The videoclips described above were used in our 16 experimental conditions defined by Task (AV speech task, visual control task), Semantic Context (coherent, incoherent), Auditory Quality (good, poor) and Visual Quality (good, poor). Note that we had three runs, each containing eight of the 24 coherent video clips (in all coherent context conditions) and eight of the 24 incoherent video clips (in all incoherent context conditions). Thus, all the participants were presented with all the 48 dialogues. Every other run started with the AV speech task, and every other with the visual control task. Within the functional runs, the AV speech task and the visual control task were presented in an alternating order. The order of conditions and dialogues presented was pseudorandomized. Because we could not entirely randomise the videos into the 16 conditions per run, we used the Latin square to construct four different versions of the experiment (see Suppl. Table 3), six participants completed the first version, five the second, four the third and four the fourth version.

Stimulus presentation was controlled using Presentation 20.0 software (Neurobehavioral Systems, Berkeley, California, USA). The auditory stimuli were presented binaurally through insert earphones (Sensimetrics model S14; Sensimetrics, Malden, Massachusetts, USA). Before the experiment, the audio volume was set to a comfortable level individually for each participant**.** It was approximately 75–86 dB SPL at the ear drum. The video clips (size 26° × 15°) were projected onto a mirror attached to the head coil and presented in the middle of the screen on a grey background. In the middle of each run, there was a break of 40 s. During the break, the participants were asked to rest and focus on a fixation cross (located in the middle of the screen, height 0.5°). The distracting audiobook (presented with a sound intensity 3 dB lower than the voices of viewed male and female speakers) started randomly 0.5–2 s before video onset and stopped at the offset of the video. The differences in dialogue durations were compensated by inserting periods with a fixation cross between the instruction and the onset of the dialogue, keeping the overall trial durations constant.

2.4. Tasks

During the AV speech task, the participants were asked to attend to the videos and ignore the background speech. After every dialogue, the participants were presented with seven statements which each related to the occurrence of a topic in each line from the dialogue by pressing the 'Yes' or 'No' button on a response pad with their right index or middle finger, for example, "Did the boy drop his phone?", "Was there a cat on the table?". A new statement was presented every 2 s. After the seven statements, the participants were provided with feedback on their performance.

During the visual control task, the participants were asked to attend to the fixation cross in the videos and calculate how many times the cross rotated from a multiplication sign (×) to a plus sign (+) and vice versa. Every time the cross disappeared, the participants were supposed to shift their attention to the other fixation cross on the other side of the frame. The participants were also instructed to ignore the dialogues and the audiobook. At the end of the videoclip, the participants were presented with seven statements about the rotating cross ("Did the cross turn X times?" the X being between 9 and 15 in an ascending order). As in the AV speech task, the response was given by pressing either the 'Yes' or 'No' button on a response pad. If the participants were not sure, they were instructed to answer 'Yes' to all the alternatives they deemed possible. After the seven statements, the participants received feedback on their performance.

Please, refer to Supplementary Videos 1-8 for schematic examples of the stimulus materials used for the two tasks.

2.5. Pre-trial

Before the experiment, the participants practised the tasks outside the scanner. In the practice phase, the participants performed the AV speech task and the visual control task, using a coherent dialogue not included in the actual experiment. The dialogue was presented with different auditory and visual qualities.

2.6. Data acquisition

Functional brain imaging was carried out with 3T MAGNETOM Skyra whole-body scanner (Siemens Healthcare, Erlangen, Germany) using a 20-channel head coil. The functional echo planar (EPI) images consisted of 43 continuous oblique axial slices (TR 2600 ms, TE 30 ms, flip angle 75°, voxel matrix 64 × 64, field of view 19.2 cm, slice thickness 3.0 mm, in-plane resolution 3 mm × 3 mm × 3 mm). The three functional runs consisted of 616 volumes, thus in total 1 848 functional volumes were measured from each participant in one session (session duration approximately 75 min). Additionally, electroencephalography was measured from each participant during the functional runs. Unfortunately, after collecting the EEG data we noticed that there was a jitter of tens of milliseconds in the fMRI pulse timing causing so much variation in the fMRI artifact that neither template matching, nor ICA-based approaches to clean the EEG data were successful. Therefore, these EEG data are not publishable. After the functional runs, high-resolution anatomical images (TE 3.3 ms, TR 2 530 ms, voxel matrix 256 × 256, in-plane resolution 1 mm × 1 mm × 1 mm) were acquired from each participant.

We collected three runs of fMRI data from 19 participants, each run containing 16 conditions outlined in Fig. 3.

2.7. Data analysis

2.7.1. Analysis of behavioural data

The total number of questions in the experiment was 336. We collected the number of correct button presses in each task block. Misses were treated as incorrect button presses. Mean and standard error of mean was used to establish that the participants were performing the task as expected. To analyse participants´ performance in the quizzes during the audiovisual and the visual control task, two separate repeated-measures analysis of variance (ANOVA) were computed with 3 factors: Semantic Context (coherent, incoherent), Auditory Quality (good, poor) and Visual Quality (good, poor). Statistical analyses were carried out with IBM 18 SPSS Statistics 25 (IBM SPSS, Armonk, New York, USA) software and the results were visualized with Matlab

R2018a (Mathworks Inc., Natick, Massachusetts, USA). Due to technical issues, the first two participants could not answer all of the statements in the AV speech tasks. There were nine values missing from the first participant and six from the second. These values were replaced with the mean from all other participants' scores in the condition in question.

2.7.2. First-level fMRI data analysis

Pre-processing of the fMRI data was carried out using SPM 12 (Statistical parametric mapping; https://www.fil.ion.ucl.ac.uk/spm/). Preprocessing included motion correction, slice timing correction, high pass filtering (1/130 Hz), and low-pass filtering. Registration to the high resolution structural images was carried out using the FLIRT BBR function (part of FSL FMRIB's Software Library, http://www.fmrib.ox.ac.uk/fsl, Jenkinson et al., 2012). First-level analysis of the preprocessed and coregistered data was performed using SPM 12. At this step, a general linear model with 120 explanatory variables was fitted to the time-series data from each voxel. The variables were [112 lines, i.e., 16 conditions × 7 lines per condition: Task (AV speech task, visual control task), Semantic Context (coherent, incoherent), Auditory Quality (good, poor), Visual Quality (good, poor)], as well as instructions, quizzes, and six motion correction parameters. To extract 'attentive audiovisual-speech-processing-related modulations' (AttnAVSMs) of brain activity, we also performed a second general linear model (GLM) analysis, where we subtracted linearly each of the dialogue line regressors during the audiovisual task from their corresponding visual task dialogue line regressors. That is, brain activity for each line of the AV speech task (8 conditions × 7 lines per dialogue) was contrasted with brain activity for the line of the visual control task *of the same run* with the same auditory and visual quality, level of coherence, and position in the dialogue. Main effects and interaction terms were built in the first-level GLM. To decrease the amount of multidimensional statistical maps we only estimated the main effects of the factors Task, Auditory-, Visual Quality and Semantic coherence for the mass group level univariate effects (see 2.7.5).

2.7.3. Psychophysical interactions (PPI) analysis

First the data was preprocessed as above. Thereafter, the fMRI data were projected onto the Freesurfer (Fischl, 2012) average surface based on the participants' own Freesurfer surface (surface smoothening: 5mm^2 full width half maximum smoothening). The PPI model was conducted using FSL with a model including a psychological regressor (the AV speech task or the visual control task), a physiological regressor (mean time series separately for the left and right hemisphere in Task main effect STG/STS clusters), and PPI (interaction between psychological and physiological regressors) as explanatory variables. The model also included all the rest of the explanatory variables of the original model (see 2.7.2).

2.7.4. Decoding analysis

Support vector machine (SVM) decoding with leave one run-out cross-validation (Cortes and Vapnik, 1995) was used to classify the AV Speech Task and the Visual Control Task. Each line constituted an exemplar and each voxel a feature in the analysis. This was conducted with the decoding toolbox (TDT, Hebart et al., 2015), using the beta-images from the first-level GLM in the participants' anatomical space. We used searchlight-based (Kriegeskorte et al., 2006) decoding with a radius of 6 mm (isotropic), and with default settings of TDT; L2-norm SVM with regularizing parameter C = 1 running in LIBSVM (Chang and Lin, 2011). All other decoding analyses were performed using the AttnAVSMs (same method as above). Here we estimated all possible main effects and interactions between Auditory Quality, Visual Quality and Semantic Context.

2.7.5. Group-level fMRI analysis

For the group-level analyses, the first-level results (contrasts, decoding accuracies etc.) were projected to the Freesurfer average (fsaverage) using the participants' own Freesurface surface (surface smoothening: 5mm^2 full width half maximum smoothening). Group level statistics were based on a two level-procedure using a one-sample t-test performed using the glm-fit function of the Freesurfer software. Clusters were defined using permutation inference (a robust method for controlling false discoveries; Greve and Fischl, 2009) in Freesurfer, with the initial cluster forming threshold $z = 4$, cluster probability $p < 0.01$ for the univariate analyses, and $z = 2.8$, cluster probability $p < 0.01$ for the multivariate analyses. Clusters smaller than 100 mm^2 were discarded.

2.7.6. Definition of left Heschl's gyrus (HG) ROIs and pairwise decoding

Heschl's gyrus (HG) was anatomically defined in each participant using their anatomical surface. In case of duplications or partial duplications of HG, the most frontal one was chosen. These ROIs were used as masks in pairwise decoding analyses where each possible AttnAVSM condition (8 conditions, i.e., Semantic Context × Auditory Quality × Visual Quality) was decoded from each other in pairs, resulting in 28 pairwise decoding analyses for each ROI. The method for decoding was the same as above.

2.7.7. Trend analysis, clustering and visualisation of ROI data

To gain an understanding on the temporal profile of the AttnAVSMs across the dialogues, we modelled linear trends (i.e. monotonic), quadratic trends and a combination of these two (linear-quadratic trends) at the first-level GLM.

We used the significant clusters from our linear-quadratic trend analyses (initial cluster threshold, $z = 4$; permutated cluster significance, $p < 0.01$) to define our ROIs for clustering analyses. We calculated the mean for each AttnAVSMs in each of the ROIs (56 AttnAVSMs in 52 ROIs). Then we calculated correlations using the 56 AttnAVSMs across the 52 ROIs. The resulting representational dissimilarity matrix (RDM; 1-correlation) is shown in Suppl. Fig. 5. This RDM was visualized in two ways: First, by 2D Multidimensional scaling (using Matlab function 'mdscale'), and then by hierarchical clustering and dendrography using Matlab function 'linkage', which applies Euclidean distances between the ROIs (Fig. 7). For further inspection, the dendrogram was thresholded using a maximum of 20 clusters.

2.8. Data availability

Due to concerns regarding participant privacy, structural MRI data and raw functional MRI data will not be made openly available. However, anonymised fMRI data which have been transformed into standard space and behavioural data may be made openly available. The data used to generate the figures in the study are shared using the Open science framework under Attention and Memory networks (https://osf.io/agxth/). Other anonymised data is available from the corresponding author on reasonable request. The computer code used to derive the findings of this study is available from the corresponding author upon reasonable request.

3. Results

3.1. Semantic context, auditory quality and visual quality modulate intelligibility of the video materials

We performed a three-way repeated measures ANOVA (Semantic Context, Auditory Quality, Visual Quality) on the performance in the AV speech task. As expected (Fig. 2, left), recollection of the dialogues was significantly affected by Semantic Context ($F_{1,18} = 107.2$, $p = 1 \times 10^{-8}$, FDR-corrected $p = 1.4 \times 10^{-7}$, $\eta^2 = 0.86$), Auditory Quality ($F_{1,18} = 40.7$, $p = 5.2 \times 10^{-6}$, FDR-corrected $p = 4.7 \times 10^{-5}$, $\eta^2 = 0.64$) and Visual Quality ($F_{1,18} = 32.4$, $p = 2.1 \times 10^{-5}$, FDR-corrected $p = 7 \times 10^{-4}$, $\eta^2 = 0.69$) presumably due to better intelligibility of coherent dialogues and dialogues with a better auditory or visual quality. The effect of Auditory Quality was, however, stronger for incoherent dialogues, indicated by an interaction between Semantic Context and Auditory

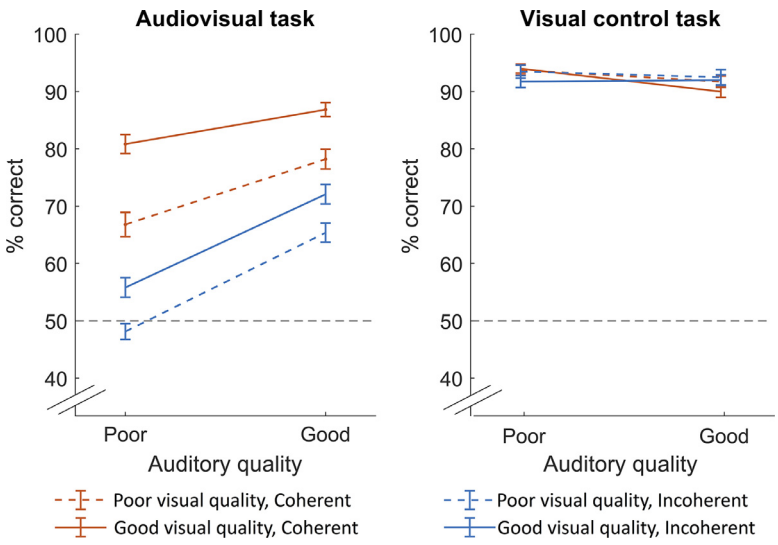


Fig. 2. Performance in the audiovisual (AV) speech task and the visual control task. In the AV speech task, performance was significantly modulated by all three factors. Further, there was an interaction between Semantic Context and Auditory Quality. There were no significant differences between the different conditions during the visual control task. Error bars denote ± standard errors of the mean.

Quality, which was not significant after FDR-correction ($F_{1,18}$ = 5.9, $p = 2.6 \times 10^{2}$, FDR-corrected $p = 9 \times 10^{-2}$, $\eta^2 = 0.25$).

Performance in the visual control task was analysed using a similar three-way ANOVA. This analysis yielded no significant effects (Fig. 2, right; F < 2.0, p > 0.17 in all cases), which was expected, as manipulations of the dialogues were irrelevant to this task. The lack of significant effects of the manipulations of the dialogues on visual task performance also suggests that the dialogues were not processed to a high degree during the visual control task, as more intelligible dialogues would increase errors in the visual task had they been covertly attended.

To further evaluate how much participants processed the dialogues during the visual control task, we conducted a separate control experiment with 18 new participants undergoing EEG data collection in a separate study (to be reported later) using the same paradigm. In this control experiment, participants were presented with a dialogue they had not seen/heard before and told to ignore the dialogue and perform the visual control task. However, at the end of the task, participants were asked one question about each of the lines of the dialogue they had been instructed to ignore. Each participant was presented with only one dialogue, as the participants might covertly start attending to the dialogues during the visual task once they knew that questions about the dialogues would be asked afterwards. In this control experiment, the dialogues were presented with good AV quality and coherent context, i.e. the dialogue that would be most difficult to ignore during the visual task. Recollection of the to-be-ignored dialogues was on average 57% correct, that is, slightly above the chance level (t_{17}=2.23, p = 0.04, d' = 0.52). Yet, it should be noted that the participants performed on average 5% better in all conditions of the control experiment compared with performance described above, presumably due to the lack of MRI scanner noise in the control experiment. Thus, also the results from the control experiment suggested that the to-be-ignored dialogues were minimally processed during the visual control task.

3.2. *Attentive processing of audiovisual speech is associated with extensive activation modulation across the brain*

First, we conducted an omnibus univariate four-way repeated measures ANOVA, with the factors Task (AV speech task, visual control task), Semantic Context (coherent, incoherent), Auditory Quality (good, poor) and Visual Quality (good, poor).

Brain regions showing significant modulation of activity in relation to attentive processing of the AV speech (in comparison to the visual control task, i.e., independently from stimulus-level activity) are shown in Fig. 3A (initial cluster threshold, z = 4; permutated cluster significance, p < 0.01). As expected, the AV speech task was associated with enhanced activations (shown in red) in relation to the visual control task in extensive brain networks covering the auditory and visual cortices. Significant effects were also found in medial frontal and parietal regions, in line with our previous study using a similar paradigm (Leminen et al., 2020). However, in corroboration with previous studies on selective listening to auditory (and not audiovisual) speech (Alho et al., 2006; Alho et al., 2003), there were no strong attention-related modulations of dorsolateral prefrontal cortex activity during the audiovisual task, although such activity is typically prominent during selective attention to non-speech sounds (Alho et al., 1999; Degerman et al., 2006; Salmi et al., 2007; Seydell-Greenwald et al., 2014; Tzourio et al., 1997).

As expected, the comparison of the visual control task with the AV speech task was associated with stronger activations (shown in blue in Fig. 3A) in the visual cortex (Alho et al., 1999; Salo et al., 2013). Activations were also seen in the superior parietal lobule and motor cortical regions presumably associated with detection of the visual targets and counting them (Stevens et al., 2000; Yoshiura et al., 1999). The other significant main effects of the omnibus ANOVA are depicted in Supplementary (Suppl.) Fig. 1.

Thereafter, we used multivariate pattern analysis (MVPA) to identify additional brain regions that contain information, which can be used to classify brain activity modulations as belonging to the AV speech task or the visual control task. Classification was performed using support vector machine decoding and leave-one-run-out cross-validation (Hebart et al., 2015). This analysis (Fig. 3B) revealed that it was possible to correctly classify the activity patterns related to either of the two tasks with a remarkable accuracy in a multitude of cortical areas that showed no univariate differences between the tasks (Fig. 3C). For example, dorsolateral, supplementary, premotor, primary motor, and superior parietal cortical regions with no univariate effect showed extensive bilateral decoding accuracies falling between 62% and 80% across participants. Moreover, the average decoding accuracies in certain regions of the visual cortex (approximately the primary and secondary visual cortices) were above 80%, with some voxels showing even 100% decoding accuracies across participants. Importantly, the participants were presented with similar stimulus materials in the two tasks, and thus, the average decoding accuracies depicted in Fig. 3B were independent of stimulus-level manipulations.

Note that small differences in the timing and control of eye movements might partly explain differences in activation patterns between the tasks in the frontal eye field (FEF) and supplementary, premotor and primary motor cortical regions involved in controlling and executing eye-movements. However, we have unpublished data from a sepa-

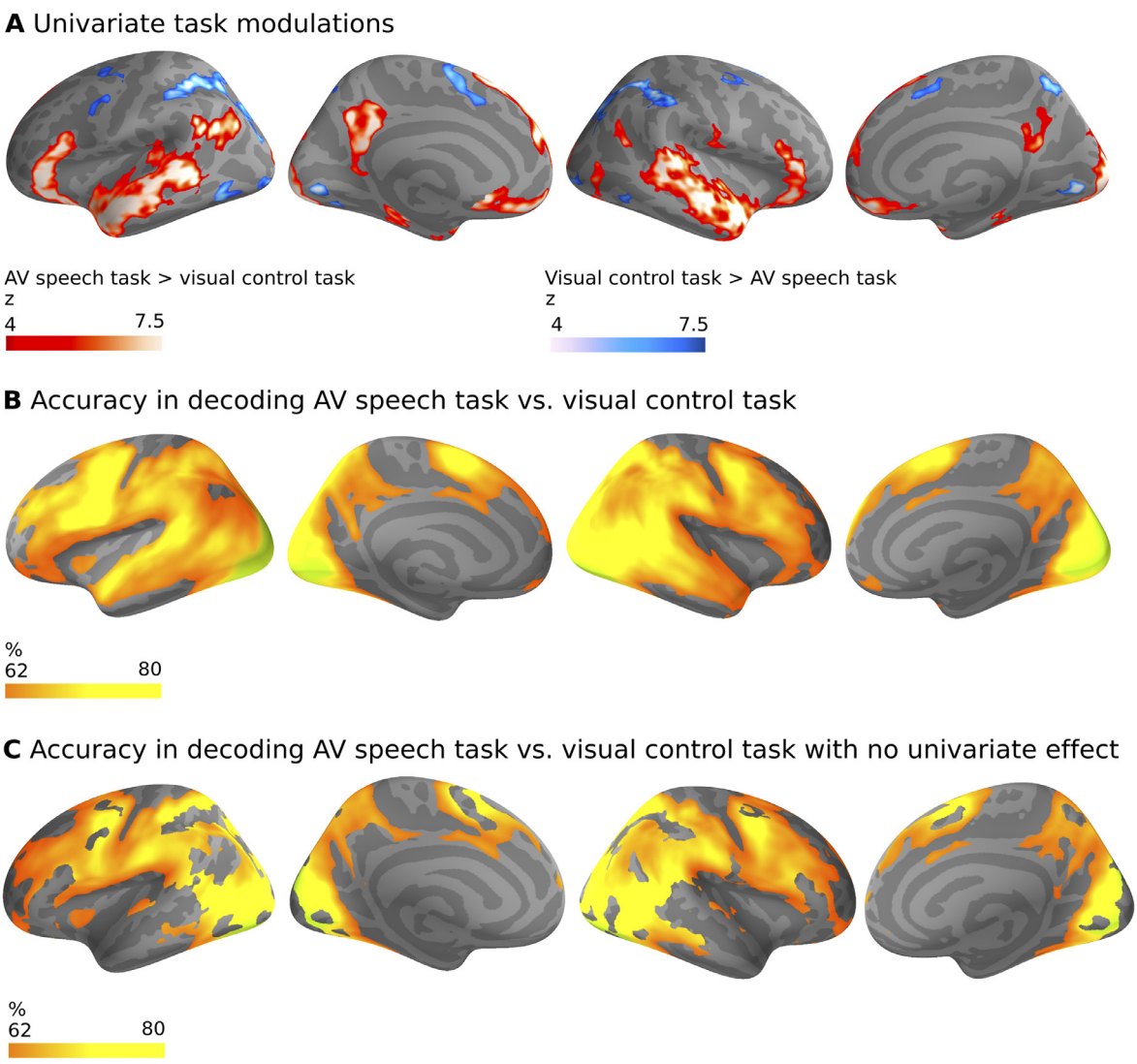


Fig. 3. The AV speech task strongly modulates activations in extensive cortical neural networks. **(A)** Significant clusters (initial cluster threshold z = 4; permutated cluster significance p < 0.01) for the univariate main effect of Task. Red denotes regions where activations were stronger for the AV speech task than for the visual control task, while blue denotes regions showing an opposite effect. **(B)** Results from an MVPA analysis (using Support Vector Machine decoding and a searchlight radius 6 mm) to classify the lines of dialogues as associated with the AV speech task or the visual control task. The average (N = 19) decoding accuracies above 62% are shown. Note that a 60% decoding accuracy is generally considered as a substantial decoding accuracy (Hebart et al., 2015), and all regions that show above 62% decoding accuracy are significantly above chance (z > 2.8). **(C)** Regions without significant univariate effects but with significant multivariate effects.

rate EEG experiment with the same experimental manipulations as the present ones (18 new participants; described in 3.1; to be reported in full later) suggesting that it is unlikely that the two tasks significantly differed in regards to eye-movements. We identified ICA components in the EEG data (128 channels, Brain Vision Analyzer 2) related to horizontal and vertical eye movements and used support vector machine to classify the eye-movement ICAs for each line of the dialogues as belonging to the AV speech task or the visual control task. The mean classification accuracies across subjects were low and non-significant for both of the ICA components: 0.497 (t_{17} = 0.702, p = 0.492, d' = 0.161) and 0.523 (t_{18} = 2.017, p = 0.059, d' = 0.4627).

3.3. The AV speech task modulates connectivity between the STG/STS and frontoparietal regions

As seen in Fig. 3A depicting results from the univariate analysis, the AV speech task was not associated with activity enhancements in frontoparietal regions. In contrast, MVPA (Fig. 3B) revealed that it is possible to classify frontoparietal activity patterns related to the lines of the dialogues as belonging either to the AV speech task or the visual control task with a high accuracy (70–80%). However, MVPA does not elucidate the reason for successful classification. Moreover, since frontal regions may participate in attention-related tasks by changing their functional connectivity with sensory regions (Braun et al., 2015; Davison et al., 2016; Smirnov et al., 2014), we conducted an exploratory psychophysical interaction (PPI) analysis to investigate which brain regions were functionally interacting with the STG/STS during the AV speech task. The first-level analysis was conducted using a model with a psychological regressor (AV speech task), a physiological regressor (mean time series of the significant main effect of Task separately for the left- and right-hemisphere STG/STS clusters), and a PPI regressor (interaction between psychological and physiological regressors) as explanatory variables. As seen in Fig. 4, the results of this PPI analysis (initial cluster threshold, z = 4; permutated cluster significance, p < 0.01) showed no regions with enhanced connectivity with STS/STG regions during the AV speech task. However, dorsolateral prefrontal regions, parietal regions, higher-level visual cortex regions, low-level (core, belt) auditory cortical regions, and multiple medial cortical regions showed decreased

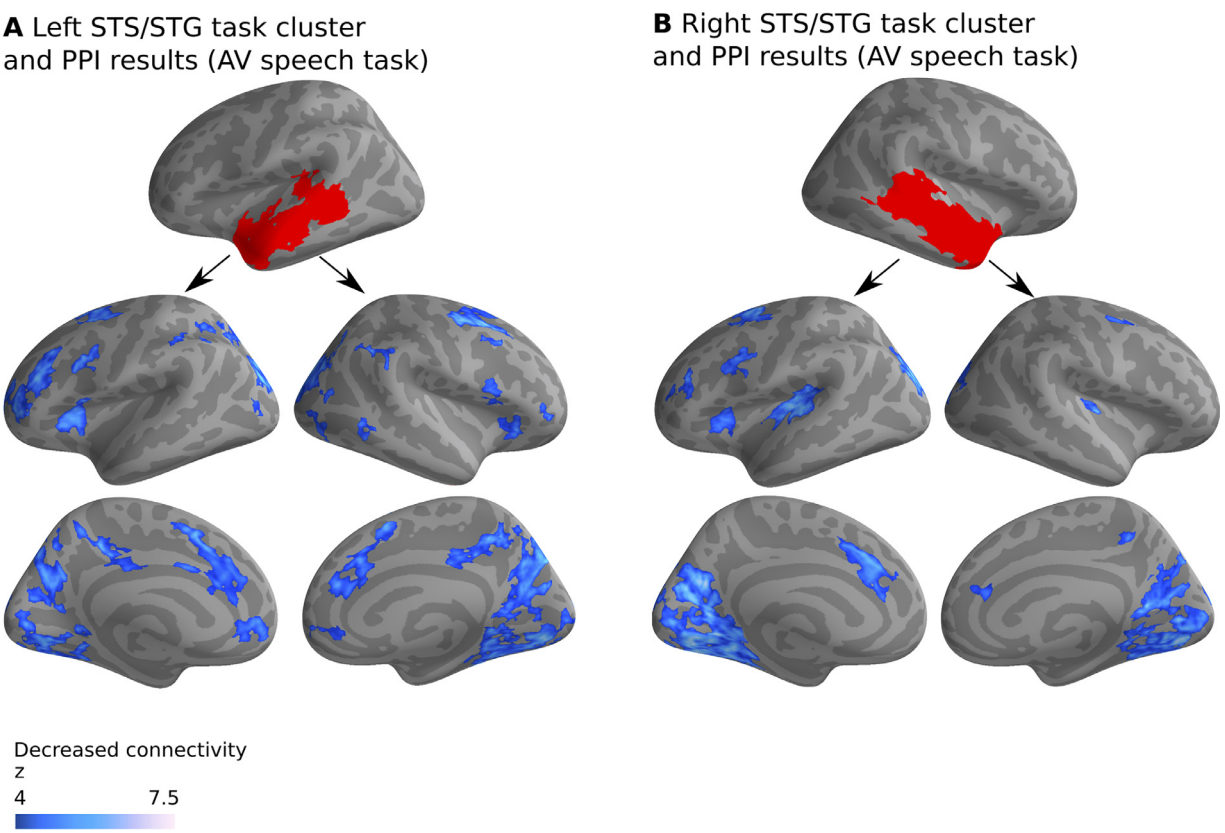


Fig. 4. Psychophysical interactions (PPI) analyses indicate decreased connectivity between the STG/STS and frontoparietal regions during the AV speech task. (A and B, top) The two STG/STS clusters (left, right) where there was a significant univariate task effect (AV speech task > visual control task, see Fig. 3A) were chosen as seed points for the PPI analysis. (A and B, bottom) Significant (initial cluster threshold $z = 4$; permutated cluster significance $p < 0.01$) results from the PPI analyses conducted separately for the left-hemisphere (A) and right-hemisphere (B) STG/STS clusters. No regions showed significantly increased connectivity with STG/STS regions, but a number of regions showed significantly decreased connectivity, including dorsolateral prefrontal, parietal, higher level visual-cortex, low-level (core, belt) auditory-cortex and medial cortical regions.

connectivity with the bilateral STG/STS clusters during the AV speech task. These decreased connectivity patterns were evidently task-specific, as similar results were not found when an analogous PPI analysis was conducted using the activity in the same cluster during the visual control task (Suppl. Fig. 2).

3.4. Modulations related to attentive processing of AV speech are dependent on semantic context and AV quality

In order to study whether attention to the AV speech causes complex dynamic modulation of neuronal responses during the dialogues, we extracted 'attentive AV-speech-processing-related modulations' (AttnAVSMs) by contrasting each line during the AV speech task with the corresponding line during the visual control task. For example, fMRI data for the first line of a dialogue in the AV speech task with a coherent context, good auditory quality and good visual quality was contrasted with fMRI data for the first line in the visual control task of the same run with coherent context, good auditory quality and good visual quality. Because similar stimulus materials (see 2.2–2.4) were presented in the two tasks, differences in line processing depended on the task performed by the participant.

We performed MVPA decoding analyses, with similar methods as above, to classify the different stimulus level manipulations (e.g., good vs. poor auditory quality) based on AttnAVSMs. Since classification analyses take into account both the mean and the variability of the activation in the region studied, significant classification may result simply due to a difference in mean activation (i.e., a simple univariate effect). To highlight the additional information provided by the MVPA, the regions showing only a significant multivariate effect and those showing both a significant univariate and a significant multivariate effect are shown in Fig. 5 using different colours.

As seen in Fig. 5A, Semantic Context was significantly classified using AttnAVSMs (one-sample t-test, initial cluster threshold $z = 2.8$, permutated cluster significance $p < 0.01$) in the left STG/STS. Importantly, in accordance with our pivotal hypothesis, semantic context modulated AttnAVSM patterns in the supratemporal plane including core auditory cortex (Moerel et al., 2014) in Heschl's gyrus (HG).

Auditory Quality, in turn, was significantly classified in bilateral STG/STS regions, with left-hemisphere STG/STS regions also showing an expected univariate effect (Evans et al., 2016; McGettigan et al., 2012); i.e. AttnAVSMs were stronger when the auditory quality of the dialogue was good than when it was poor (Fig 5B, red). AttnAVSMs were also modulated by Auditory Quality in extensive regions of the right frontal cortices. This suggests that although frontal regions do not seem to be automatically strongly activated when performing speech streaming tasks (see Fig. 3A), the level of auditory quality modulates task signals in these regions.

As expected, Visual Quality could be accurately classified based on AttnAVSMs in right-hemisphere STG/STS regions. However, many regions that showed significant decoding accuracies between good and poor visual quality, also showed a univariate difference. That is, visual, frontal, and medial cortical regions all showed both a multivariate decoding effect and a univariate task effect with stronger activations when the AV speech task was performed with poor visual quality compared to good visual quality (Fig. 5C). These findings likely reflect increased effort needed to process bimodal speech with noisy visual inputs.

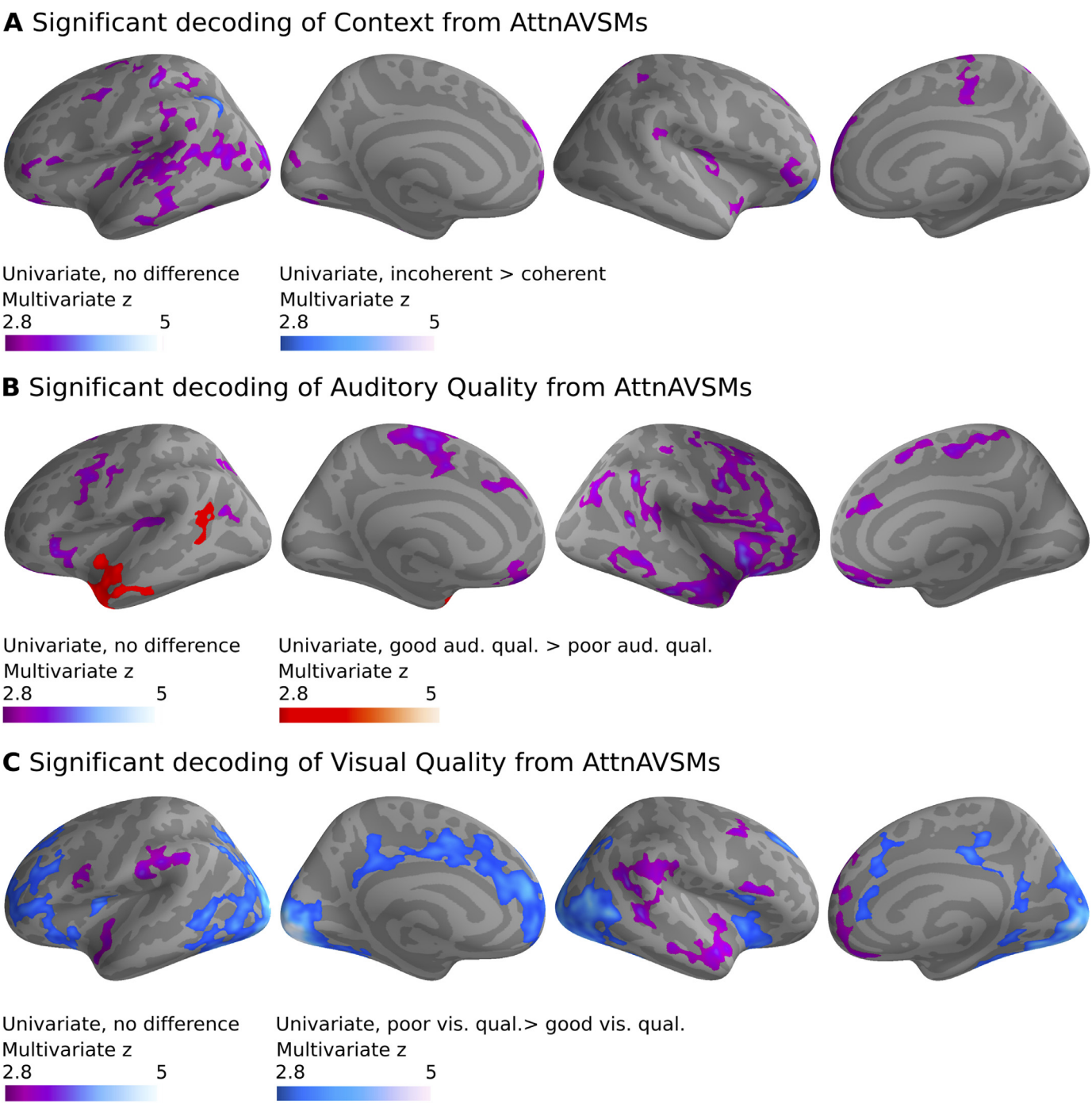


Fig. 5. Semantic Context, Auditory Quality and Visual Quality can be significantly decoded from attentive audiovisual-speech-processing-related modulations (AttnAVSMs). **(A)** Based on AttnAVSMs and using the same MVPA methods as in Fig. 3B-C, we first classified AttnAVSMs as being associated with coherent or incoherent dialogues. The regions where this classification analysis yielded significant decoding accuracy based on a one-sample t-test (initial cluster threshold $z = 2.8$, permutated cluster significance $p < 0.01$) are shown in purple and blue. We also tested whether there was a significant univariate effect in each cluster separately. Two clusters (left parietal, right frontal pole) showed a significant univariate effect, because task modulation signals were stronger for the events with incoherent context than those with coherent context (blue). (B) Auditory Quality (poor vs. good) was classified using the same MVPA method as above. Here two clusters in the left STS region also showed a significant univariate effect (good > poor; red). (C) Classification of Visual Quality (poor vs. good) where many clusters also showed a significant univariate effect (poor > good; blue).

Previous studies (e.g. McGettigan et al., 2012) have suggested that there are complex interactions between auditory and visual quality as well as intelligibility of speech input in the neural processing of speech in various brain regions. Therefore, we also classified the interaction effects of the Semantic Context, Auditory Quality and Visual Quality manipulations. For example, in the Auditory Quality × Visual Quality interaction, we used AttnAVSMs to classify each line as belonging to one of two categories. Category 1 contained all the conditions with good auditory and good visual quality (gagv) and the conditions with poor auditory and poor visual quality (papv). Category 2 contained all the conditions with poor auditory and good visual quality (pagv) and all the conditions with good auditory and poor visual quality (gapv). See also

Suppl. Table 1 for the different combinations of conditions used in each of the interaction analyses. Significant decoding accuracies for the interactions Semantic Context × Auditory Quality, Semantic Context × Visual quality, and Auditory Quality × Visual Quality can be found in Suppl. Fig. 3.

Importantly, significant decoding accuracies for the three-way interaction Semantic Context × Auditory Quality × Visual Quality were found, amongst other regions, in bilateral HG (Fig. 6A). Because the location of HG in normalized anatomical space shows significant interindividual variation (Kang et al., 2004), we wanted to confirm that the cluster centred on left-hemisphere HG in Fig. 6A was indeed due to signals originating from HG and not from adjacent regions. To this end, we defined HG individually on the cortical surfaces (see 2.7.6) of each of the 19 participants. HG was thereafter divided into a medial and a lateral region. As can be seen in Fig. 6B (top), the two subdivisions of HG showed substantial overlap across participants and were centred on HG in the mean cortical surface. We performed the same classification analysis as in Fig. 6A separately in these two regions of interest (ROIs). We found that multivariate classification was significant in the medial left HG ($t_{18} = 2.4$, familywise-error rate corrected $p = 0.03$) approximately where the primary auditory cortex putatively is located (Moerel et al., 2014), but not in the lateral left HG (Fig. 6B).

To explore possible explanations for the interaction between Semantic Context, Auditory Quality, and Visual Quality in HG, we conducted pairwise classification between each of the eight AttnAVSM levels (i.e., combination of the three factors, see Fig. 6B, lower). Thereafter, we tested which of the four pairs with the two levels of Semantic Context (coherent vs. incoherent) resulted in decoding accuracy significantly above chance levels. The only pair to show significant classification accuracy of Semantic Context was the pair with good auditory and visual quality. This suggests that the interaction effect was related to the fact that AttnAVSMs in HG were mostly modulated by semantic context when the auditory and visual qualities were good.

3.5. Different brain networks are revealed based on temporal changes in AttnAVSMs

Next, we evaluated temporal modulation of AttnAVSMs over the timecourse of the dialogue using polynomial trends, that is, linear, quadratic or linear-quadratic trends across the dialog lines. We only included the first six lines in each dialogue, because the seventh was immediately followed by the questions related to the video (see 2.4), which might confound the fit of our models (e.g., because of the visual presentation of the questions).

All three analyses yielded extensive significant clusters, but the best fit was found for the linear-quadratic models, and therefore the linear and quadratic models are not reported. That is, in most cortical regions, AttnAVSMs did neither monotonically decrease nor increase, but instead generally showed an inverted u-shape profile (see Fig. 7 and Suppl. Fig 4). We also tested whether the linear-quadratic task modulations interacted with Semantic Context, Auditory Quality, and Visual Quality (the same comparisons were not made for the linear and quadratic trends for brevity and to constrain the amount of effects studied). The results of these analyses are shown in the Suppl. Fig. 4.

Next, we explored these clusters in more detail. We used the significant clusters from our linear-quadratic trend analyses (initial cluster threshold, $z = 4$; permutated cluster significance, $p < 0.01$; see 2.7.7 for details) to define our ROIs for multidimensional scaling analyses (see Fig. 7, bottom, left, for the brain regions associated with the labels, and Suppl. Table 2 for the label names). First, we performed a dissimilarity analysis (see 2.7.7) on the extracted average AttnAVSMs in each of the ROIs (Suppl. Fig. 5). To visualize these data, we projected the dissimilarity matrix on a 2D plane using multidimensional scaling (Fig. 7, top, right). We also calculated linkages (see 2.7.7.) between the ROIs based on the dissimilarity matrix that were visualized with a dendrogram (Fig. 7 top, left).

As can be seen in the dendrogram (Fig. 7 top), the similarity between the regions forms a hierarchical structure. On the first level of the hierarchy, two branches could be identified, the right-hemisphere temporoparietal junction (TPJ) which differed from all other brain structures. On the second level of the hierarchy, five clusters could be identified, starting from the bottom: (1) a TPJ cluster, (2) a cluster of occipital regions, (3) a cluster of regions previously associated with processing of conceptual information (e.g., Binder et al., 2000), (4) a large cluster of regions related to sensory and control processes in speech processing (Friederici, 2011), and (5) a large cluster of frontoparietal and medial brain regions. For further inspection, we chose to threshold the dendrogram using an intermediate height threshold (dashed grey line in Fig. 7, top, left). This yielded six networks, with two or more anatomical regions, and 14 networks with only one anatomical region. For the sake of brevity, we chose the six networks for further inspection, excluding the 14 single-region networks. The anatomical locations and labels for these six brain networks are shown in Figure 7 (bottom, left).

The average AttnAVSM profiles are plotted for an exemplar ROI from each of the six networks identified. The first network ('Secondary control network'; blue) consists of frontal and medial regions. In this network, the AttnAVSMs first rise up to the third line and thereafter plateau until the end of the dialogue. A similar profile was found for the second network ('Sensomotor control network'; green) comprising adjacent superior frontal regions. The bilateral visual cortices comprised their own sub-network ('Visual network'; purple). Here AttnAVSMs were initially strong, monotonically decreasing after the second line. A slightly different profile was found for a network consisting of premotor regions and the bilateral inferior frontal gyri ('Primary control network'; orange). Here the AttnAVSMs were initially strong, but started to steadily decrease half-way into the dialogue. The most central network ('Multidimensional speech network'; red), consisted of regions previously associated with low-level auditory processing of speech (Friederici, 2011), such as the auditory cortex and STG/STS, but also regions associated with somato-motor processing. Here AttnAVSMs first steadily increased but started to decrease after the third line. The last and smallest network ('Conceptual network'; light green) consisted of the left anterior middle temporal gyrus and orbitofrontal cortex. This network showed a similar temporal profile as the Multidimensional speech network.

4. Discussion

In line with previous studies (e.g., Alho et al., 2006; Alho et al., 2003; Leminen et al., 2020), we observed that attentive processing of AV speech strongly modulates neural activity in STG/STS as well as in other regions, including the frontal and visual cortical areas. Specifically, we found that it is possible to classify activity patterns as belonging to the AV speech or the visual control task with a classification accuracy falling between 80 and 100% across participants in many regions. This is a substantially higher accuracy than previously reported for fMRI multivariate classification studies of selective attention (Bonte et al., 2009; Haynes and Rees, 2005; Häkkinen and Rinne, 2018; Kamitani and Tong, 2005, 2006; Rosenberg et al., 2015). The high accuracy is probably due to the complex lifelike AV stimuli that are more variable than simple ones (e.g., tones), and therefore cause less neural adaptation (Richter et al., 2018). Also, lifelike AV speech may engage participants more than less naturalistic tasks. Additionally, we found that semantic context modulated attentional processing of speech in the STG/STS as expected, but, interestingly, also in the belt and core regions of the auditory cortex, supporting the notion that semantic information influences low-level auditory processing. Thus, the present results provide us with substantial evidence for resolving the long-standing debate between the early and late selection theories of attention. Lastly, attentional modulations of brain activity were not uniform during the course of the dialogue. Instead, a novel combination of polynomial trend analysis and multidimensional scaling revealed that attentional modulation has distinct temporal profiles in different brain regions. Taken together,

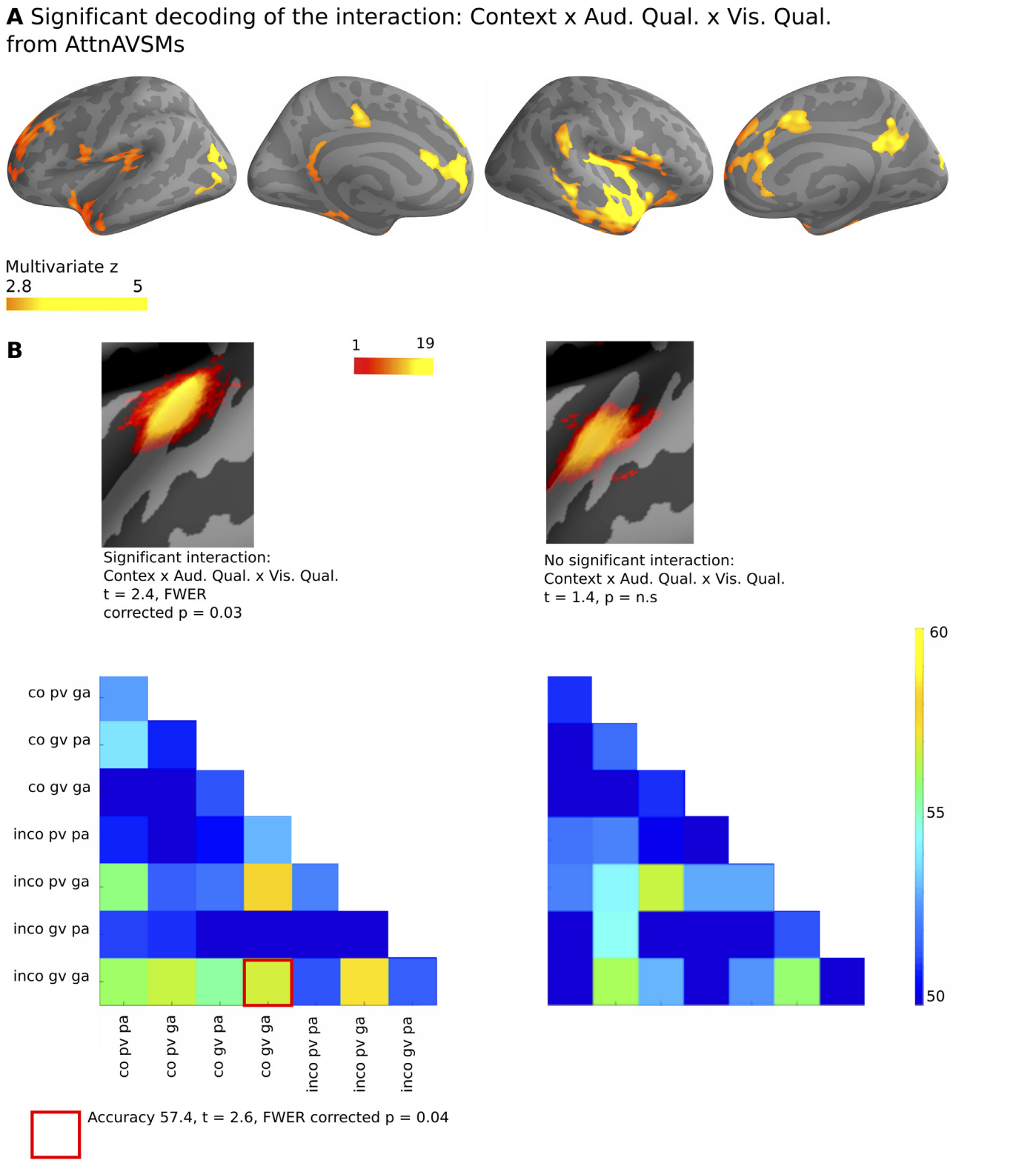


Fig. 6. The interaction effect between Semantic Context, Auditory Quality and Visual Quality could be decoded with significant accuracy using AttnAVSMs. (A) The interaction between Semantic Context, Auditory Quality and Visual Quality was decoded with significant accuracy (one-sample t-tests, initial cluster threshold $z = 2.8$, permutated cluster significance $p < 0.01$) in, amongst other regions, bilateral HG (i.e., putative core/belt auditory cortex) based on AttnAVSMs. (B, top) Overlap (yellow) of the extracted medial and lateral part of HG in the left hemisphere across all 19 participants. The same interaction effect as in (A) was decoded in the two HG ROIs in each participant's anatomically defined ROI. The medial part of HG showed a significant effect, while the lateral part did not. (B, bottom) Pairwise decoding for each of the conditions revealed that the interaction effect was related to the fact that in HG, the semantic context effect resulted in significant classification accuracy only in conditions when both the auditory and visual quality was good. This combination is highlighted in the classification matrix with a red square. Abbreviations: co, coherent; inco, incoherent, pv, poor visual quality; gv, good visual quality; pa, poor auditory quality; ga, good auditory quality; FWER, Familywise-error rate.

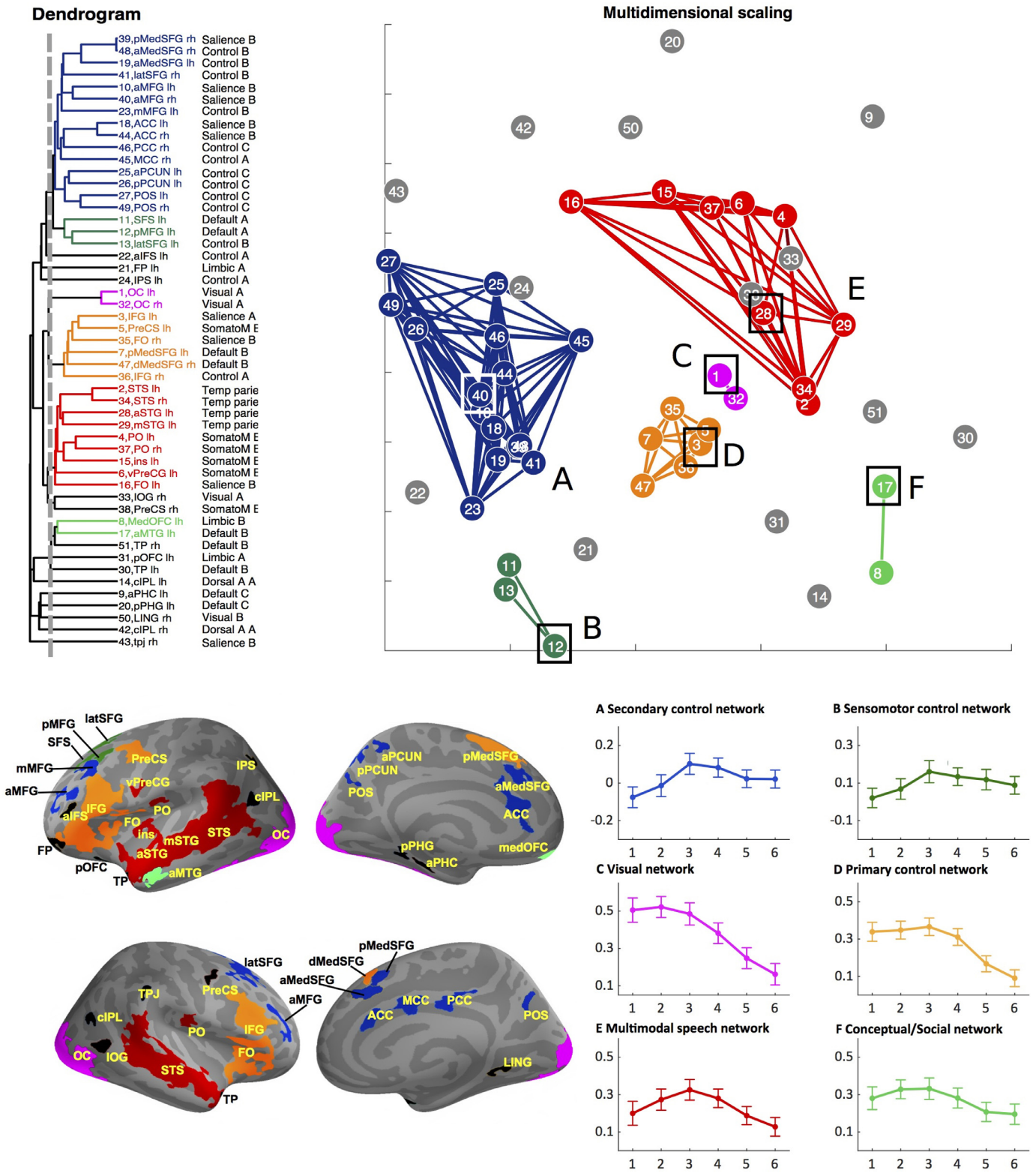


Fig. 7. Multidimensional scaling based on AttnAVSMs reveals six brain networks. We used the clusters identified by our linear-quadratic trend analyses (see Suppl. Fig. 4) as ROIs for a subsequent multidimensional scaling analysis. First, we calculated dissimilarities for each ROI-ROI pair (Suppl. Fig. 5). Next, we calculated the linkage between each ROI based on the dissimilarity matrix. Top: The linkages are visualized both using a dendrogram and multidimensional scaling. Based on the dendrogram, we used an intermediate threshold (grey dashed line), which yielded six sub-networks with two or more regions (14 regions comprised their own cluster). In the dendrogram, we also show the YEO17 (Yeo et al., 2011) brain networks containing each ROI. All regions belonging to our first network are control/salience regions in YEO17, but the remaining five networks show some divergence from the YEO17. Bottom, left: The anatomical locations of these 6 different networks are visualized with anatomical labels. Bottom, right: The temporal profiles in an exemplar region from each of the networks (see top, A–F), with the order based on the similarity distance between the networks. Abbreviations can be found in Supplementary Table 1. Error bars denote ± standard errors of the mean.

our results support a new conceptualization of attention-related modulations, where attention not just simply increases the gain and fidelity of neuronal responses in the sensory cortex, but attentional modulations instead arise as neuronal networks strive to solve complex tasks in demanding environments.

There is a common conception that selective attention modulates stimulus processing in sensory cortices through fairly simple non-specific mechanisms that increase the gain and fidelity of neuronal responses to task-relevant stimuli (Briggs et al., 2013). Yet, in the current study we found that activity patterns related to attentive AV speech processing (AttnAVSMs) in the auditory cortex and its vicinity were modulated by both the semantic coherence and the AV quality of the video stimuli. Accordingly, it is important to note that these activity pattern modulations were not explained by simple activity enhancement or decrease during the different conditions, but instead semantics and AV quality modulated how the information structure in the auditory cortex was utilized during attentive processing of AV speech. Thus, our results suggest that attention-related modulation in the sensory cortex is much more task- and context-specific than previously assumed, corroborating observations that attention-related modulation in the sensory cortex is fine-tuned to offer the most efficient neural coding strategies for the task at hand (Maunsell and Treue, 2006; Navalpakkam and Itti, 2007). The current study supports the emerging view that attention-related modulations of neural activity patterns arise as neuronal networks strive to solve and/or learn how to solve specific tasks in different contexts (Angeloni and Geffen, 2018; Kilgard, 2012; Scheich et al., 2007; Scheich et al., 2011).

There is a long-standing debate concerning the role of semantics in modulating low-level neural processing of speech (Broderick et al., 2019). While neurophysiological studies show that the primary auditory cortex is modulated by speech intelligibility (Luo and Poeppel, 2007; Peelle et al., 2013; Sohoglu et al., 2012; Tuennerhoff and Noppeney, 2016; Wild et al., 2012), evidence of semantic modulations has remained elusive (with a few exceptions, Broderick et al., 2019; de Heer et al., 2017). In the present study, we manipulated semantic coherence at the sentence level. Surprisingly, these semantic manipulations were associated with modulation of neuronal processes in left-lateralized STG/STS regions including HG, providing evidence that in certain situations semantic content in fact modulates low-level speech processing. From a predictive coding framework (Kragel et al., 2018; Näätänen et al., 2001; Schroeger et al., 2015; Winkler et al., 2009), one could argue that this semantic influence at the low-level auditory cortex reflects predictive feedback from higher-level semantic networks.

Interestingly, significant semantic modulation of HG AttnAVSMs was limited to the conditions with good AV quality (Fig. 6). This unexpected result could be interpreted to suggest that the semantic modulation of attentive speech processing in HG was related to the increased effort needed to make sense of the input when successive lines were unrelated to each other. However, several observations argue against a simple task difficulty explanation. First, note that there were similar performance differences related to AV quality (Fig 2, left), but AV quality was not associated with strong modulations of AttnAVSMs in the supratemporal plane. Second, previous studies have consistently shown that general task difficulty does not simply enhance activations in auditory cortex. For example, during auditory discrimination tasks, task difficulty does not modulate STG activation (Harinen et al., 2013; Harinen and Rinne, 2013; Häkkinen et al., 2015; Rinne et al., 2009; Rinne et al., 2012), while in auditory working memory tasks, a higher memory load causes decreased, not increased STG activation (Harinen et al., 2013; Harinen and Rinne, 2013; Häkkinen et al., 2015; Rinne et al., 2009; Rinne et al., 2012; Wikman and Rinne, 2018). Thus, is it possible that the semantic effect in the auditory cortex is related to higher demands on working memory when the content of the dialogue was semantically unrelated? We also find this unlikely as in the current study the semantic content changed AttnAVSM activation patterns in the auditory cortex, without causing general decrease or increase in absolute activations in the auditory cortex, while previous studies (see above) have shown that increasing auditory working memory load attenuates activations in the auditory cortex (also outside the primary auditory areas). It should be noted, however, that the univariate main effect of semantic coherence (observed across tasks; see Suppl. Fig. 1 a) might be related to decreased working memory load for the semantically coherent dialogues. Also, it is important to note that poor audiovisual quality of speech might also increase working memory load (which also partly might explain the memory-task performance in our task, see Fig. 2). Yet, neither good auditory nor good visual quality increased activations in primary auditory regions.

Instead of the simple task difficulty explanation for the semantic modulations of HG activity there are at least three other post hoc alternatives to explain these results. First, using the framework presented above for attention-related modulations, we suggest that the information structure or changes in the information structure of core auditory neurons might be used to help with semantic uncertainty when the AV speech is clear and one can see the faces of the speakers. However, when the AV quality is poor, the low intelligibility of speech input cannot be used to enhance comprehension, leading to recruitment of alternative brain networks and neuronal strategies to solve the task instead of HG. Second, it is possible that the comprehension of speech input is supported more strongly by low-level auditory regions only when the intelligibility of the speech input is optimal, that is, with good AV speech quality and semantic context. However, we want to underline that speech intelligibility is also influenced by auditory and visual quality (see the behavioral results in Fig 2). Yet, these factors did not strongly modulate AttnAVSMs in the auditory cortex and were associated with distinct and different higher-order cortical activation patterns than the semantic influences (Fig. 3A–C). Thus, if the semantic influence on AttnAVMs is driven through its effect on speech intelligibility, it is probably caused by distinct mechanisms from the supposedly bottom-up effects on speech intelligibility caused by audiovisual quality. Third, it may be that the general conception of primary auditory cortex as a simple, low-level sound-analyser (Scheich et al., 2007) might be flawed. For example, recent studies in mice have suggested that there are neurons or neuronal assemblies in the primary auditory cortex that store long-term auditory engrams (Weinberger, 2011). Therefore, as humans are a considerably more complex species than mice, it is not inconceivable to envision that semantic information, at least in a rudimentary form, might be processed at the level of the primary auditory cortex. Such semantic effects could be driven by neural connections formed between higher-level semantic regions (such as the temporal pole) and the auditory cortex when learning semantic associations (see for example, Patterson et al., 2007). Such 'pointers' can be used to reactivate the whole memory trace, including representations stored in the auditory cortex (Cowan, 2008).

In our previous study, we found that attentive AV speech processing was associated with activity modulations in the medial frontal cortex and particularly in the orbitofrontal cortex (Leminen et al., 2020). The ad hoc explanation for these activation enhancements was that they were related to the social or emotional nature of the listening task (Buckley et al., 2009). To resolve the function of orbitofrontal AV speech task-related activations we reasoned that mixing the lines from different dialogues would destroy the social context and thus reduce activity in brain regions specifically involved in social cognition. However, the level of semantic coherence in the dialogues did not strongly modulate AttnAVSMs in the orbitofrontal cortex in the current study (Fig. 5A). We nevertheless replicated the result of our previous study, as attentive processing of AV speech modulated orbitofrontal activity. Furthermore, both auditory and especially visual quality (Fig. 5B–C), that is, the amount of social sensory information, modulated orbitofrontal AttnAVSMs. We therefore suggest that the AV presentation of speech (in contrast to auditory-only presentation, Alho et al., 2006; Alho et al., 2003) enhances social engagement of participants regardless of semantic coherence per se, resulting in the recruitment of orbitofrontal cortex. This is further supported by the fact that our multidimensional scaling

and dendrography analyses (Fig. 7) found similarities between the activity pattern in the orbitofrontal cortex with other nodes previously suggested to be involved in the processing of social/conceptual information (Binder et al., 2009).

One perplexing finding in fMRI studies on speech-related attention is the lack of frontal activation in such tasks. Although the frontal cortex is routinely activated in studies on selective attention (Gazzaley and Nobre, 2012) including selective auditory attention that use simple auditory stimuli (Alho et al., 1999; Degerman et al., 2006; Salmi et al., 2007; Seydell-Greenwald et al., 2014; Tzourio et al., 1997), selective attention to speech does not strongly activate the frontal cortex (Alho et al., 2006; Alho et al., 2003; Evans et al., 2016; Leminen et al., 2020). We suggest three possible explanations for the lack of strong frontal activations in AV speech tasks: 1) The degree of frontal involvement in different tasks may depend on task automaticity (Chein and Schneider, 2012), with highly automatised tasks, like listening to speech in noise, recruiting mainly sensory networks, while novel tasks recruit the prefrontal cortex (Chein and Schneider, 2012). Thus, although the present AV speech task was deemed very demanding by the participants even in the easiest condition (coherent dialogue, good AV quality), it was not associated with strong activation modulations in prefrontal regions (Fig 3A). In contrast, AV quality modulated AttnAVSMs in dorsolateral prefrontal regions presumably because listening to noise-vocoded speech with blurred faces is a more novel task (Cole et al., 2016). 2) The prefrontal cortex might not be involved in perceptual tasks through changes in its absolute activation per se, but rather by changing its connectivity with sensory regions (Braun et al., 2015; Davison et al., 2016; Smirnov et al., 2014). The results of our PPI analysis (Fig. 3) supports this notion, since the connectivity between speech-related sensory areas and the frontal cortex decreased during the AV speech task, but not during the visual control task. This pattern might be due to certain frontal neurons connecting to inhibitory interneurons in the sensory speech processing areas. Thus, activation of frontal neurons during the AV speech task would cause transient inhibition in sensory regions, leading to an overall decrease in functional connectivity between the two regions. Such connections have been suggested to reorganize the pattern of activity in local neuronal assemblies in a task-dependent manner (Braun et al., 2015). 3) The decreased functional connectivity between speech sensory regions and frontal regions during the AV speech task might also reflect that the two regions are performing entirely different types of neural computations during the AV speech task. In other words, functions in these two regions become less related to each other than during rest, causing a desynchronization during the AV speech task. Further studies using for example TMS during AV speech tasks could be used to resolve these three possible theories.

Lastly, the present study shows that it is important not to consider attention-related modulations as static, but rather as processes that change over time in a complex manner. The present combination of linear-quadratic trend analyses, multidimensional scaling and dendrography revealed that AttnAVSMs changed during the dialogue with distinct temporal profiles in different clusters (Fig. 7). We propose that the central network for performing the AV speech task was the network comprising speech sensory and sensomotor regions (Multimodal speech network). In this network, the attention-related modulations were initially weak, gradually increasing to the middle of the dialogue, and then gradually decreasing back to the initial levels (Fig. 7). One possibility is that these temporal changes were due to task performance becoming more automatic and less effortful during AV speech. Kilgard (2012) has suggested a Darwinian framework to explain similar perplexing findings related to plasticity and changes in sensory cortex related to complex cognitive/perceptual tasks. According to Kilgard, when the task is initially learned, all possible neuronal networks that might be useful to solve the task at hand are recruited. Gradually, the unnecessary, less informative neuronal networks are pruned out, and the most efficient network ends up performing the task (sparse coding). Thus, it might be that the temporal profiles of AttnAVSMs in the Multimodal speech network reflect such dynamics in a micro-time scale. That is, each new dialogue could be considered as a new opportunity for neurons in the speech-sensitive networks to refine their performance in the AV speech task, seen as a gradual increase and thereafter decrease in recruited neuronal resources in this network. The activity profiles in the remaining five networks support this notion: The Visual network and the Primary control network showed different profiles from the Multimodal speech network. In these two networks, AttnAVSMs peaked two lines before the Multimodal speech network. This might indicate that these two networks act as primary controllers of the Multimodal speech network, providing strategic top-down control and contextual information. This kind of control in the Multimodal speech network would be critical in the beginning of the AV speech task when the task is not yet automatic in this network. The Secondary and Sensomotor control networks showed an opposite time profile to the Primary control and Visual networks. Here AttnAVSMs did not emerge until halfway into the dialogue. These Secondary and Sensomotor control networks are related to computations of a higher abstraction level, such as task strategies (Chein and Schneider, 2012). Thus, we propose that in the beginning of the dialogue, the AV speech task networks (the Multimodal speech network and the Visual and Primary control networks) are first recruited automatically, and then towards the end of the dialogue, the task performance gradually automatises and the amount of neural resources recruited in these networks wanes. This enables resources to be allocated to neural networks that are critical for forming new task performance strategies, that is, the Secondary and Sensomotor control networks. Therefore, these frontal and medial networks are activated towards the end of the dialogue. Such strategies could henceforth be used to optimize task performance in later task occurrences. To explore whether our current data give support for this notion, we compared AttnAVSM profiles in the left STG/STS region (Suppl. Fig. 4, red) for the first run (little optimization) to those of the third run (optimized performance). Indeed, there was a non-significant trend for stronger AttnAVSMs in the left STG/STS during the first run compared to the third run ($F_{1,18} = 3.6$, $p = 0.07$, $\eta^2 = 0.168$; Suppl. Fig. 6A). This is probably not related to fatigue or decreased effort towards the end of the experiment as there was a slight (non-significant) increase in the accuracy of behavioral performance in the third run compared with the first ($F_{1,18} = 1.51$, $p = 0.23$, $\eta^2 = 0.08$; Suppl. Fig. 6B). It should be noted, however, that we did not design the current experiment to compare data from different runs. Therefore we hope that these results will inspire later studies regarding temporal changes in task effects on speech-related brain activity.

Another possible explanation for the temporal profiles in the Secondary and Sensomotor control networks is that many regions in these networks are part of the so-called default mode network (see also Fig. 7 for YEO17 networks). These regions have been implemented in, for example, attention towards internal representations and non-task related thoughts (Zvyagintsev et al., 2013). Thus, as the present AV listening task became more automatic towards the end of each dialogue, the participants might have, for example, started to monitor their own behavior more strongly.

5. Conclusions

In conclusion, the present results suggest that attentive processing of AV speech in a cocktail-party-like setting is associated with distinct modulation of neuronal responses in both sensory and other cortical regions that follow predictable temporal profiles. Further, our results reveal semantic influence on attentional selection of speech sounds in regions associated with low level speech processing (already at HG) – a finding, which despite great efforts, has been hard to verify in previous studies. Together, our results support emerging notions that speech processing in the auditory cortex can be modulated by high-level semantic information and that task-specificity, context, and temporal dynamics influence how, when, and whether attention modulates the neural responses to sound. The exact micro scale neural dynamics that underlie the perfor-

mance of abstract speech tasks, however, remain unresolved. Therefore, computational and neuronal level work that takes the abstract dynamics of attentive processing of speech into account will ultimately help us to understand the complex interplay between sensory, frontal control, and other cortical networks underlying the remarkable human capability of understanding speech in noisy "cocktail-party" settings – an ability that we often take for granted.

Declaration of Competing Interest

The authors declare no competing interests.

CRediT authorship contribution statement

Patrik Wikman: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Visualization. **Elisa Sahari:** Data curation, Investigation, Visualization. **Viljami Salmela:** Formal analysis, Methodology, Visualization. **Alina Leminen:** Supervision. **Miika Leminen:** Software. **Matti Laine:** Funding acquisition, Supervision. **Kimmo Alho:** Funding acquisition, Project administration, Resources, Supervision.

Acknowledgments

This work is supported by the Academy of Finland (grant #297848, "Modulations of brain activity patterns during selective attention to speech", 2016–2020). The founder had no role in the conceptualization, design, data collection, analysis, decision to publish, or preparation of the manuscript.. Additionally, authors want to thank Prof. David L. Woods and Dr. Satu Saalasti for comments.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.neuroimage.2020.117365.

References

Ahissar, E., Nagarajan, S., Ahissar, M., Protopapas, A., Mahncke, H., Merzenich, M.M., 2001. Speech comprehension is correlated with temporal response patterns recorded from auditory cortex. PNAS 98, 13367–13372.
Alho, K., Medvedev, S.V., Pakhomov, S.V., Roudas, M.S., Tervaniemi, M., Reinikainen, K., Zeffiro, T., Näätänen, R., 1999. Selective tuning of the left and right auditory cortices during spatially directed attention. Cognitive Brain Res. 7, 335–341.
Alho, K., Vorobyev, V.A., Meduedev, S.V., Pakhomov, S.V., Starchenko, M.G., Tervaniemi, M., Näätanen, R., 2006. Selective attention to human voice enhances brain activity bilaterally in the superior temporal sulcus. Brain Res. 1075, 142–150.
Alho, K., Vorobyev, V.A., Medvedev, S.V., Pakhomov, S.V., Roudas, M.S., Tervaniemi, M., van Zuijen, T., Näätanen, R., 2003. Hemispheric lateralization of cerebral blood-flow changes during selective listening to dichotically presented continuous speech. Cognitive Brain Res. 17, 201–211.
Angeloni, C., Geffen, M.N., 2018. Contextual modulation of sound processing in the auditory cortex. Curr. Opin. Neurobiol. 49, 8–15.
Binder, J.R., Desai, R.H., Graves, W.W., Conant, L.L., 2009. Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. Cereb. Cortex 19, 2767–2796.
Binder, J.R., Frost, J.A., Hammeke, T.A., Bellgowan, P.S.F., Springer, J.A., Kaufman, J.N., Possing, E.T., 2000. Human temporal lobe activation by speech and nonspeech sounds. Cereb. Cortex 10, 512–528.
Boersma, P., Weenink, D., 2001. Praat Speech Processing Software. Institute of Phonetics Sciences of the University of Amsterdam http://www.praat. org.
Bonte, M., Valente, G., Formisano, E., 2009. Dynamic and task-dependent encoding of speech and voice by phase reorganization of cortical oscillations. J. Neurosci. 29, 1699–1706.
Braun, U., Schafer, A., Walter, H., Erk, S., Romanczuk-Seiferth, N., Haddad, L., Schweiger, J.I., Grimm, O., Heinz, A., Tost, H., Meyer-Lindenberg, A., Bassett, D.S., 2015. Dynamic reconfiguration of frontal brain networks during executive cognition in humans. PNAS 112, 11678–11683.
Briggs, F., Mangun, G.R., Usrey, W.M., 2013. Attention enhances synaptic efficacy and the signal-to-noise ratio in neural circuits. Nature 499, 476–480.
Broadbent, D.E., 1954. The role of auditory localization in attention and memory span. J. Exp. Psychol. 47, 191–196.
Broderick, M.P., Anderson, A.J., Lalor, E.C., 2019. Semantic context enhances the early auditory encoding of natural speech. J. Neurosci. 39, 7564–7575.
Buckley, M.J., Mansouri, F.A., Hoda, H., Mahboubi, M., Browning, P.G.F., Kwok, S.C., Phillips, A., Tanaka, K., 2009. Dissociable components of rule-guided behavior depend on distinct medial and prefrontal regions. Science 325, 52–58.
Chang, C.-C., Lin, C.-J., 2011. LIBSVM: a library for support vector machines. ACM TIST 2, 1–27.
Chein, J.M., Schneider, W., 2012. The brain's learning and control architecture. Curr. Dir. Psychol. Sci. 21, 78–84.
Cherry, E.C., 1953. Some experiments on the recognition of speech, with one and with two ears. J. Acoust. Soc. Am. 25, 975–979.
Cole, M.W., Ito, T., Braver, T.S., 2016. The behavioral relevance of task information in human prefrontal cortex. Cereb. Cortex 26, 2497–2505.
Connolly, J.F., Service, E., D'Arcy, R.C.N., Kujala, A., Alho, K., 2001. Phonological aspects of word recognition as revealed by high resolution spatio-temporal brain mapping. Neuroreport 12, 237–243.
Cortes, C., Vapnik, V., 1995. Support-vector networks. Machine Learn. 20, 273–297.
Cowan, N., 2008. What are the differences between long-term, short-term, and working memory? Prog. Brain Res. 169, 323–338.
Davison, E.N., Turner, B.O., Schlesinger, K.J., Miller, M.B., Grafton, S.T., Bassett, D.S., Carlson, J.M., 2016. Individual differences in dynamic functional brain connectivity across the human lifespan. PLoS Comput. Biol. 12.
de Heer, W.A., Huth, A.G., Griffiths, T.L., Gallant, J.L., Theunissen, F.E., 2017. The hierarchical cortical organization of human speech processing. J. Neurosci. 37, 6539–6557.
Degerman, A., Rinne, T., Salmi, J., Salonen, O., Alho, K., 2006. Selective attention to sound location or pitch studied with fMRI. Brain Res. 1077, 123–134.
Deutsch, J.A., Deutsch, D., 1963. Attention - some theoretical considerations. Psychol. Rev. 70, 80–90.
Ding, N., Chatterjee, M., Simon, J.Z., 2014. Robust cortical entrainment to the speech envelope relies on the spectro-temporal fine structure. Neuroimage 88, 41–46.
Ephrat, A., Mosseri, I., Lang, O., Dekel, T., Wilson, K., Hassidim, A., Freeman, W.T., Rubinstein, M., 2018. Looking to listen at the cocktail party: a speaker-independent audio-visual model for speech separation. Acm Trans. Graphics 37.
Evans, S., McGettigan, C., Agnew, Z.K., Rosen, S., Scott, S.K., 2016. Getting the cocktail party started: masking effects in speech perception. J. Cogn. Neurosci. 28, 483–500.
Fischl, B., 2012. FreeSurfer. Neuroimage 62, 774–781.
Friederici, A.D., 2011. The brain basis of lanuage processing: from structure to function. Physiol. Review. 91, 1357–1392.
Friederici, A.D., Gierhan, S.M.E., 2013. The language network. Curr. Opin. Neurobiol. 23, 250–254.
Gazzaley, A., Nobre, A.C., 2012. Top-down modulation: bridging selective attention and working memory. Trends Cogn. Sci. 16, 129–135.
Golumbic, E.M.Z., Ding, N., Bickel, S., Lakatos, P., Schevon, C.A., McKhann, G.M., Goodman, R.R., Emerson, R., Mehta, A.D., Simon, J.Z., Poeppel, D., Schroeder, C.E., 2013. Mechanisms underlying selective neuronal tracking of attended speech at a "cocktail party". Neuron 77, 980–991.
Greve, D.N., Fischl, B., 2009. Accurate and robust brain image alignment using boundary-based registration. Neuroimage 48, 63–72.
Häkkinen, S., Ovaska, N., Rinne, T., 2015. Processing of pitch and location in human auditory cortex during visual and auditory tasks. Front. Psychol. 6, 1–12.
Häkkinen, S., Rinne, T., 2018. Intrinsic, stimulus-driven and task-dependent connectivity in human auditory cortex. Brain Struct. Funct. 223, 2113–2127.
Hansen, J.C., Hillyard, S.A., 1988. Temporal dynamics of human auditory selective attention. Psychophysiology 25, 316–329.
Harinen, K., Aaltonen, O., Salo, E., Salonen, O., Rinne, T., 2013. Task-dependent activations of human auditory cortex to prototypical and nonprototypical vowels. Hum. Brain Mapp. 34, 1272–1281.
Harinen, K., Rinne, T., 2013. Activations of human auditory cortex to phonemic and nonphonemic vowels during discrimination and memory tasks. Neuroimage 77, 279–287.
Haynes, J.D., Rees, G., 2005. Predicting the stream of consciousness from activity in human visual cortex. Curr. Biol. 15, 1301–1307.
Hebart, M.N., Gorgen, K., Haynes, J.D., 2015. The Decoding Toolbox (TDT): a versatile software package for multivariate analyses of functional imaging data. Front. Neuroinformatics 8.
Jenkinson, M., Beckmann, C.F., Behrens, T.E., Woolrich, M.W., Smith, S.M., 2012. Fsl. Neuroimage 62, 782–790.
Kamitani, Y., Tong, F., 2005. Decoding the visual and subjective contents of the human brain. Nat. Neurosci. 8, 679–685.
Kamitani, Y., Tong, F., 2006. Decoding seen and attended motion directions from activity in the human visual cortex. Curr. Biol. 16, 1096–1102.
Kang, X., Bertrand, O., Alho, K., Yund, E.W., Herron, T.J., Woods, D.L., 2004. Local landmark-based mapping of human auditory cortex. Neuroimage 22, 1657–1670.
Kilgard, M.P., 2012. Harnessing plasticity to understand learning and treat disease. Trends Neurosci. 35, 715–722.
Kong, Y.-Y., Somarowthu, A., Ding, N., 2015. Effects of spectral degradation on attentional modulation of cortical auditory responses to continuous speech. J. Assoc. Res. Otolaryngol. 16, 783–796.
Kragel, P.A., Koban, L., Barrett, L.F., Wager, T.D., 2018. Representation, Pattern Information, and Brain Signatures: From Neurons to Neuroimaging. Neuron 99, 257–273.
Kriegeskorte, N., Goebel, R., Bandettini, P., 2006. Information-based functional brain mapping. PNAS 103, 3863–3868.
Kutas, M., Hillyard, S.A., 1980. Reading Senseless Sentences - Brain Potentials Reflect Semantic Incongruity. Science 207, 203–205.
Lau, E.F., Phillips, C., Poeppel, D., 2008. A cortical network for semantics: (de)constructing the N400. Nat. Rev. Neurosci. 9, 920–933.
Leminen, A., Verwoert, M., Moisala, M., Salmela, V., Wikman, P., Alho, K., 2020. Modulation of brain activity by selective attention to audiovisual dialogues. Front. Neurosci. 14, 1–16.

Lewis, J.L., 1970. Semantic Processing of Unattended Messages Using Dichotic Listening. J. Exp. Psychol. 85, 225–228.
Luo, H., Poeppel, D., 2007. Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex. Neuron 54, 1001–1010.
Maunsell, J.H.R., Treue, S., 2006. Feature-based attention in visual cortex. Trends Neurosci. 29, 317–322.
McGettigan, C., Faulkner, A., Altarelli, I., Obleser, J., Baverstock, H., Scott, S.K., 2012. Speech comprehension aided by multiple modalities: Behavioural and neural interactions. Neuropsychologia 50, 762–776.
Mesgarani, N., Chang, E.F., 2012. Selective cortical representation of attended speaker in multi-talker speech perception. Nature 485, 233–236.
Moerel, M., De Martino, F., Formisano, E., 2014. An anatomical and functional topography of human auditory cortical areas. Frontiers in Neuroscience 8.
Näätänen, R., 1990. The role of attention in auditory information processing as revealed by event-related potentials and other brain measures of cognitive function. Behav. Brain Sci. 13, 201–233.
Näätänen, R., Teder, W., Alho, K., Lavikainen, J., 1992. Auditory attention and selective input modulation: a topographical ERP study. Neuroreport 3, 493–496.
Näätänen, R., Tervaniemi, M., Sussman, E., Paavilainen, P., Winkler, I., 2001. 'Primitive intelligence'in the auditory cortex. Trends Neurosci. 24, 283–288.
Navalpakkam, V., Itti, L., 2007. Search goal tunes visual features optimally. Neuron 53, 605–617.
Oldfield, R.C., 1971. The assessment and analysis of handedness: the Edinburgh inventory. Neuropsychologia 9, 97–113.
O'Sullivan, J., Herrero, J., Smith, E., Schevon, C., McKhann, G.M., Sheth, S.A., Mehta, A.D., Mesgarani, N., 2019. Hierarchical Encoding of Attended Auditory Objects in Multi-talker Speech Perception. Neuron 104, 980–991.
O'Sullivan, J.A., Power, A.J., Mesgarani, N., Rajaram, S., Foxe, J.J., Shinn-Cunningham, B.G., Slaney, M., Shamma, S.A., Lalor, E.C., 2015. Attentional Selection in a Cocktail Party Environment Can Be Decoded from Single-Trial EEG. Cereb. Cortex 25, 1697–1706.
Patterson, K., Nestor, P.J., Rogers, T.T., 2007. Where do you know what you know? The representation of semantic knowledge in the human brain. Nat. Rev. Neurosci. 8, 976–987.
Peelle, J.E., Gross, J., Davis, M.H., 2013. Phase-locked responses to speech in human auditory cortex are enhanced during comprehension. Cereb. Cortex 23, 1378–1387.
Richter, D., Ekman, M., de Lange, F.P., 2018. Suppressed Sensory Response to Predictable Object Stimuli throughout the Ventral Visual Stream. J. Neurosci. 38, 7452–7461.
Riecke, L., Snipes, S., van Bree, S., Kaas, A., Hausfeld, L., 2019. Audio-tactile enhancement of cortical speech-envelope tracking. Neuroimage 202, 116134.
Rinne, T., Koistinen, S., Salonen, O., Alho, K., 2009. Task-dependent activations of human auditory cortex during pitch discrimination and pitch memory tasks. J. Neurosci. 29, 133–138.
Rinne, T., Koistinen, S., Talja, S., Wikman, P., Salonen, O., 2012. Task-dependent activations of human auditory cortex during spatial discrimination and spatial memory tasks. Neuroimage 4126–4131.
Rosenberg, M.D., Finn, E.S., Constable, R.T., Chun, M.M., 2015. Predicting moment-to–moment attentional state. Neuroimage 114, 249–256.
Rutten, S., Santoro, R., Hervais-Adelman, A., Formisano, E., Golestani, N., 2019. Cortical encoding of speech enhances task-relevant acoustic information. Nature Human Behav. 3, 1125 -1125.
Salmi, J., Rinne, T., Degerman, A., Salonen, O., Alho, K., 2007. Orienting and maintenance of spatial attention in audition and vision: multimodal and modality-specific brain activations. Brain Struct. Funct. 212.
Salo, E., Rinne, T., Salonen, O., Alho, K., 2013. Brain activity during auditory and visual phonological, spatial and simple discrimination tasks. Brain Res. 1496, 55–69.
Scheich, H., Brechmann, A., Brosch, M., Budinger, E., Ohl, F.W., 2007. The cognitive auditory cortex: task-specificity of stimulus representations. Hear. Res. 229, 213–224.
Scheich, H., Brechmann, A., Brosch, M., Budinger, E., Ohl, F.W., Selezneva, E., Stark, H., Tischmeyer, W., Wetzel, W., 2011. Behavioral semantics of learning and crossmodal processing in auditory cortex: the semantic processor concept. Hear. Res. 271, 3–15.
Schroeger, E., Marzecova, A., SanMiguel, I., 2015. Attention and prediction in human audition: a lesson from cognitive psychophysiology. Eur. J. Neurosci. 41, 641–664.
Seydell-Greenwald, A., Greenberg, A.S., Rauschecker, J.P., 2014. Are you listening? Brain activation associated with sustained nonspatial auditory attention in the presence and absence of stimulation. Hum. Brain Mapp. 35, 2233–2252.
Smirnov, D., Glerean, E., Lahnakoski, J.M., Salmi, J., Jaaskelainen, I.P., Sams, M., Nummenmaa, L., 2014. Fronto-parietal network supports context-dependent speech comprehension. Neuropsychologia 63, 293–303.
Sohoglu, E., Peelle, J.E., Carlyon, R.P., Davis, M.H., 2012. Predictive top-down integration of prior knowledge during speech perception. J. Neurosci. 32, 8443–8453.
Stevens, A.A., Skudlarski, P., Gatenby, J.C., Gore, J.C., 2000. Event-related fMRI of auditory and visual oddball tasks. Magn. Reson. Imaging 18, 495–502.
Sumby, W.H., Pollack, I., 1954. Visual contribution to speech intelligibility in noise. J. Acoust. Soc. Am. 26, 212–215.
Treisman, A., Squire, R., Green, J., 1974. Semantic processing in dichotic listening? A replication. Mem. Cognit. 2, 641–646.
Tuennerhoff, J., Noppeney, U., 2016. When sentences live up to your expectations. Neuroimage 124, 641–653.
Tylen, K., Christensen, P., Roepstorff, A., Lund, T., Ostergaard, S., Donald, M., 2015. Brains striving for coherence: Long-term cumulative plot formation in the default mode network. Neuroimage 121, 106–114.
Tzourio, N., El Massioui, F., Crivello, F., Joliot, M., Renault, B., Mazoyer, B., 1997. Functional anatomy of human auditory attention studied with PET. Neuroimage 5, 63–77.
Weinberger, N.M., 2011. Reconceptualizing the primary auditory cortex: learning, memory and specific plasticity. In: The Auditory Cortex, pp. 465–491.
Wikman, P., Rinne, T., 2018. Interaction of the effects associated with auditory-motor integration and attention-engaging listening tasks. Neuropsychologia.
Wild, C.J., Davis, M.H., Johnsrude, I.S., 2012. Human auditory cortex is sensitive to the perceived clarity of speech. Neuroimage 60, 1490–1502.
Winkler, I., Denham, S.L., Nelken, I., 2009. Modeling the auditory scene: predictive regularity representations and perceptual objects. Trends Cogn. Sci. 13, 532–540.
Yeo, B.T.T., Krienen, F.M., Sepulcre, J., Sabuncu, M.R., Lashkari, D., Hollinshead, M., Roffman, J.L., Smoller, J.W., Zoller, L., Polimeni, J.R., Fischl, B., Liu, H.S., Buckner, R.L., 2011. The organization of the human cerebral cortex estimated by intrinsic functional connectivity. J. Neurophysiol. 106, 1125–1165.
Yoshiura, T., Zhong, J., Shibata, D.K., Kwok, W.E., Shrier, D.A., Numaguchi, Y., 1999. Functional MRI study of auditory and visual oddball tasks. Neuroreport 10, 1683–1688.
Zion Golumbic, E., Cogan, G.B., Schroeder, C.E., Poeppel, D., 2013. Visual input enhances selective speech envelope tracking in auditory cortex at a "cocktail party". J. Neurosci. 33, 1417–1426.
Zvyagintsev, M., Clemens, B., Chechko, N., Mathiak, K.A., Sack, A.T., Mathiak, K., 2013. Brain networks underlying mental imagery of auditory and visual information. Eur. J. Neurosci. 37, 1421–1434.