Ylinen, Artturi; Wikman, Patrik; Leminen, Miika; Alho, Kimmo

Task-dependent cortical activations during selective attention to audiovisual speech

Published in:
Brain Research

DOI:
10.1016/j.brainres.2021.147739

Published: 15/01/2022

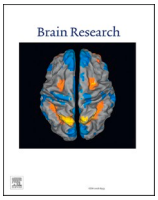
Document Version
Publisher's PDF, also known as Version of record

Brain Research 1775 (2022) 147739

Contents lists available at ScienceDirect

Brain Research

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



Task-dependent cortical activations during selective attention to audiovisual speech

Artturi Ylinen [a,*], Patrik Wikman [a,b], Miika Leminen [c], Kimmo Alho [a,d]

[a] Department of Psychology and Logopedics, University of Helsinki, Helsinki, Finland
[b] Department of Neuroscience, Georgetown University, Washington D.C., USA
[c] Analytics and Data Services, HUS Helsinki University Hospital, Helsinki, Finland
[d] Advanced Magnetic Imaging Centre, Aalto NeuroImaging, Aalto University, Espoo, Finland

ARTICLE INFO

Keywords:
Selective attention to speech
Audiovisual speech processing
Task-dependent effects
Functional magnetic resonance imaging

ABSTRACT

Selective listening to speech depends on widespread networks of the brain, but how the involvement of different neural systems in speech processing is affected by factors such as the task performed by a listener and speech intelligibility remains poorly understood. We used functional magnetic resonance imaging to systematically examine the effects that performing different tasks has on neural activations during selective attention to continuous audiovisual speech in the presence of task-irrelevant speech. Participants viewed audiovisual dialogues and attended either to the semantic or the phonological content of speech, or ignored speech altogether and performed a visual control task. The tasks were factorially combined with good and poor auditory and visual speech qualities. Selective attention to speech engaged superior temporal regions and the left inferior frontal gyrus regardless of the task. Frontoparietal regions implicated in selective auditory attention to simple sounds (e.g., tones, syllables) were not engaged by the semantic task, suggesting that this network may not be not as crucial when attending to continuous speech. The medial orbitofrontal cortex, implicated in social cognition, was most activated by the semantic task. Activity levels during the phonological task in the left prefrontal, premotor, and secondary somatosensory regions had a distinct temporal profile as well as the highest overall activity, possibly relating to the role of the dorsal speech processing stream in sub-lexical processing. Our results demonstrate that the task type influences neural activations during selective attention to speech, and emphasize the importance of ecologically valid experimental designs.

1. Introduction

The human ability to focus on one speech stream among many might seem like a mundane skill, but its actualization relies upon complex neurocognitive processing (Shinn-Cunningham, 2008; Wikman et al., 2021). How listeners solve the so-called cocktail party problem has been investigated ever since the classic studies of Cherry (1953), and more recent research has also begun to determine the neural basis of this phenomenon. Electrophysiological experiments have revealed that the neural processing of attended and unattended speech streams is differentiated already in the non-primary auditory cortex (AC) in the supratemporal plane extending to the superior temporal gyrus (STG), as the activation of neuronal populations becomes tuned to the attended speech stream (Mesgarani and Chang, 2012; O'Sullivan et al., 2019; Zion Golumbic et al., 2013; see also Teder et al., 1993; Woods et al., 1984). However, the functional organization of cortical auditory processing is complex already at the AC; studies have shown that activations in this region are profoundly influenced by, for example, attention and the task performed by a listener (e.g., pitch discrimination or pitch memory tasks; Ahveninen et al., 2006; Petkov et al., 2004; Scheich et al., 2007; Wikman and Rinne, 2019; Wikman et al., 2015). These observations highlight the importance of considering how the choice of an experimental task may affect the results of studies on speech processing. However, to our knowledge, how different attention-engaging tasks affect activations in or outside the AC during selective attention to continuous speech has not been systematically studied before.

Task-dependent effects in the auditory processing of relatively simple stimuli, such as tones, phonemes, and syllables, have been experimentally demonstrated by varying the task performed by a listener while keeping stimulus type constant. For example, in a study by Harinen and

* Corresponding author.
E-mail address: artturi.ylinen@helsinki.fi (A. Ylinen).

https://doi.org/10.1016/j.brainres.2021.147739
Received 14 June 2021; Received in revised form 21 October 2021; Accepted 21 November 2021
Available online 26 November 2021

Rinne (2013), a vowel discrimination task enhanced activation in STG regions involved in phonetic and phonological processing (Leonard and Chang, 2014; Mesgarani et al., 2014; Turkeltaub and Branch, 2010), while a memory task performed on the same stimuli enhanced activation in more lateral and posterior STG regions and the inferior parietal lobule (IPL). Moreover, increasing memory load in the memory task was associated with deactivation in portions of the STG and superior temporal sulcus (STS), even though the stimulus type remained constant. Other experiments have corroborated and complemented these results, demonstrating that the organization of cortical processing of sounds depends strongly on the kind of task performed (Ahveninen et al., 2006; Alho et al., 2014; Rinne et al., 2009; Wikman and Rinne, 2019; Wikman et al., 2015).

Whereas the experiments described above focused on the processing of relatively simple sounds, the choice of task and stimulus type may also be crucial in studies on the processing of continuous speech (see, e.g., Price, 2012). Experiments with simple stimuli and artificial but well-controlled tasks provide invaluable information on the different levels of speech processing (see, e.g., Davis and Johnsrude, 2003), but concerns have also been raised that artificial tasks and stimuli may engage neurocognitive mechanisms that are not essential in more naturalistic settings (Hickok and Poeppel, 2007). For instance, one long lasting debate concerns whether the motor regions involved in speech production are also involved in speech perception (Arsenault and Buchsbaum, 2016; Hickok and Poeppel, 2007; Liberman and Mattingly, 1985; Möttönen and Watkins, 2009; Mugler et al., 2018; Pulvermüller et al., 2014; Schomers and Pulvermüller, 2016). These motor areas form part of the so-called dorsal stream of speech processing, flowing from the posterior STG/STS to the temporoparietal cortex and further to the motor and prefrontal areas. The dorsal stream has been suggested to be involved in mapping between auditory and motor speech information (Fridriksson et al., 2016; Hickok and Poeppel, 2007; Rauschecker and Scott, 2009), and the role of these regions in speech perception has been demonstrated when listening to relatively simple speech stimuli (e.g., single syllables; Fadiga et al., 2002; Möttönen et al., 2013; Möttönen and Watkins, 2009; Murakami et al., 2015; Pulvermuller et al., 2006). However, it has also been suggested that these regions might mainly be recruited when participants focus on the sub-lexical aspects of speech, questioning whether these results can be generalized to more ecologically valid situations (Hickok and Poeppel, 2007). One reason for this suggestion is that patients with lesions in these brain regions may show deficits in tasks where sub-lexical processing is required, but fare well when listening to normal, continuous speech (Hickok and Poeppel, 2004). It has also been suggested that the motor regions in the dorsal processing stream may participate in speech perception in a dynamic and compensatory manner, even when listening to continuous speech, when speech intelligibility is poor because of, for example, overlaid noise (Du et al., 2014; Osnes et al., 2011; Wild et al., 2012). Taken together, the debate surrounding the role of these dorsal stream regions in speech perception suggests that the brain may process speech in a highly context-dependent fashion.

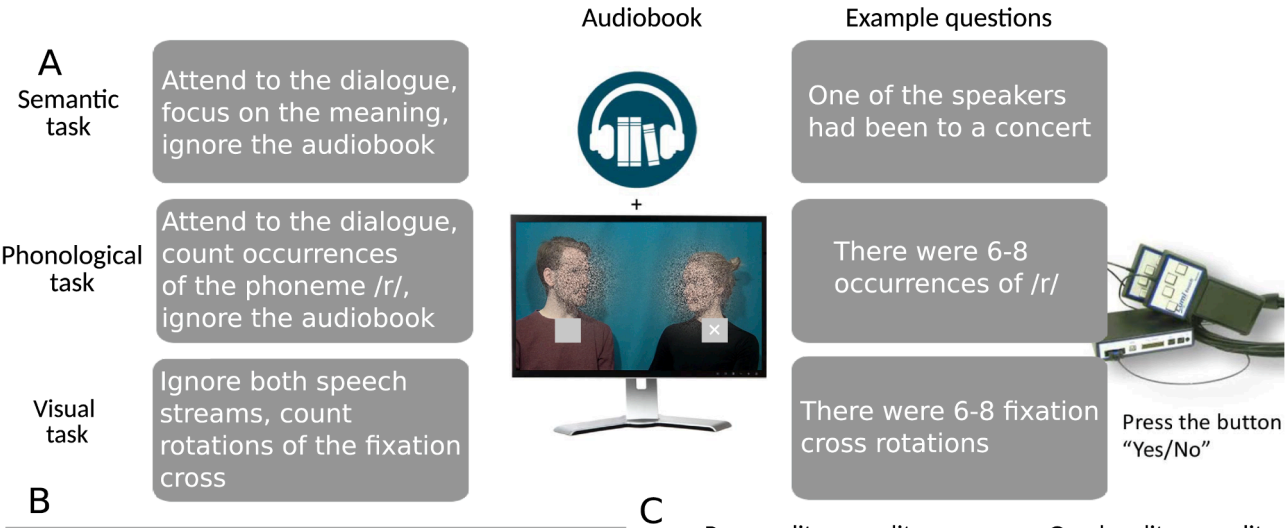
An additional consideration on task-dependent effects on speech processing stems from the fact that widely distributed networks of the brain are recruited to process different aspects of speech (Binder et al., 2009; Friederici and Gierhan, 2013; Huth et al., 2016; Vigneau et al., 2011; Vigneau et al., 2006), but simple artificial settings are unlikely to capture all the processes that are relevant in more naturalistic ones. For example, processing the social aspects of speech (e.g., emotion, talker relations) is essential in everyday situations, but obviously absent from studies with stimuli that lack any social or even semantic meaning. Recent fMRI experiments with relatively naturalistic audiovisual dialogues have found speech processing to engage the medial orbitofrontal cortex (OFC; Leminen et al., 2020; Wikman et al., 2021), which has been suggested to play a role in social cognition (e.g., during tasks evoking empathy or demanding moral judgments or theory of mind; Adolphs, 2009; Alcalá-López et al., 2019; Bzdok et al., 2013). However, how this region contributes to the processing of speech remains poorly understood.

Real-life speech processing situations routinely demand selective attention, as environmental noises or speech from irrelevant speakers are often present in the auditory background. A large portion of studies on the neural basis of selective auditory attention have been conducted with relatively simple sounds, such as tones and single syllables. Selectively attending to these simple sounds has been associated with activity in the dorsolateral prefrontal cortex (DLPFC) and parietal regions (Alho et al., 1999; Degerman et al., 2006; Salmi et al., 2007; Zatorre et al., 1999), that is, canonical attention-related brain networks (Alho et al., 2014; Näätänen, 1990). Accordingly, it has been suggested that when selectively attending to sounds, the frontoparietal regions contribute to maintaining attention, and facilitate the processing of attended sounds in the sensory systems (e.g., Alho et al., 1999; Näätänen, 1990). Somewhat surprisingly, however, in studies on selective attention to continuous speech with distracting speech in the background (i.e., listening to speech in a 'cocktail party situation'), these frontoparietal activations have not been as consistently found. Instead, selective attention to continuous speech has mainly been associated with enhanced activity in the STG, STS, and left inferior frontal gyrus (LIFG; Alho et al., 2006, 2003; Leminen et al., 2020; Wikman et al., 2021), regions that are implicated also in the processing of speech in quiet (i.e., when selective attention is not needed; Friederici and Gierhan, 2013). This pattern of results seems to hold even though the task may not be subjectively especially easy (see, e.g., Leminen et al., 2020; Wikman et al., 2021), and even though speech perception performance in general is often reduced when noise or other speakers appear in the background, as compared to situations where these distractors are not present (see, e.g., Miller, 1947; Sumby and Pollack, 1954; Treisman, 1964). Thus it seems that when it comes to selective auditory attention, results from experiments conducted with simple stimuli do not necessarily generalize to more naturalistic settings. It has been suggested that the lack of frontoparietal activations in more naturalistic experimental settings might be due to over-learning and automatization of processing (Alho et al., 2006; Wikman et al., 2021). The reasoning behind this suggestion is that selectively listening to one speech stream among many is something we do daily, and extensive practice diminishes the need for cognitive control in task performance (Chein and Schneider, 2012).

While frontoparietal contributions may not be crucial in maintaining selective attention to continuous speech, evidence suggests that they may be dynamically involved in guiding processing at the initial stages of 'tuning in' to a speaker or speakers (i.e., directing attention to speech), or when orienting to the task at hand (Hill and Miller, 2010; Näätänen, 1990). For example, our recent study (Wikman et al., 2021) examined temporal modulations in activity levels across the time course of an audiovisual dialogue, and found that a network involving portions of the bilateral inferior frontal gyrus and precentral gyrus (PreCG) were activated mainly at the beginning of a dialogue. In the STG and STS, activation was observed to increase for the first half of a dialogue, and then decrease toward the end of the dialogue. These results are consistent with the idea that the frontal regions may be dynamically recruited in guiding processing in a goal-dependent manner, and that their contribution diminishes after processing becomes more independent in the sensory regions (Wikman et al., 2021).

The present experiment was designed to study how the brain mechanisms of selective attention to continuous audiovisual speech depend on the task performed by a listener. Specifically, we wanted to examine how performing a relatively naturalistic speech listening task compares to situations where it is the phonological content of speech that is focused on (as is often done in studies on the processing of sub-lexical speech stimuli), as well as to situations where speech is not attended to. Moreover, we wanted to study the time course of activations in these different tasks, as evidence suggests that the involvement of different brain regions varies dynamically throughout the time course of task performance (Hill and Miller, 2010; Näätänen, 1990; Wikman et al.,

2021). To meet these goals, we used functional magnetic resonance imaging (fMRI) to monitor neural activations while participants viewed audiovisually presented dialogues, in which concurrent task-irrelevant speech was always present in the background. Participants performed three different tasks (Fig. 1): In the Semantic (S) task, the participants focused on the dialogues and responded to questions regarding the semantic content of the conversation. This task was aimed to mimic naturalistic speech processing as closely as possible while also controlling for performance. In the Phonological (P) task, the participants selectively attended to the dialogues, monitored the speech stream for occurrences of the phoneme /r/, and reported their number. This task was intended to require selective attention and the processing of sub-lexical aspects of speech, but not the meaning of speech. In the Visual (V) control task, the participants were instructed to ignore the speakers altogether, focus on a fixation cross below the speaker's face, and count how many times the cross rotated. This task was intended to control for stimulus-related effects while speech is not attended to (cf. Leminen et al., 2020; Wikman et al., 2021). To examine possible interactions



B

Example lines from a dialogue	Example questions
A: I bought a new phone and it has so many features that I am completely lost.	One of the speakers had a new phone (True).
B: Oh yes, you used to have that ancient phone, which was not even a smart phone.	The speaker mentioned his or her watch (False)
A: Yes, and it was a perfectly good phone until my cat dropped it on the floor from a table, and that was it.	The pet of one of the speakers had broken the phone (True)

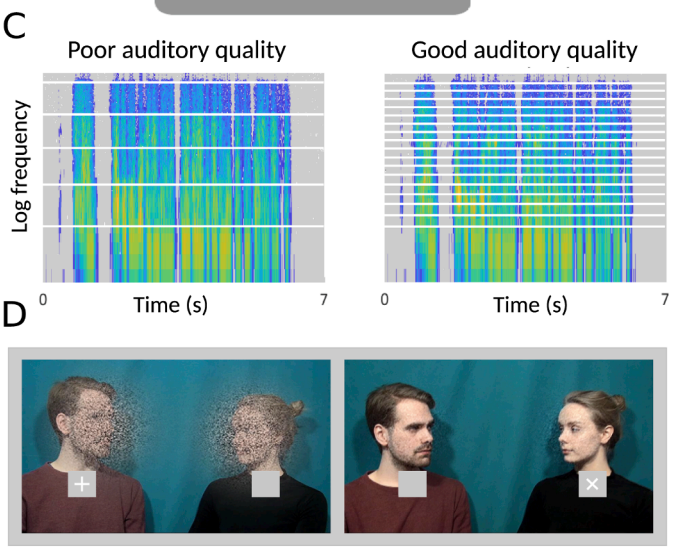


Fig. 1. The experimental setting of the present study. (A) Participants viewed audiovisual dialogues of a male and female speaker discussing neutral everyday topics, while a concurrent speech stream from an audiobook was played in the background. Three different tasks were performed by the participants: 1) In the Semantic (S) task, the participants selectively attended to the dialogue, ignored the audiobook (and the cross below the speaker's face), and, after the dialogue, answered seven questions related to its semantic content. 2) In the Phonological (P) task, the participants selectively attended to the dialogue, searched the speech stream for occurrences of the phoneme /r/, and ignored the audiobook (and the cross). 3) In the Visual (V) task, the participants ignored the speech streams altogether and counted the number of rotations of the fixation crosses below the faces of the speakers. In the P and V tasks, after the presentation of the dialogue video, the participants reported the number of occurrences of /r/:s or fixation cross rotations by answering 'yes' or 'no' to seven statements of the form *'There were x-y /r/:s' or 'There were x-y fixation cross rotations'*. The number of task-relevant occurrences was matched between the P and V tasks (i.e., /r/:s and cross rotations). (B) Example lines and S task questions from one of the dialogues (see Supplementary Table 2 for an example of a full dialogue). (C) The voices of the two attended speakers were manipulated using noise-vocoding (Shannon et al., 1995) on two levels. In the poor auditory quality, frequencies above 0.3 kHz were noise-vocoded with four logarithmically spaced frequency bands shown with the white lines, resulting in speech with poor intelligibility (note that the fundamental frequencies of the speakers were not noise-vocoded). In the good auditory quality, frequencies above 0.3 kHz were noise-vocoded on 16 logarithmically spaced frequency bands, maintaining good intelligibility (see Section 5.2 for details). (D) Visual speech intelligibility was modulated on two levels by masking the faces of the speakers with more or less white noise.

between task and speech intelligibility, the tasks were factorially combined with good or poor auditory and visual speech qualities (Fig. 1).

It was hypothesized that the S task would recruit regions associated with selective attention to speech, such as the STG, STS, and LIFG, as well as the medial orbitofrontal cortex, a region implicated in social cognition (see Leminen et al., 2020; Wikman et al., 2021). The P task was expected to recruit partially overlapping areas with the S task in the STG, STS, and LIFG, as it, too, involves selective attention to speech. However, since the P task requires sub-lexical processing of speech, we expected it to recruit the AC regions processing phonological information (i.e., the anterior and posterior STG) more strongly than either of the other tasks. Moreover, following the dual-stream model of Hickok and Poeppel (2007, 2004), we expected regions in the dorsal stream of speech processing (i.e., left temporoparietal, premotor, and inferior prefrontal areas) to show most activity during the P task. As these dorsal stream regions have also been previously implicated in the processing of degraded speech, we further expected these regions to show higher activation in conditions with poor speech intelligibility during the S and P tasks, but not during the V task (i.e., an interaction between task and the audiovisual qualities). Regarding activation modulations across the time course of a dialogue, we expected to replicate our previous results (Wikman et al., 2021) in that the beginnings of the dialogues would be associated with the highest activity in the bilateral IFG and precentral areas during the S and P tasks, and activity in the sensory auditory regions in the STS/STG would first increase for the first half of the dialogue, and then decrease.

2. Results

2.1. Behavioral performance

After each dialogue video, the participants responded to questions related to the task they had performed during that dialogue. In the S task, performance was measured as the percentage of correct responses (Fig. 2, left). In the P and V tasks, performance was measured as the distance of the participant's answer from the correct answer (Fig. 2, middle and right; detailed descriptive statistics on task performance are reported in Supplementary Table 1).

To study task performance in conditions with varying speech intelligibility, separate 2 × 2 repeated-measures ANOVAs with factors Auditory Quality and Visual Quality were conducted for the S and P tasks. Differences in behavioral performance across the audiovisual qualities in the V task were assessed with the non-parametric Friedman test, because the data did not meet the assumption of normality of model residuals made by the repeated measures ANOVA. In the S and V tasks, no significant effects of audiovisual quality on task performance were found (Fig. 2; S task: $F(1,17) < 1.0$, $p > .36$ for all effects; V task: $\chi^2(3) = 1.5$, $p > .68$). In the P task, a significant main effect of Auditory Quality was found ($F(1,17) = 18.9$, $p < .001$, $\eta_p^2 = 0.53$), while the main effect of Visual Quality and the interaction of Auditory and Visual Quality were not significant ($F(1,17) < 2.0$, $p > .18$ for both effects).

2.2. fMRI results

Two whole-brain group-level analyses were performed on the fMRI data. In both, a general linear model (GLM) was fit to the time series data of each voxel in each run. The first analysis was performed to assess block-level effects related to the different tasks and audiovisual qualities (i.e., effects averaged over all the dialogue lines). In this analysis, the GLM included one regressor modeling all dialogue lines in each task and audiovisual quality combination (e.g., all lines spoken during the S task with good auditory and good visual condition were modeled with one regressor, while another regressor modeled the lines during the S task with poor auditory and good visual condition, etc.). Main effects and interactions of Task, Auditory Quality, and Visual Quality were coded into the first-level GLM for separate 2 × 2 × 2 repeated-measures ANOVAs with factors Task, Auditory Quality, and Visual Quality for all task pairs (i.e., S and V, S and P, P and S; Fig. 3).

The second whole-brain analysis was aimed to assess whether the task-related effects were dynamically modulated across the time course of a dialogue (see also our previous study, Wikman et al., 2021). To this end, another GLM was constructed that modeled each dialogue line in each task and audiovisual quality condition with a separate regressor. The first six lines (out of seven) of each dialogue were included in subsequent analyses, because the last line was always followed by questions related to the task performed by the participant, potentially confounding the fit of the model for that line. Based on this GLM, contrasts were formed for each task pair (i.e., S vs. V, P vs. V, and S vs. P) that modeled linear, quadratic, and linear-quadratic (i.e., a combination of linear and quadratic effects) trends in the data. That is, each line during one task and audiovisual quality combination was contrasted with the corresponding line from another task (e.g., the first line of the S task with the first line of the V task), with the linear, quadratic, and linear-quadratic trends modeling changes in activity levels across the

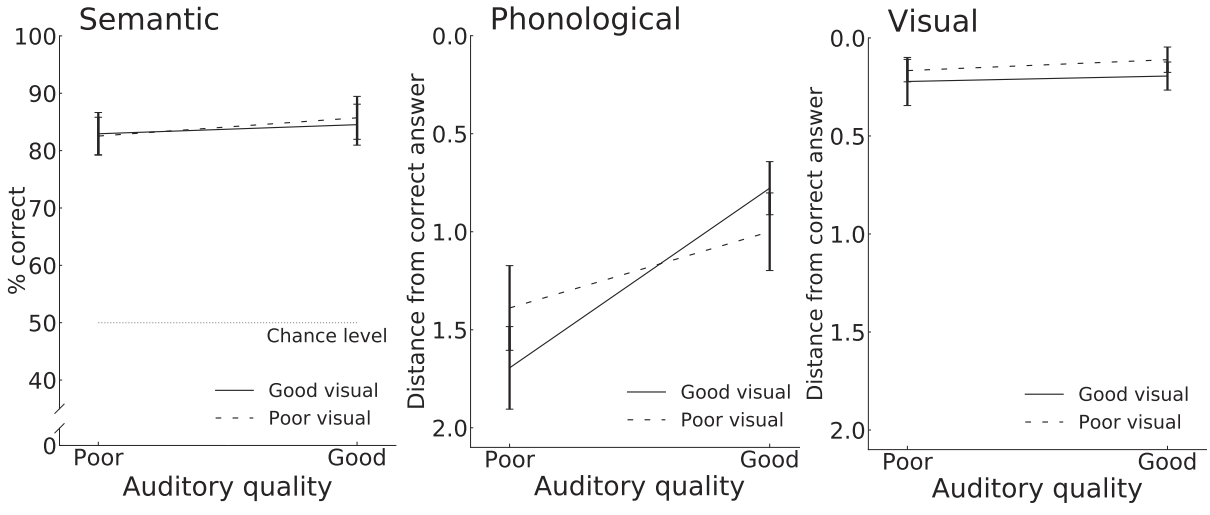


Fig. 2. Behavioral performance accuracy of participants in the different tasks and audiovisual conditions (±SEM). In the Semantic (S) task, performance was above chance level in all conditions, but the audiovisual qualities had no significant effect on performance. In the Phonological (P) task, better Auditory Quality improved performance significantly ($F(1,17) = 18.9$, $p < .001$, $\eta_p^2 = 0.53$), while a significant main effect of Visual Quality or interactions were not found. In the Visual (V) task, audiovisual qualities had no effect on task performance.

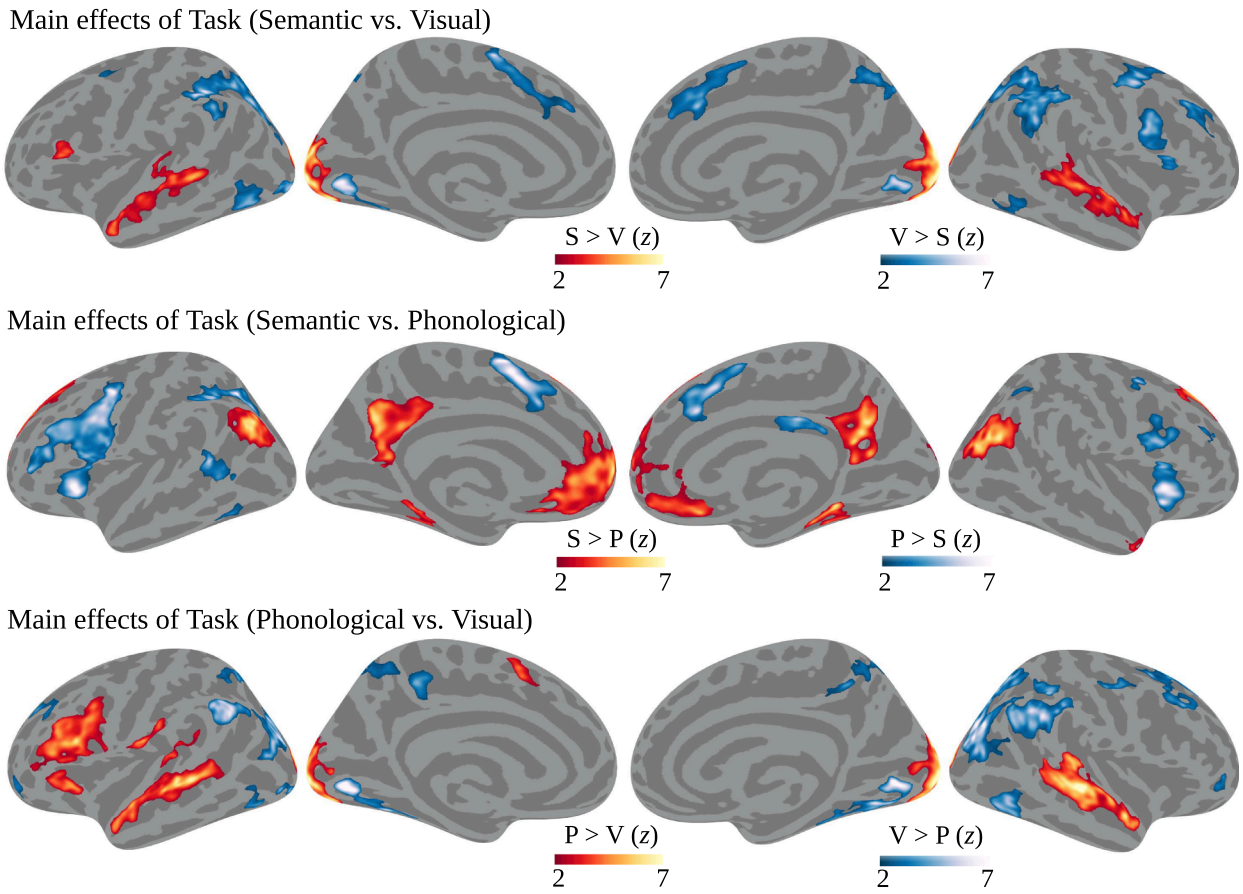


Fig. 3. Main effects of Task were observed in widespread regions of the brain in all three task comparisons (initial cluster forming threshold z = 3.1, permutated cluster significance p < .05, FDR-corrected across all fMRI comparisons). In the top row, red/yellow denotes areas with significantly higher activity during the Semantic (S) task than during the Visual (V) task, and blue/white denotes the opposite effects. In the middle row, red/yellow denotes areas with significantly higher activity during the S task than during the Phonological (P) task, and blue/white denotes the opposite effects. In the bottom row, red/yellow denotes areas with significantly higher activity during the P task than during the V task, and blue/white denotes the opposite effect.

lines of a dialogue. Main effects and interactions of Task, Auditory Quality, and Visual Quality were coded into the first-level GLM.

All fMRI analyses were performed with surface-projected fMRI data, as our focus was on cortical activation modulations. Moreover, as the present study aimed to examine task-related effects, all contrasts not involving the Task factor (i.e., main effects of Auditory and Visual Quality and the interaction of Auditory and Visual Quality) were left unanalyzed to increase statistical power. Initial cluster forming threshold z = 3.1 was used in both analyses, with permutated cluster significance p < .05, FDR-corrected across all fMRI comparisons (see Section 5.8. for details).

2.2.1. *Task-dependent effects*

In the ANOVA contrasting the S and V tasks, the S task was expectedly associated with significantly higher activity in the LIFG and the STG and STS bilaterally, as well as in the occipital regions (Fig. 3, top, red). Contrary to our hypothesis, the S task was not associated with significantly higher activity than the V task in the medial prefrontal regions. Regions showing significantly higher activity during the V task than during the S task were found bilaterally in occipital, superior parietal, inferior temporal, superior frontal, and lingual regions, as well as in the right precentral gyrus, IFG, middle frontal gyrus, and the right precuneus (Fig. 3, top, blue).

In the ANOVA contrasting the S and P tasks, the S task was associated with significantly higher activity in the bilateral medial prefrontal cortex, as well as in the bilateral angular gyrus (AG) and adjoining inferior parietal regions, parahippocampal gyrus, and the posterior cingulate cortex (PCC), and in the right anterior temporal lobe (ATL) and right occipital cortex (Fig. 3, middle, red). Regions with significantly higher activity during the P task than during the S task were observed in a large cluster encompassing parts of the LIFG, left inferior frontal sulcus (LIFS), left PreCG, and left insula. Activity during the P task was also observed in the left posterior STG/STS and left inferior temporal gyrus (ITG), bilateral parietal and superior medial frontal regions, and right precentral, inferior and middle frontal regions, insula, and retrosplenial cortex (Fig. 3, middle, blue). Contrary to our hypothesis, the P task was not associated with significantly higher activity than the S task in AC regions of the anterior STG.

In the ANOVA contrasting the P and V tasks, the P task elicited significantly higher activity in a large cluster in the LIFG and LIFS, and in the bilateral STG and STS, as expected. Activity was also observed in the left insula, left postcentral/supramarginal gyrus, left medial superior frontal region, and in the occipital regions bilaterally (Fig. 3, bottom, red). Clusters with significantly higher activity during the V task than during the P task were found bilaterally in occipital and parietal regions, ITG, superior and anterior prefrontal areas, the lingual gyri, precuneus, and in the right superior PreCG (Fig. 3, bottom, blue).

To illustrate regions that were more strongly activated by both of two different tasks as compared to the remaining task (e.g., regions where both S > V effects and P > V effects were significant), a conjunction image was formed based on the results observed in this analysis (Fig. 4). Areas with higher activity during the S and P tasks than during the V task

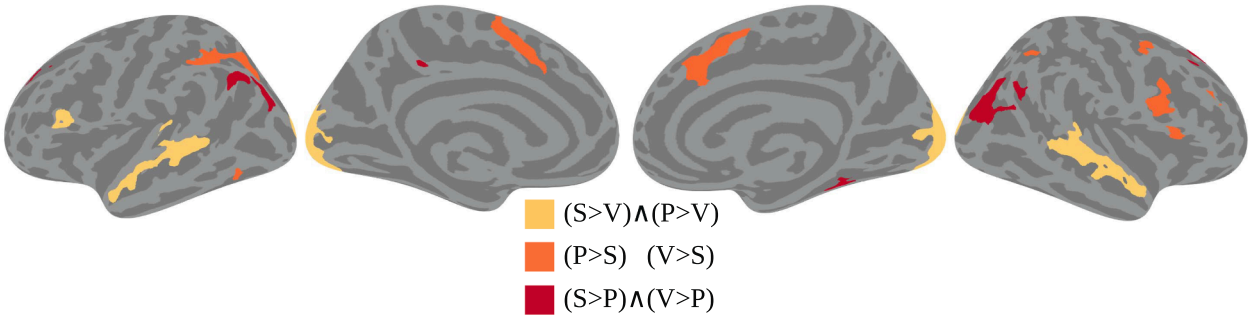


Fig. 4. Conjunctions of regions with significantly higher activity during the Semantic (S) and Phonological (P) tasks than during the Visual (V) task (yellow), regions with significantly higher activity during the P and V task than during the S task (orange), and regions with significantly higher activity during the S and V tasks than during the P task (red). The conjunctions were formed based on the clusters illustrated in Fig. 3.

overlapped in the bilateral STG and STS, LIFG, and occipital regions (Fig. 4, yellow). When contrasted with the S task, the P and V task activated overlapping areas in the right precentral region, MFG and IFG, as well as in superior medial frontal regions, bilateral intraparietal sulcus, and in the left ITG (Fig. 4, orange). When contrasted with the P task, the S and V tasks activated overlapping areas bilaterally in the superior frontal, inferior parietal and occipital regions, as well as in the left PCC and right parahippocampal area (Fig. 4, red).

Clusters with significant interactions between Task and Visual Quality were found in the ANOVA with the S and V tasks and in the ANOVA with the P and V tasks; these clusters were located in the visual areas of the occipital cortex, which are not of essential importance to the research questions of the present study; therefore, they are not discussed further here (see Electronic Supplementary Materials for result images). No significant Task × Auditory Quality or Task × Auditory Quality × Visual Quality interactions were found.

2.2.2. Activations in the OFC

Based on our earlier experiments (Leminen et al., 2020; Wikman et al., 2021), we hypothesized that the OFC would be most strongly engaged by the S task. A region-of-interest (ROI) analysis was performed to test this hypothesis. The ROI was defined based on the cortical parcellation of Yeo and colleagues (Yeo et al., 2011), from where the medial orbitofrontal region was selected from each hemisphere (Fig. 5; cortical network '17Networks_10' in the parcellation). This ROI approximately encompasses the OFC activations observed in the S > V contrast in the studies by Wikman and colleagues (2021) and Leminen and colleagues (2020). A 3 × 2 × 2 repeated-measures ANOVA with factors Task, Visual Quality, and Auditory Quality was performed on the percent signal change values obtained from this ROI (averaged over the two hemispheres). The main effect of Task reached statistical significance ($F(2,16) = 3.30$, $p = .049$, $\eta_p^2 = 0.16$), while the main effects of Visual and Auditory Quality and interaction effects did not (although the interaction between Task and Visual Quality approached significance, $F(1,17) = 2.95$, $p = .066$; for all other effects $F(1,17) < 1.17$, $p > .30$). Therefore, the values were averaged across all four Auditory and Visual Quality combinations for plotting (Fig. 5). The plot indicates that the S task was associated with the highest level of activity, whereas the P task was associated with the lowest level of activity.

2.2.3. Modulations in activation levels across the lines of a dialogue

Clusters with significant linear effects were found in all task comparisons. Clusters with significant quadratic effects were found in the contrast including the S and V tasks and in the contrast including the P and S tasks. Clusters with significant linear-quadratic effects were found in the contrast including the S and V task and in the contrast including the P and V tasks. No interaction effects were found in this analysis. We describe here only clusters found in regions that are essential to the present research questions (all other results can be found in the Electronic Supplementary Materials). Note that while these results were obtained in comparisons between two tasks, we also extracted and analyzed the activity values of the third task to better understand what the observed effects may relate to.

In the contrast including the P and V tasks, a cluster with a significant linear effect was found in the LIFS/LIFG (Fig. 6, left). As indicated by the plot, during the P task, activity in this cluster was highest for the first

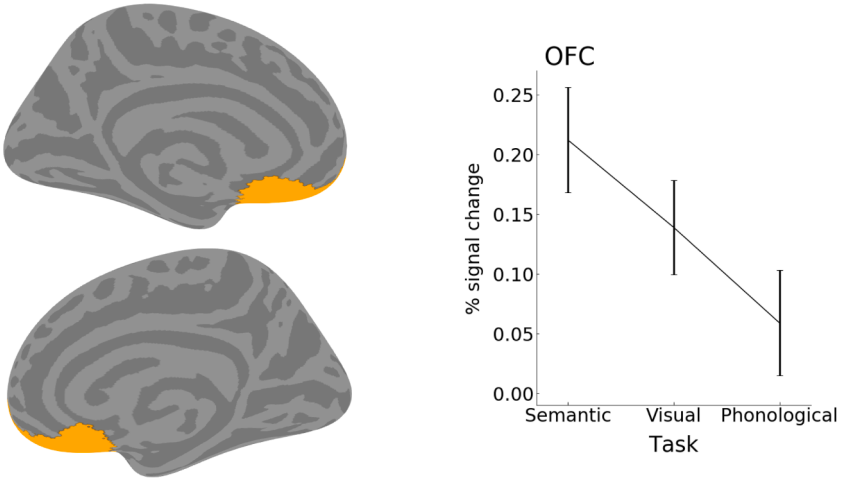


Fig. 5. A region-of-interest (ROI) analysis was performed in the medial orbitofrontal regions. The y-axis shows percent signal change (±SEM) relative to a resting baseline.

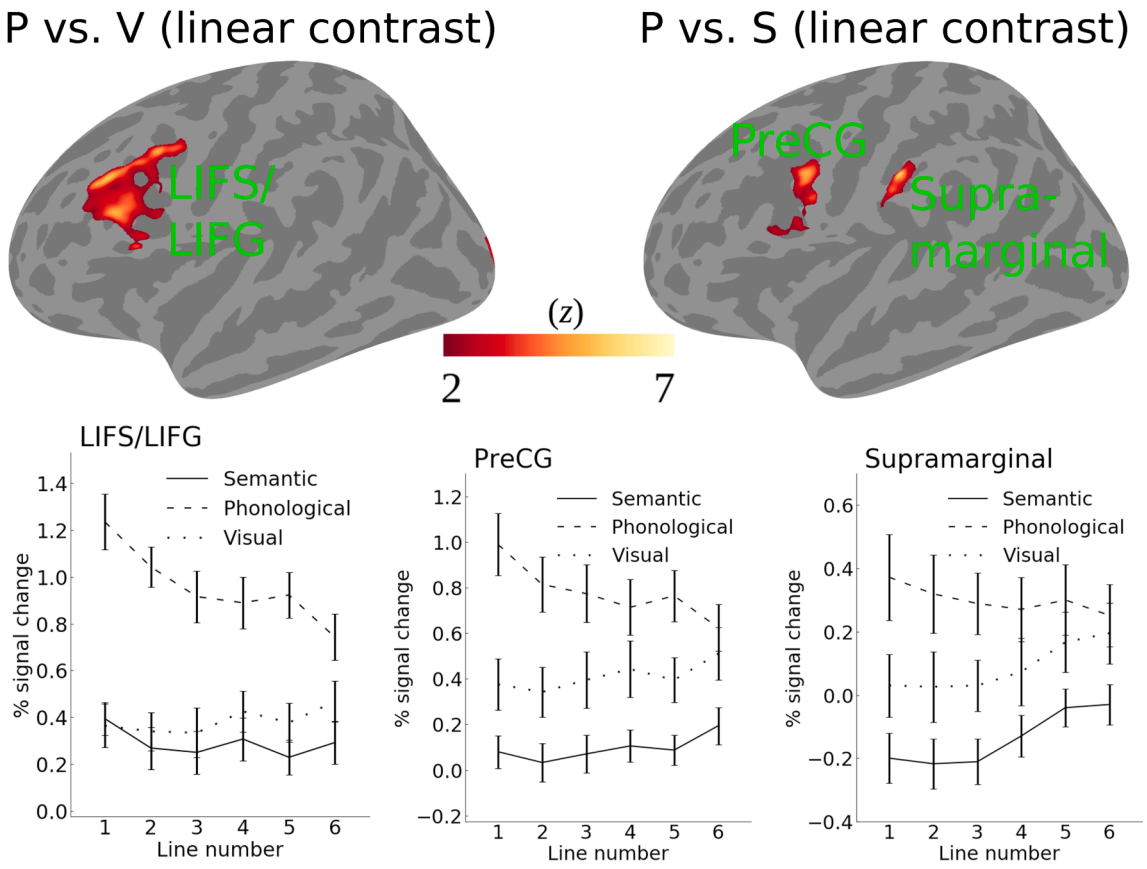


Fig. 6. Effects observed in the linear contrast including the P and V tasks (left) and in the linear contrast including the P and S tasks (right; initial cluster forming threshold z = 3.1, permutated cluster significance p < .05, FDR-corrected across all fMRI comparisons). The y-axis shows percent signal change (±SEM) relative to a resting baseline.

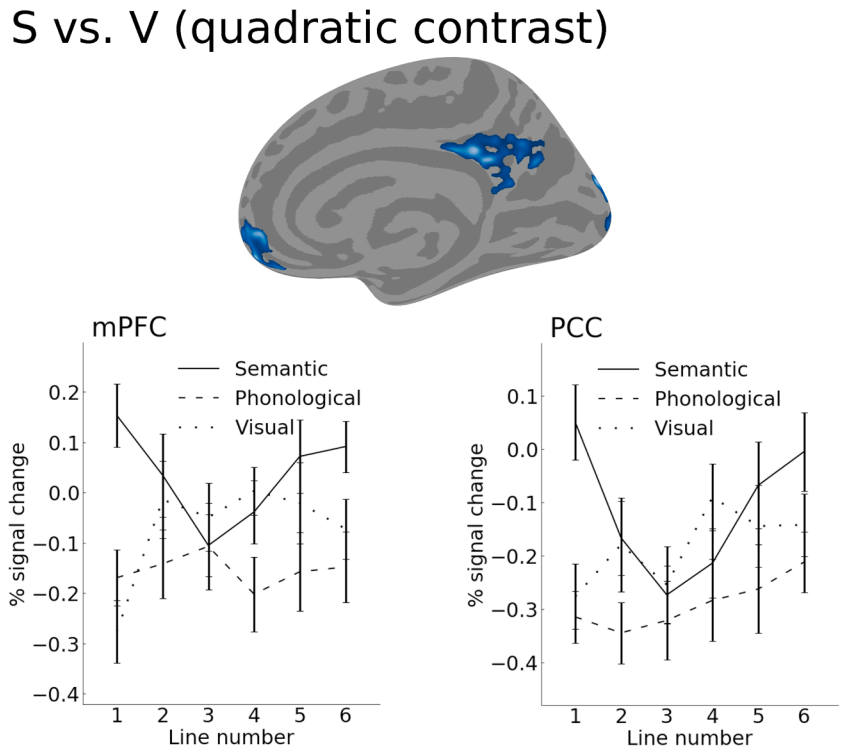


Fig. 7. Effects observed in the quadratic contrast including the S and V tasks (initial cluster forming threshold z = 3.1, permutated cluster significance p < .05, FDR-corrected across all fMRI comparisons). The y-axis shows percent signal change (±SEM) relative to a resting baseline.

line of the dialogue, with a subsequent decrease during the following lines. To analyze whether the temporal activity profiles in this cluster during the P and S tasks also differed from each other, a 2 × 6 repeated measures ANOVA with factors Task and Line was performed. This ANOVA showed significant main effects of Task ($F(1,17) = 107$, $p < .001$, $\eta_p^2 = 0.86$) and Line ($F(5,85) = 10.9$, $p < .001$, $\eta_p^2 = 0.39$), as well as a significant Task × Line interaction ($F(5,85) = 5.0$, $p < .001$, $\eta_p^2 = 0.29$), indicating that the temporal modulations in activity levels during these tasks were different.

In the contrast including the P and S tasks, clusters with significant linear effects were found in the left PreCG/ventral IFS, and in the left supramarginal region (Fig. 6, right). To analyze whether the temporal activity profiles of the P and V tasks also differed from each other in these clusters, 2 × 6 repeated measures ANOVAs with factors Task and Line were performed. In both clusters, the main effect of Task was significant (PreCG: $F(1,17) = 19.4$, $p < .001$, $\eta_p^2 = 0.53$; Supramarginal: $F(1,17) = 18.8$, $p < .001$, $\eta_p^2 = 0.53$), with activity being higher during the P task. The main effects of Line were not significant ($F(5,85) < 2.2$, $p > .067$ in both cases). In both clusters, however, there was a significant Task × Line interaction (PreCG: $F(5,85) = 6.1$, $p < .001$, $\eta_p^2 = 0.26$; Supramarginal: $F(5,85) = 3.2$, $p = .01$, $\eta_p^2 = 0.16$), indicating differences between the temporal activity modulations in the two tasks.

In the contrast including the S and V tasks, significant quadratic effects were observed in the right medial prefrontal cortex and the PCC (Fig. 7). In these clusters, activity was initially higher during the S task than during the V task, but the difference decreased during the first lines of the dialogue, and then increased again towards the end of the dialogue.

3. Discussion

The present study was designed to examine the neural basis of selective attention to continuous audiovisual speech. The experimental design enabled us to study neural activations related to selectively attending to speech when focusing either on the meaning (S task) or the phonological content (P task) of the attended speech stream, or when ignoring speech altogether and performing a visual control task (V task). To examine interactions between task-dependent effects and speech intelligibility, the tasks were factorially combined with good and poor auditory and visual speech qualities.

3.1. Behavioral performance

The present behavioral results indicate that the participants were able to understand the dialogues and respond successfully to questions related to the semantic content of the dialogues (Fig. 2, left). Contrary to the results of previous experiments (Leminen et al., 2020; McGettigan et al., 2012; Wikman et al., 2021), auditory and visual speech qualities had no significant effects on performance in the S task. This possibly relates to the fact that unlike the participants of our previous studies (Leminen et al., 2020; Wikman et al., 2021), the participants of the present study practiced the tasks for 1–2 h before the actual experiment (the practice session was arranged to maximize behavioral performance in a shadowing task that was also performed but is not reported in the present study; see Section 5.3). Consequently, the present participants were relatively experienced at listening also to the degraded speech qualities. This may have lead to a ceiling effect in the behavioral scores with regard to the effects of the audiovisual qualities. Nonetheless, in the P task, better auditory quality did significantly improve performance (Fig. 2, middle). Contrary to our hypothesis, better visual quality did not affect performance in the P task either. In earlier research, better visual quality has been associated with better performance in speech tasks especially in situations with poor auditory quality (McGettigan et al., 2012; Peelle and Sommers, 2015; Sumby and Pollack, 1954). The relatively long training session might partially explain why no such effects are seen in the present study. Another possible explanation relates to the fact that the phoneme /r/ is not very visible on the face of a speaker (see, e.g., Files et al., 2015) and, therefore, visual quality most likely helped the participants mainly indirectly through the increased overall intelligibility of speech, enabling the use of secondary detection strategies, such as covert repetition of speech.

3.2. Task and attention-related effects in the STG/STS

Performing different tasks was associated with activity modulations in widespread regions of the brain. As hypothesized, both the S and P tasks were associated with enhanced activity in the STG and STS, as well as the LIFG and occipital areas (Fig. 3, top and bottom; Fig. 4, yellow). Previous research has implicated these areas in many aspects of speech and voice processing, such as mapping between auditory and conceptual information (Hickok and Poeppel, 2007; Rauschecker and Scott, 2009) and audiovisual integration (Jääskeläinen, 2010; Venezia et al., 2017), as well as in selective attention to speech and voice (Alho et al., 2014, 2006; Hill and Miller, 2010; Wikman et al., 2021). The present study corroborates and extends earlier findings by showing that these regions are involved in selective attention to continuous speech regardless of whether it is the meaning or the phonological content of speech that is focused on.

Previous research has implicated the anterior and posterior portions of the STG in phonological processing (Leonard and Chang, 2014; Mesgarani et al., 2014; Turkeltaub and Branch, 2010), and task-dependent activity modulations related to phoneme discrimination have also been shown in these regions (Harinen and Rinne, 2013; Wikman and Rinne, 2019). Therefore, we expected the P task to engage these regions even more than the S task. In the anterior STG, we found no support for this hypothesis, suggesting that task-specific activations in the anterior STG may not be as eminent when attending to naturalistic stimuli. However, this cannot, of course, be conclusively determined based on the present results. In the left posterior STG/STS, the P task was associated with higher activity than the S task. This result is possibly related to our hypothesis regarding the role of the dorsal stream of speech processing in the P task, which is discussed below (Section 3.4.).

In our previous study with a setting similar to the present one (Wikman et al., 2021), activity modulations across the lines of a dialogue were observed in the STG and STS so that activity first increased for the first half of a dialogue, and then decreased toward the end of the dialogue. In that study, we hypothesized that this modulation might reflect the process of gradual automatization of task-performance: at first, all neuronal networks are recruited that might be useful in performing a given task (possibly with the help of top-down control from frontal regions; see Section 3.3. below), but gradually, the most efficient mechanism is arrived at, and less useful neuronal activations are pruned out (see also Kilgard, 2012). In the present study, no significant activity modulations were observed in the STG/STS. While this may relate to statistical power, as we had fewer data per task in the present study than in the previous one (Wikman et al., 2021), it should also be noted that the result is consistent with the above-mentioned 'efficiency hypothesis', too. As stated earlier, the present participants practiced the tasks more extensively than the participants of our previous study, which may have lead to less prominent activity levels and their modulations. Moreover, in the previous study, there was a non-significant trend toward the STG/STS activation decreasing from the first to the last run, consistent with the hypothesis that the use of neural resources becomes more efficient with more practice (note also that this effect was most likely not due to fatigue, as indicated by behavioral performance; see Wikman et al., 2021, Supplementary Materials).

3.3. Effects in the frontoparietal regions

Studies on selective attention to simple sounds, such as tones and single syllables, have found attention-related activations in the frontoparietal network. These activations have been proposed to be associated

with maintaining selective attention (Alho et al., 1999; Degerman et al., 2006; Salmi et al., 2007; Zatorre et al., 1999). It would seem reasonable to expect these regions to show activity also when maintaining selective attention to continuous speech in the presence of other, task-irrelevant speech streams, but this seems not to be the case (Alho et al., 2006, 2003; Leminen et al., 2020; Wikman et al., 2021). It has been suggested that this pattern of results may relate to the fact that selectively attending to speech is something most humans rehearse throughout their lives (Alho et al., 2006), which is likely to diminish the need for cognitive control in performing the task (Chein and Schneider, 2012). The results of the present study corroborate this view in that no DLPFC or parietal activations were observed during the S task when compared with the other tasks. On the other hand, both the P and V tasks did engage these frontoparietal regions when compared with the S task, especially in the right DLPFC and left parietal regions (Fig. 3, top and middle; Fig. 4, orange). It should be noted that similar activations have also been observed in response to increasing working memory (WM) load (Emch et al., 2019; Manoach et al., 1997). However, it seems unlikely that WM load would be a decisive factor in accounting for the DLPFC and parietal activations observed during the P and V tasks, because updating and maintaining a single number in WM (i.e., the number of task-relevant occurrences) is a relatively low-load WM task (cf. Manoach et al., 1997). Moreover, the S task also included a WM component (i.e., keeping in mind what was said in the dialogue), although the load imposed by the S task is harder to quantify. Thus, our results are in line with the view that the frontoparietal processes may not be as crucial when selectively attending to the meaning of continuous speech as in maintaining selective attention in less naturalistic situations.

Frontal and parietal regions may still contribute at the initial stages of tuning in to the speaker or speakers, or to the task at hand, or both (Hill and Miller, 2010; Näätänen, 1990; Wikman et al., 2021). In our previous study (Wikman et al., 2021), a frontal network including the bilateral inferior frontal gyri and premotor regions was observed, where activity was high at the beginning of a dialogue and decreased toward the end. These regions were termed the 'primary control network', in line with the hypothesis that they orchestrate and facilitate processing in sensory regions. It was suggested that these processes are required especially at the initial stages of the task, and that activity then decreases as processing in the sensory regions becomes more independent. Accordingly, as discussed above (Section 3.2.), activity in the sensory speech regions was observed to increase throughout the first lines of the dialogue, possibly reflecting top-down influences from the frontal regions, and then decreased subsequently (Wikman et al., 2021). The present study did not replicate these results with the linear, quadratic, or linear-quadratic contrasts including the S and V tasks. However, in the linear contrast including the P and V tasks, a cluster in the LIFS/LIFG was observed, where activity during the P task was high at the beginning of a dialogue, and then decreased subsequently (Fig. 6, left). During the S and V tasks, activity levels in this cluster were lower overall and did not undergo such modulations. The cluster observed here partially overlaps with the 'primary control network' of Wikman and colleagues (2021), and the result might reflect the top-down control processes postulated there. That is, this region could be involved in control processes that help guide processing in the sensory regions in a task-dependent manner, with activity decreasing as processing in the sensory regions becomes more automatized. That the present participants practiced the tasks more extensively than the participants of Wikman and colleagues might partially account for why the effect was not seen during the S task, but was seen during the more novel P task. Additionally, more extensive practice might also explain why the effect in the present study was observed with a linear contrast, whereas previously (Wikman et al., 2021), it was observed with a linear-quadratic contrast (i.e., in the previous study, activity was observed to persist on a high level for a longer time at the beginning of a dialogue).

While the observed LIFS/LIFG modulations may reflect frontal top-down control processes, as discussed above, it should also be noted that the observed LIFS/LIFG cluster partially overlaps with the dorsal stream regions of Hickok and Poeppel (2007). Moreover, this region has multiple proposed functions, including, for example, phonological working memory (Bohland and Guenther, 2006; Papoutsi et al., 2009). Therefore, a complementary interpretation to these results can be given in the context of the dual-stream theory of speech processing (Hickok and Poeppel, 2007).

3.4. *Effects in dorsal stream regions*

The P task was designed to require sub-lexical processing while selectively attending to continuous speech. We expected this task to engage areas in the dorsal stream of speech processing, which is suggested to connect the left AC regions with inferior frontal and premotor regions via the posterior STG and IPL. The dorsal stream is proposed to have a role in mapping between auditory and motor speech representations (Fridriksson et al., 2016; Hickok and Poeppel, 2007; Mugler et al., 2018; Rauschecker and Scott, 2009), but its role in normal speech perception is debated (Schomers and Pulvermüller, 2016). It has also been proposed that the dorsal stream regions participate in speech processing in a compensatory manner when speech is degraded or overlaid with noise (Du et al., 2014; Osnes et al., 2011; Wild et al., 2012). The present study found no support for this hypothesis, however, as interactions between task and the audiovisual qualities were not found.

As expected, the P task was associated with higher activity than the other tasks in the left posterior STG/STS, as well as in the posterior LIFG and LIFS (Fig. 3, middle, bottom; Fig. 6, left). This result is consistent with the hypothesis that these regions are especially recruited during sub-lexical processing. In left ventral premotor regions, the P task was associated with higher activity than the S task in areas 6r and 6v, whereas activations in this region were less extensive when contrasting the P task with the V task (for the anatomical labels, see Glasser et al., 2016). This result was somewhat surprising, as we expected that the left premotor regions would be especially engaged by sub-lexical processing, and would therefore show higher activity during the P task when contrasted with either of the other tasks. We hypothesize, a posteriori, that this result relates to common cognitive processes involved in performing the P and V tasks. In particular, both tasks involved detecting occurrences of task-relevant events in a continuous stimulus stream, as well as mentally counting these occurrences, and maintaining the number in WM. Relatedly, the P and V tasks engaged partially overlapping areas in the right precentral and dorsal medial frontal regions, as well as in bilateral parietal regions (Fig. 4, orange), and previous fMRI research has implicated a similar network of areas to be active during target detection in both auditory and visual modalities (Stevens et al., 2000; Yoshiura et al., 1999). Moreover, mental calculation has also been associated with activation in the premotor regions (Kansaku et al., 2007; Kansaku et al. 2006; Tschentscher et al., 2012). These findings may partially account for the overlapping activations during the P and V tasks, and the fact that in the block-level analyses, the left premotor cortex was not more activated by the P than by the V task.

When contrasted with the V task, the P task showed higher activity at the border of the superior temporal and inferior parietal cortices, as well as in the ventral postcentral gyrus (Fig. 3, bottom). The former cluster likely partially coincides with area Spt of the dual-stream model of Hickok and Poeppel (2007, 2004), which is proposed to function as an audiomotor interface. The latter cluster is at least partially located in the secondary somatosensory cortex (SII; area OP4 and PF opercular), possibly coinciding with somatotopic representations of the orofacial area (Matelli and Luppino, 2001; Sanchez Panchuelo et al., 2018; note, however, that since we had no functional localizers in the present study, we cannot be sure about the exact somatotopic location of this activation). This region has been implicated in aspects of articulation and phonological working memory (among other things; Binkofski et al.,

2016; Gierhan, 2013). Relatedly, the linear contrast including the P and S tasks also revealed a cluster at least partially located in the SII (area PFop; Fig. 6, right), as well as another cluster in the left precentral areas (extending ventrally from area 6 to area 44). As mentioned earlier, these precentral regions have been associated with the representation of articulatory gestures and phonemes (Clos et al., 2013; Mugler et al., 2018). Activity during the P task in these clusters was highest at the beginning of a dialogue, and then decreased approximately linearly throughout the dialogue, similarly as in the LIFG/LIFS cluster discussed above (Section 3.3.). In these clusters, activity was highest overall during the P task, and the temporal profile of activity during the P task was different from those of the S and V tasks. In the precentral and SII clusters, the V task was also associated with higher overall activity than the S task (Fig. 6). While this may relate to common cognitive components in the P and V tasks, such as target-detection or mental calculation, it is unlikely that these functions could completely account for the finding that activity during the P task was highest. This is because the temporal activity profile of the P task was also different from both other tasks in these clusters, indicating processes specific to the P task.

It has been shown that motor and somatosensory regions can play a role in speech perception (D'Ausilio et al., 2009; Ito et al., 2009; Möttönen et al., 2005; Möttönen and Watkins, 2009), and it seems plausible that the participants of the present study, or some of them, made use of motor and somatosensory information especially during the P task. Moreover, because of the decreasing trend in activity levels during the P task in these regions, it seems possible that this information was utilized especially at the beginning of task-performance, with other strategies becoming more important subsequently. An alternative explanation, as per the 'efficiency hypothesis' discussed above, is that neural processing became more efficient during the time course of the dialogues. In any case, it is possible that motor and somatosensory information were used in conjunction, possibly, for example, in mentally simulating speech. Studies in the macaque have also identified anatomical and functional connections between the macaque areas F4 and PF/ventral intraparietal area (VIP), which have similar properties with the human areas 6 and PF, respectively (Duhamel et al., 1998; Gentilucci et al., 1988; Goulas et al., 2017; Luppino et al., 1999). This provides some support for the idea that the activations and their temporal profiles in these regions might indicate co-operation between speech motor and somatosensory regions during the P task.

Taken together, our results indicate that the motor regions of the dorsal stream of speech processing (Fridriksson et al., 2016; Hickok and Poeppel, 2007; Rauschecker and Scott, 2009), as well as regions in the SII possibly corresponding to the somatotopic representations of the face and mouth, were especially engaged by the P task. The present results do not, of course, rule out the possibility that motor and somatosensory regions may play a role in normal speech perception. Multivoxel pattern analysis (MVPA) methods might also reveal task-related differences in the activity patterns of these regions, which we cannot see with univariate methods. For example, in our previous study (Wikman et al., 2021), MVPA did find significant differences in the premotor regions between the S and V tasks, although it is difficult to pin down exactly what this reveals about neural processing in this region during the different tasks. Note also that, in the present study, we were unable to use multivariate analysis methods, because due to time limitations of fMRI measurements, we only had two presentations of each task in each audiovisual quality condition, which is not enough for multivariate analyses. Nonetheless, the present results do suggest that the involvement of the motor and somatosensory regions is especially prominent when the task focuses on sub-lexical aspects of speech. Moreover, our results demonstrate that when considering the role of these regions in speech processing, inspecting the time course of activations may provide additional information not attainable with only block or event-related analyses.

3.5. Effects in regions associated with social and semantic cognition

In our two previous experiments with an experimental setting similar to the present one, selective attention to continuous audiovisual speech was associated with activation in the angular gyrus (AG), anterior temporal lobe (ATL), PCC, and OFC (Leminen et al., 2020; Wikman et al., 2021). The AG and ATL have been suggested to act as 'semantic hubs' that receive and integrate inputs from various parts of the brain (Binder et al., 2009; Jefferies, 2013), while the PCC is implicated in a wide variety of functions including memory and aspects of attentional regulation (Leech and Sharp, 2014). That these regions showed activation while the participants viewed meaningful continuous audiovisual speech was, therefore, not surprising. Activation in the OFC was somewhat unexpected, however. Evidence suggests that this region has a role in semantic and social cognition (i.e., the processing of affective information, moral judgments, and theory of mind; Adolphs, 2009; Alcalá-López et al., 2019, 2019; Mitchell et al., 2006), but it has not been consistently implicated in studies on speech processing. Moreover, the stimuli of our studies (i.e., the audiovisual dialogues) were deliberately written to be as emotionally neutral as possible, and the tasks only related to the semantic content of speech, not any specifically social aspect of the situation (e.g., speakers' emotions or relationship). It seems unlikely that the OFC activity could be completely explained by semantic processing either, because in our other previous studies, no activity in these regions was observed while selectively attending to continuous speech and focusing on its meaning (Alho et al., 2006; Alho et al., 2003). In these studies, however, the speech stimuli were spoken by one speaker only, and the speech was presented only auditorily. In our recent study (Wikman et al., 2021), we also tested the hypothesis that the OFC activations would be affected by the semantic and social coherence of an audiovisually presented dialogue (i.e., whether or not the lines of the attended dialogue formed a coherent conversation), but no OFC modulations related to this manipulation were observed. We therefore put forth the hypothesis that it might be the mere audiovisual presentation of the dialogues that enhances social engagement and recruits the OFC.

One of the aims of the present study was to assess whether activity in the OFC is modulated by the task type even when both tasks involve viewing and attending to audiovisual speech. Contrary to our expectations, the whole-brain analysis contrasting the S and V tasks did not replicate our previous results (Leminen et al. 2020; Wikman et al., 2021) regarding activation in the OFC (Fig. 3, top). However, in the contrast between the S and P tasks, the OFC (along with other medial prefrontal areas) did show significantly higher activity during the S task than during the P task (Fig. 3, middle). The reason for the contrast between the S and V tasks not replicating our previous results probably relates to statistical power, as we had fewer fMRI data per task in the present study than in our previous ones. An ROI analysis in the OFC nonetheless found a significant main effect of Task, with the S task being associated with the highest level of activity, and the P task with the lowest level (Fig. 5). This finding shows that the task performed by a listener affects activations in this region even when both tasks involve attentive processing of audiovisual speech. Higher OFC activation during the V task than during the P task might be due to the participants having covertly processed the content of the dialogues to a larger degree during the V task, as it was arguably easier than the P task (Fig. 2). This might also explain the pattern of present results observed in the AG, right ATL, and PCC, where the S task was associated with higher activity than the P task (Fig. 3, middle), but not with higher activity than the V task (Fig. 3, top), contrary to our previous results (Leminen et al., 2020; Wikman et al., 2021). It should be noted, that the mPFC regions, as well as the PCC and AG, are also implicated as important nodes in the default mode network (DMN; Buckner and DiNicola, 2019). Therefore, it could be argued that lower activation during the more demanding P task reflects suppression of DMN processing. However, it seems unlikely that suppression related to task difficulty would entirely explain this pattern of results, as the V task

was the easiest task, and still not associated with higher activity than the S task in these regions. Future experiments could nonetheless test this hypothesis by systematically varying the difficulty of the control task performed while viewing socially engaging stimuli.

In the quadratic contrast with the S and V tasks, significant effects were observed in the right mPFC and the right PCC (Fig. 7). This indicates that the time course of activity during the S task was different than that during the V task, even though block-level effects were not observed in these regions. The mPFC cluster observed in this quadratic contrast is mostly distinct from our OFC ROI. However, the medial prefrontal regions also outside the OFC, as well as the PCC, have been implicated in aspects of social cognition (Bzdok et al., 2013; Leech and Sharp, 2014); these functions are, therefore, one possible factor accounting for the activity modulations observed here. The temporal profiles of activity modulations across the different tasks in the mPFC and PCC clusters are somewhat complex (Fig. 7). During the S task in both clusters, activity is high in the beginning of a dialogue, decreases subsequently, and then rises again. In the mPFC during the V task, activity seems to be low in the beginning, and then rise to approximately constant levels for the rest of the dialogue, whereas activity during the P task seems to remain approximately constant throughout the dialogue. In the PCC, the temporal profile of activity during the V and P tasks is somewhat similar to the mPFC, although not quite as clear-cut. The initially low but subsequently rising activity profile of the V task could reflect initial suppression of processing in these regions, with the subsequent increase in activity possibly correlating with task-performance becoming more automatized (e.g., more resources can be allocated to processing the dialogue, although it was to be ignored and is task-irrelevant during the V task). During the S task, on the other hand, high activity at the beginning with a subsequent decrease might be related to the above-mentioned 'efficiency hypothesis', although it is less clear why activity then returns to higher levels.

3.6. Limitations of the present study

There are certain aspects of the tasks that were not controlled for in the present experimental setting. For example, the P task was arguably more difficult than the other tasks, which may complicate the interpretation of some of our results. Note, however, that degrading the auditory quality of speech was also associated with increasing task difficulty in the P task (Fig. 2), but no interactions of Task and Auditory Quality were observed in the fMRI results. This speaks against any strong influences of task difficulty, as effects related to audiovisual quality would have been expected to be seen during the S and P tasks, but not during the V task.

Attention skills vary from person to person, possibly resulting in increased between-participant variation in our results. Due to the relatively low number of participants, however, analysis of individual differences would not have been feasible in the present study. Thus, individual differences in attention skills were not assessed with any standard psychological test, nor were they taken into account in the analyses. Moreover, the present design did not assess effects related to talker adaptation or normalization (i.e., how the perceptual processing of speech is adapted to specific speakers), which are known to influence speech processing also in situations with multiple speakers and degraded speech (see, e.g., Bent et al., 2009; Bradlow et al., 1999; Stilp and Theodore, 2020). Note, however, that while the participants of the present study likely adapted to the speakers, these effects should not play a significant role in our results, because the order of the tasks was randomized across participants, and performance in all of the tasks should therefore benefit from talker adaptation and normalization effects to a similar extent.

Finally, the present study has discussed the benefits of employing as ecologically valid experimental settings as possible. Accordingly, one of the goals of the present study was to assess speech processing in a relatively naturalistic setting (i.e., S task), and compare it to speech processing during an artificial experimental task performed on the same stimuli (i.e., P task). Yet, it must be acknowledged that the S task, even in the good auditory and good visual condition, is not natural when compared with everyday speech processing situations. While we still believe the ecological validity of the S task to be relatively high when compared to many brain imaging studies on speech processing, it remains for future studies to determine the extent to which even this sort of experimental setting suffers from effects introduced by the artificial nature of the situation.

4. Conclusions

The present study examined the neural basis of selective attention to continuous audiovisual speech in the presence of task-irrelevant speech. The participants either focused on the meaning or on the phonological content of the attended speech stream, or ignored speech altogether. The observed task-dependent modulations indicate that selective attention to speech mostly relies on areas involved in the processing of speech in quiet as well, such as the STG, STS, and LIFG. Frontoparietal systems involved in attentional control may not be critical in sustaining selective attention in more ecologically valid speech listening conditions, although they may be recruited to a larger extent in more novel tasks, and particularly at the beginning of task performance. Regions in the dorsal stream of speech processing, as well as in the secondary somatosensory cortex, were found to be especially engaged by the P task, with distinct temporal profiles during the different tasks also indicating task-dependent differences in how the processes unfold during the time course of a dialogue. While these results do not demonstrate that the dorsal stream regions have no role in normal speech perception, they are consistent with the idea that these regions are especially activated when processing the sub-lexical aspects of speech (Hickok and Poeppel, 2007). The present study further shows that activity in the OFC is modulated by the task performed by a listener during attentive processing of audiovisual speech, corroborating previous research suggesting that this region has a role in social cognition. In conclusion, the present results demonstrate that the neural mechanisms of speech processing and selective attention to speech are strongly modulated by the task performed by a listener. This encourages the use of as ecologically valid experimental settings as possible, if the goal is to understand speech processing as it occurs in everyday contexts.

5. Materials and methods

5.1. Participants

Nineteen participants were recruited from the University of Helsinki mailing lists to take part in the study. One participant was excluded from all analyses due to a technical error in data collection, resulting in a total n of 18 participants (9 females, mean age 25.6 years, range 19–39 years, all university students). All participants were healthy native Finnish speakers who had self-reported normal hearing, normal or corrected-to-normal vision (in which case they wore contact lenses during the fMRI session), and no self-reported history of psychiatric or neurological disorders. All participants were right-handed as verified by the Edinburgh Handedness Inventory (Oldfield, 1971). The participants gave written consent and were monetarily compensated for their time. The experimental protocol was approved by the University of Helsinki Ethical Review Board in the Humanities and Social and Behavioral Sciences, and the study was conducted in accordance with the Declaration of Helsinki.

5.2. Stimuli

Audiovisually presented dialogues between a female and a male speaker were used as stimuli. The dialogues were spoken in Finnish by native speakers, and they were about emotionally neutral everyday

topics (e.g., hobbies; see Supplementary Table 2 for an example of an entire dialogue). The videos were created for a previous study, which also describes the process of their creation in more detail (Leminen et al., 2020). Each dialogue consisted of seven lines spoken alternately by the two speakers. The gender of the first speaker varied from one video to another. The dialogue lines had an average duration of 5.4 s (range 4.9–6.1 s) and were always followed by a pause (mean duration 3.4 s, range 2.9–3.9 s). This resulted in a total duration of approximately one minute per dialogue (range 55–65 s).

For the dialogues to require selective attention to the speech of the two speakers, passages from an audiobook (a Finnish translation of *The Autumn of the Middle Ages*, by Johan Huizinga, originally published in 1919) were added to the background of the videos as an auditory distractor. The audiobook was spoken in Finnish by a female native Finnish-speaking actor, and is distributed freely by the Finnish National Broadcasting Company (Yleisradio, YLE; https://areena.yle.fi/1–3529001). The pitch of the voice of the audiobook's speaker was lowered to an average of 0.16 kHz for it to be clearly distinguishable from both speakers' voices (with F0 frequencies around 0.12 kHz and 0.2 kHz, for the male and female speaker, respectively), and it was low-pass filtered with a cut-off of 5 kHz (for details, see Leminen et al., 2020). The volume of the audiobook was attenuated by 3 dB in comparison to the attended dialogues.

The intelligibility of the attended speech streams was manipulated by noise-vocoding (Shannon et al., 1995). In noise-vocoding, the amplitude envelopes of a speech stream are used to modulate white noise in logarithmically divided frequency bands. The intelligibility of noise-vocoded speech depends on the number of frequency bands used (Davis and Johnsrude, 2003). In the present study, to retain information on the gender of the speakers, the fundamental frequencies (F0; < 0.3 kHz) of the speech streams were left intact, while frequencies from 0.3 to 5 kHz were noise-vocoded on two levels (4 bands vs. 16 bands) using Praat (version 6.0.27; Boersma, 2001; bandwidth frequency boundaries the 4 frequency bands in kHz: 0.300, 0.684, 1.385, 2.665, 5.000; bandwidth frequency boundaries for the 16 bands in kHz: 0.300, 0.376, 0.463, 0.565, 0.684, 0.822, 0.982, 1.168, 1.385, 1.637, 1.929, 2.269, 2.665, 3.124, 3.658, 4.279, 5.000; for further details, see Leminen et al., 2020). This manipulation resulted in two auditory qualities, one with relatively poor intelligibility and the other with good intelligibility. Intelligibility was measured in a separate behavioral pilot experiment by having participants ($n = 5$, not included in the actual experiment) listen to the audio track of the attended speech stream line by line and transcribe what they heard. Performance was assessed by calculating the number of words correctly transcribed per auditory quality condition. In the 4 bands condition, on average 76.4% of words were correctly transcribed (SD = 10.3%), whereas in the 16 bands condition, on average 98.5% of words were correctly transcribed (SD = 18.6%; for more details, see Leminen et al., 2020). The amount of visual information available in aiding speech processing was also varied on two levels. This was accomplished by masking the faces of the speakers with visual noise (for details, see Leminen et al., 2020). In the poor visual condition, the faces were almost completely masked by noise, whereas in the good condition there was very little noise (Fig. 1).

A light gray box containing a fixation cross was added below the face of each speaker in all videos (for purposes of the V task; see Fig. 1). At the beginning of each dialogue video, only one cross was present, and it was placed on the side of the speaker who would utter the first line. The cross below the face of the other speaker faded in 1500 ms after the start of the first line. 500 ms after the end of each line, the fixation cross below whoever had been speaking faded out. The cross that faded out then faded back in 1500 ms after the next line started. Fading in instead of the sudden appearance of the cross was used to avoid bottom-up triggered attention (see, e.g., Corbetta and Shulman, 2002) to the appearing cross. This pattern repeated until the end of the last line, so that most of the time there were two fixation crosses present on the screen. The number of rotations from + to × and vice versa was always between 1 and 15 per dialogue, with a mean of 7.2 rotations. The timings of the rotations were randomly distributed throughout the videos, but there was always at least 1.25 s between each rotation.

5.3. Tasks

In the experimental session, the participants performed five different tasks: 1) a Semantic (S) task, 2) a Phonological (P) task, 3) a Visual (V) control task, 4) a Shadowing task, and 5) a Motor control task. In the shadowing and motor control tasks (to be reported elsewhere), the participants immediately repeated the speech of the speaker that was of the same gender as the participant (shadowing task) or counted out loud numbers whenever the same-gender speaker was speaking (motor control task). The participants were instructed to ignore the audiobook during all tasks. The fixation crosses were present in all conditions, but the participants were instructed to ignore the crosses in all but the V task. An example of a full dialogue, along with questions related to the S and P tasks can be found in Supplementary Table 2.

In the S task, the participants were instructed to watch and listen to the dialogue and direct their gaze to whoever was speaking (unavoidable eye movements were allowed to keep the situation as realistic as possible). After the presentation of each dialogue video, the participants answered seven 'yes'/'no' statements relating to the content of the dialogue. The statements concerned the semantic content of each of the seven lines in the dialogue (e.g., "One of the speakers had attended a concert") and were presented in order of the lines they concerned.

In the P task, the participants were instructed to listen to the two speakers and search for the phoneme /r/ in the speech stream, the number of which they reported after the dialogue. The reporting was done with seven yes–no statements of the form 'There were *x* occurrences of /r/ in this dialogue', with *x* being '1–2', '3–4', '5–6', '7–8', '9–10', '11–12', and '13–15', presented in this order. The participants were to press 'yes' to the statement with the number interval that contained the number they had counted, and 'no' to other intervals. They were also instructed that if they were unsure about the number, they could press 'yes' to several statements. The number intervals were determined based on the constraints that the number of /r/:s in the dialogues was always between 1 and 15, and the number of the yes–no statements was set to seven across all tasks.

In the V task, the participants were instructed to focus on the fixation cross that was below the face of whichever speaker was speaking and to ignore the dialogue. Their task was to count how many times the crosses rotated from + to ×, and vice versa. After the dialogue, the number of rotations was reported like in the P task (i.e., statements of the form 'There were *x* rotations of the cross' with *x* being '1–2′, '3–4′, '5–6′, '7–8′, '9–10′, '11–12′, and '13–15′, in this order). The cross that was below the face of the speaker not speaking at any one time was not to be attended and never rotated, and the fading out of the cross that was fixated on as a line ended acted as a cue for the participants to shift their gaze onto the other cross. To match the number of task-relevant events across the V and P tasks, the number of cross rotations was matched to the number of occurrences of /r/ in the dialogues.

5.4. Procedure

One or two days before the fMRI session, all participants underwent a training session of approximately one hour. The purpose of this session was to maximize behavioral performance in the scanner by familiarizing the participants with the tasks, as the P and shadowing tasks were deemed somewhat difficult. In the training session, the participants first received instructions concerning the stimuli and the tasks, after which they practiced performing the tasks on a laptop. A set of six videos separate from those used in the actual experiment were used as training stimuli. Of the videos used in the training session, five were 'shuffled' dialogue videos in which the dialogue lines did not form a coherent conversation (these videos were also used in the motor control task not

reported here; see Wikman et al., 2021, for details on the shuffled videos).

The experiment consisted of two runs, both including 20 dialogue blocks. The blocks consisted of one instance of each of the five tasks in each of the four audiovisual quality combinations. The order of tasks within each run was randomized. For all tasks except the motor control task, the dialogue videos were selected from a pool of 36 dialogues originally created for a previous experiment (see Leminen et al., 2020), and they were randomly paired with the tasks. The same dialogue video was never presented more than once within a run.

Before the presentation of each dialogue, instructions on which task the participant was to perform next appeared on the screen. A quiz followed immediately after each video, consisting of seven yes–no statements shown for 2 s each, to which the participants were instructed to answer by pressing a button with their right index finger for 'yes' and with their right middle finger for 'no'. After the statements, feedback on performance was shown (i.e., how many questions out of 7 they answered correctly). The stimulus videos were presented on light gray background. The length of a block was always 85 s, which consisted of 1) instructions (2 s), 2) a gray screen with a fixation cross indicating which speaker would speak first (2–12 s), 3) the dialogue video (55–65 s), 4) a quiz (14 s), and 5) feedback (2 s). The audiobook clips started randomly 500–2000 ms before the video onset and stopped at video offset. A rest block of 40 s occurred between the 10th and 11th blocks. During the rest block, the participants were to look at a small fixation cross in the middle of the screen.

The experiment was controlled using Presentation 20.0 (Neurobehavioral Systems, Berkeley, CA, USA). Stimulus videos were presented with a mirror mounted on the head coil. The approximate size of the video in visual angles was 26° horizontally and 15° vertically, from a viewing distance of approximately 38 cm. Sounds were delivered binaurally through earphones including canal tips that also acted as earplugs (Sensimetrics Model S14; Sensimetrics, Malden, MA, USA). The intensity of the sounds was determined individually so that it was pleasant but loud, approximately 80 dB SPL at the tip of the earphone. Scanner noise (approximately 102 dB SPL, as measured in the head coil) was further attenuated with viscoelastic mattresses around and under the head of the participant inside the coil, and by the earplugs. Verbal responses during the shadowing and motor control tasks (not reported here) were recorded with a noise-canceling MRI safe microphone (FOMRI II, Optoacoustics Ltd., Or-Yehuda, Israel) that was attached to the head coil and reached in front of the mouth of the participant.

5.5. fMRI data acquisition

fMRI imaging was carried out using a 3 Tesla Magnetom Skyra whole-body scanner (Siemens Healthcare, Erlangen, Germany) with a 20-channel head coil. Two functional runs of 703 volumes were acquired per participant, except for the first two participants for whom the runs consisted of 714 volumes (the excessive 11 volumes at the end of a run were deleted for the rest of the participants). Functional data consisted of 43 oblique axial slices of T2*-weighted echo-planar images (EPI; TR 2600 ms, TE 30 ms, flip angle 75°, field of view 192 mm, slice thickness 3.0 mm., 64 × 64 voxel matrix; in-plane resolution 3 mm isotropic). After the functional runs, a high-resolution anatomical image was obtained (MPRAGE sequence, 176 × 256 × 256 voxel matrix, in-plane resolution 1 mm isotropic). Simultaneous electroencephalography (EEG) was recorded from all participants during the fMRI session using a 32-channel MR compatible EEG cap (Braincap MR 32-ch, Easycap, Herrsching, Germany) and an MR compatible amplifier (BrainAmp MR plus, Brain Products GmbH, Gilching, Germany). Unfortunately, these data could not be utilized, as after data collection it was noticed that there was a jitter of tens of milliseconds in the fMRI pulse timings, rendering futile our attempts to remove the effects of magnetic artifacts produced by MRI scanning from the EEG data. Note that simultaneous measurement of fMRI and EEG with a low-density EEG cap has little effect on the signal-to-noise ratio of fMRI data in field strengths of 3 Tesla (Mullinger et al., 2008).

5.6. Behavioral data analysis

Percentage of correct answers per dialogue was used as a measure of task performance in the S task. In the P and V tasks, the distance of the participant's answer from the correct answer was used instead. For example, if a dialogue contained 7 or 8 occurrences of the /r/:s, and the participant answered that there were 5 or 6 occurrences, the distance from the correct answer was 1. This measure was used because it reflects performance more accurately than the simple number of correct and incorrect answers: detecting 7 out of 10 task-relevant events is better than detecting only 5, which is reflected in the distance to the correct answer but not in the simple number of correct answers to the yes–no questions. Note, however, that chance level performance cannot be simply assessed in the distance measure, as it varies depending on the number of task-relevant events per video. Missing responses were treated as errors in all tasks. In none of the participants were the 'yes' responses missing altogether in the P or V task, and the distance could always be defined. Behavioral performance in the S and P tasks was analyzed using separate 2 × 2 repeated-measures ANOVAs conducted with IBM SPSS Statistics 27 (IBM SPSS, Armonk, NY, USA). In these ANOVAs, the two factors were Auditory Quality (poor, good) and Visual Quality (poor, good). The behavioral data for the V task did not meet the assumption of normality of model residuals that repeated measures ANOVA makes. Therefore, whether performance was affected by the auditory and visual qualities in the V task was assessed with the non-parametric Friedman test, also conducted with IBM SPSS Statistics 27 (IBM SPSS, Armonk, NY, USA). The results were visualized with custom-made Python scripts.

5.7. (f)MRI data preprocessing

Preprocessing and first-level analyses of the fMRI data were performed using FEAT (FMRI Expert Analysis Tool) Version 6.00, part of FSL (FMRIB's Software Library, http://www.fmrib.ox.ac.uk/fsl). Registration of fMRI volumes to the high-resolution structural image of the participant was carried out using FLIRT (Jenkinson et al., 2002; Jenkinson and Smith, 2001), and preprocessing included motion correction using MCFLIRT (Jenkinson et al., 2002), slice-timing correction, non-brain removal with BET (Smith, 2002), and highpass temporal filtering (with a cutoff of 130 Hz). For all further fMRI analyses, the data were then projected to the Freesurfer average surface space (fsaverage) using the Freesurfer function mri_vol2surf (Fischl, 2012).

5.8. Whole-brain analysis of fMRI data

In the first-level analysis of the whole-brain data, two separate GLMs were formed and fit to the time series data of each voxel in each run. The first GLM consisted of 28 regressors: one for all lines in each combination of task and audiovisual quality (5 tasks × 2 auditory quality conditions × 2 visual quality conditions), and one for instructions, quizzes, and the six basic motion parameters. For group-level analysis, the main effects and interaction terms of three separate 2 × 2 × 2 repeated-measures ANOVA with factors Task (S vs. V, S vs. P, and P vs. V), Auditory Quality (good, poor), and Visual Quality (good, poor), were built into the first-level GLM. This analysis was used to assess block-level effects of task and audiovisual quality on neural activations (Fig. 3).

The second whole-brain analysis was aimed to assess whether activity levels during the different tasks were dynamically modulated across the time course of a dialogue. This GLM included a separate regressor for each dialogue line in each task and audiovisual quality combination, resulting in a total of 148 regressors (7 lines × 5 tasks × 2 auditory quality conditions × 2 visual quality conditions, plus regressors

for instructions, quizzes, and the six basic motion parameters). In subsequent analyses, only the first six lines of each dialogue were included, because the seventh line was always immediately followed by questions related to the video, which might confound the fit of the model due to a temporal overlap of activations associated with the task during the dialogues and those associated with the quiz. Based on this GLM, contrasts were formed that modeled linear, quadratic, and linear-quadratic trends in the activity differences between two tasks and across the lines of a dialogue. This analysis was again conducted separately for all task pairs (see also Wikman et al., 2021). Main effects and interaction terms (Task × Auditory Quality, Task × Visual Quality, and Task × Auditory Quality × Visual Quality) were built into the first-level GLM to test for these effects in the linear, quadratic, and linear-quadratic contrasts.

In both GLMs, the lines were modeled with a boxcar function starting at the beginning of the line and ending at the end of the line, which was convolved with FSL's gamma function (mean lag 6 s, SD 3) as the hemodynamic response function. In both analyses, all contrasts not including Task as a factor (i.e., main effects of Auditory and Visual Quality and the interaction of Auditory and Visual Quality) were left unanalyzed in order to increase statistical power, as the main focus of this experiment was in assessing task-dependent effects.

Group-level analyses were performed using Freesurfer version 6.0.0 and a one-sample *t*-test performed with the mri_glmfit function. Clusters were defined using permutation inference with the initial cluster forming threshold z set at 3.1. Clusters smaller than 50 mm^2 were discarded. Statistical significance was inferred based on false discovery rate (FDR) corrected cluster statistics across all fMRI comparisons with a threshold of $p < .05$ (Benjamini and Hochberg, 1995).

5.9. Region-of-interest analysis of fMRI data

ROI analysis was used to study activity levels in the orbitofrontal cortex, as well as to visualize interaction, linear, quadratic, and linear-quadratic effects observed in the whole-brain GLMs. For the orbitofrontal cortex, the ROI was defined based on the cortical parcellation of Yeo and colleagues (Yeo et al., 2011), from where the medial orbitofrontal region was selected from each hemisphere (Fig. 5; cortical network '17Networks_10′ in the parcellation). Percent signal change values from the ROIs were extracted using FSL's Featquery and plotted using custom-made Python scripts. The signal change was calculated relative to a resting baseline. For the plots of the linear, quadratic, and linear-quadratic effects, percent signal change values were averaged per line over all audiovisual qualities. For 2–way interaction plots (shown in Electronic Supplementary Materials), the percent signal change values were averaged over the non-significant factor (i.e., when plotting for interactions of Task and Visual Quality, percent signal change values were averaged over the poor and good Auditory Quality conditions).

CRediT authorship contribution statement

Artturi Ylinen: Conceptualization, Methodology, Formal analysis, Investigation, Data curation, Writing – original draft, Writing - review & editing, Visualization. **Patrik Wikman:** Conceptualization, Methodology, Formal analysis, Investigation, Writing - review & editing. **Miika Leminen:** Software. **Kimmo Alho:** Conceptualization, Funding acquisition, Project administration, Resources, Supervision, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We would like to thank Marita Kattelus and Elisa Sahari for their help in data collection, and two anonymous reviewers for their helpful comments on the first version of this manuscript.

Funding

This work received funding from the Academy of Finland (Grant #297848, "Modulation of brain activity patterns during selective attention to speech", 2016–2021) and the Emil Aaltonen Foundation. The funders had no role in study design, collection, analysis, or interpretation of data, or in the writing of the article.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.brainres.2021.147739.

References

Adolphs, R., 2009. The social brain: neural basis of social knowledge. Annu. Rev. Psychol. 60 (1), 693–716. https://doi.org/10.1146/annurev.psych.60.110707.163514.

Ahveninen, J., Jaaskelainen, I.P., Raij, T., Bonmassar, G., Devore, S., Hamalainen, M., Levanen, S., Lin, F.-H., Sams, M., Shinn-Cunningham, B.G., Witzel, T., Belliveau, J. W., 2006. Task-modulated "what" and "where" pathways in human auditory cortex. Proc. Natl. Acad. Sci. U. S. A. 103 (39), 14608–14613. https://doi.org/10.1073/pnas.0510480103.

Alcalá-López, D., Vogeley, K., Binkofski, F., Bzdok, D., 2019. Building blocks of social cognition: mirror, mentalize, share? Cortex 118, 4–18. https://doi.org/10.1016/j.cortex.2018.05.006.

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. Cogn. Brain Res. 7 (3), 335–341. https://doi.org/10.1016/S0926-6410(98)00036-6.

Alho, K., Rinne, T., Herron, T.J., Woods, D.L., 2014. Stimulus-dependent activations and attention-related modulations in the auditory cortex: a meta-analysis of fMRI studies. Hear. Res. 307, 29–41. https://doi.org/10.1016/j.heares.2013.08.001.

Alho, K., Vorobyev, V.A., Medvedev, S.V., Pakhomov, S.V., Roudas, M.S., Tervaniemi, M., van Zuijen, T., Näätänen, R., 2003. Hemispheric lateralization of cerebral blood-flow changes during selective listening to dichotically presented continuous speech. Cogn. Brain Res. 17 (2), 201–211. https://doi.org/10.1016/S0926-6410(03)00091-0.

Alho, K., Vorobyev, V.A., Medvedev, S.V., Pakhomov, S.V., Starchenko, M.G., Tervaniemi, M., Näätänen, R., 2006. Selective attention to human voice enhances brain activity bilaterally in the superior temporal sulcus. Brain Res. 1075 (1), 142–150. https://doi.org/10.1016/j.brainres.2005.11.103.

Arsenault, J.S., Buchsbaum, B.R., 2016. No evidence of somatotopic place of articulation feature mapping in motor cortex during passive speech perception. Psychon. Bull. Rev. 23 (4), 1231–1240. https://doi.org/10.3758/s13423-015-0988-z.

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. Ser. B Methodol. 57 (1), 289–300.

Bent, T., Buchwald, A., Pisoni, D.B., 2009. Perceptual adaptation and intelligibility of multiple talkers for two types of degraded speech. J. Acoust. Soc. Am. 126 (5), 2660–2669. https://doi.org/10.1121/1.3212930.

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. https://doi.org/10.1093/cercor/bhp055.

Binkofski, F.C., Klann, J., Caspers, S., 2016. On the Neuroanatomy and Functional Role of the Inferior Parietal Lobule and Intraparietal Sulcus. In: Hickok, G., Small, S.L. (Eds.), Neurobiology of Language. Academic Press, San Diego, pp. 35–47. https://doi.org/10.1016/B978-0-12-407794-2.00004-3.

Boersma, P., 2001. Praat, a system for doing phonetics by computer. Glot Int. 5, 341–345.

Bohland, J.W., Guenther, F.H., 2006. An fMRI investigation of syllable sequence production. NeuroImage 32 (2), 821–841. https://doi.org/10.1016/j.neuroimage.2006.04.173.

Bradlow, A.R., Nygaard, L.C., Pisoni, D.B., 1999. Effects of talker, rate, and amplitude variation on recognition memory for spoken words. Percept. Psychophys. 61 (2), 206–219. https://doi.org/10.3758/BF03206883.

Buckner, R.L., DiNicola, L.M., 2019. The brain's default network: updated anatomy, physiology and evolving insights. Nat. Rev. Neurosci. 20 (10), 593–608. https://doi.org/10.1038/s41583-019-0212-7.

Bzdok, D., Langner, R., Schilbach, L., Engemann, D.A., Laird, A.R., Fox, P.T., Eickhoff, S., 2013. Segregation of the human medial prefrontal cortex in social cognition. Front. Hum. Neurosci. 7 https://doi.org/10.3389/fnhum.2013.00232.

Chein, J.M., Schneider, W., 2012. The brain's learning and control architecture. Curr. Dir. Psychol. Sci. 21 (2), 78–84.

Cherry, E.C., 1953. Some experiments on the recognition of speech, with one and with two ears. J. Acoust. Soc. Am. 25 (5), 975–979. https://doi.org/10.1121/1.1907229.
Clos, M., Amunts, K., Laird, A.R., Fox, P.T., Eickhoff, S.B., 2013. Tackling the multifunctional nature of Broca's region meta-analytically: Co-activation-based parcellation of area 44. NeuroImage 83, 174–188. https://doi.org/10.1016/j.neuroimage.2013.06.041.
Corbetta, M., Shulman, G.L., 2002. Control of goal-directed and stimulus-driven attention in the brain. Nat. Rev. Neurosci. 3 (3), 201–215. https://doi.org/10.1038/nrn755.
D'Ausilio, A., Pulvermüller, F., Salmas, P., Bufalari, I., Begliomini, C., Fadiga, L., 2009. The motor somatotopy of speech perception. Curr. Biol. 19 (5), 381–385. https://doi.org/10.1016/j.cub.2009.01.017.
Davis, M.H., Johnsrude, I.S., 2003. Hierarchical processing in spoken language comprehension. J. Neurosci. Off. J. Soc. Neurosci. 23 (8), 3423–3431.
Degerman, A., Rinne, T., Salmi, J., Salonen, O., Alho, K., 2006. Selective attention to sound location or pitch studied with fMRI. Brain Res. 1077 (1), 123–134. https://doi.org/10.1016/j.brainres.2006.01.025.
Du, Y., Buchsbaum, B.R., Grady, C.L., Alain, C., 2014. Noise differentially impacts phoneme representations in the auditory and speech motor systems. Proc. Natl. Acad. Sci. U. S. A. 111 (19), 7126–7131. https://doi.org/10.1073/pnas.1318738111.
Duhamel, J.-R., Colby, C.L., Goldberg, M.E., 1998. Ventral intraparietal area of the macaque: congruent visual and somatic response properties. J. Neurophysiol. 79 (1), 126–136. https://doi.org/10.1152/jn.1998.79.1.126.
Emch, M., von Bastian, C.C., Koch, K., 2019. Neural correlates of verbal working memory: an fMRI meta-analysis. Front. Hum. Neurosci. 13, 180. https://doi.org/10.3389/fnhum.2019.00180.
Fadiga, L., Craighero, L., Buccino, G., Rizzolatti, G., 2002. Speech listening specifically modulates the excitability of tongue muscles: a TMS study. Eur. J. Neurosci. 15 (2), 399–402. https://doi.org/10.1046/j.0953-816x.2001.01874.x.
Files, B.T., Tjan, B.S., Jiang, J., Bernstein, L.E., 2015. Visual speech discrimination and identification of natural and synthetic consonant stimuli. Front. Psychol. 6 https://doi.org/10.3389/fpsyg.2015.00878.
Fischl, B., 2012. FreeSurfer. NeuroImage 62, 774–781. https://doi.org/10.1016/j.neuroimage.2012.01.021.
Fridriksson, J., Yourganov, G., Bonilha, L., Basilakos, A., Den Ouden, D.-B., Rorden, C., 2016. Revealing the dual streams of speech processing. Proc. Natl. Acad. Sci. U. S. A. 113 (52), 15108–15113. https://doi.org/10.1073/pnas.1614038114.
Friederici, A.D., Gierhan, S.ME., 2013. The language network. Curr. Opin. Neurobiol Macrocircuits 23 (2), 250–254. https://doi.org/10.1016/j.conb.2012.10.002.
Gentilucci, M., Fogassi, L., Luppino, G., Matelli, M., Camarda, R., Rizzolatti, G., 1988. Functional organization of inferior area 6 in the macaque monkey. Exp. Brain Res. 71 (3), 475–490. https://doi.org/10.1007/BF00248741.
Gierhan, S.M.E., 2013. Connections for auditory language in the human brain. Brain Lang. 127 (2), 205–221. https://doi.org/10.1016/j.bandl.2012.11.002.
Glasser, M.F., Coalson, T.S., Robinson, E.C., Hacker, C.D., Harwell, J., Yacoub, E., Ugurbil, K., Andersson, J., Beckmann, C.F., Jenkinson, M., Smith, S.M., Van Essen, D. C., 2016. A multi-modal parcellation of human cerebral cortex. Nature 536 (7615), 171–178. https://doi.org/10.1038/nature18933.
Goulas, A., Stiers, P., Hutchison, R.M., Everling, S., Petrides, M., Margulies, D.S., 2017. Intrinsic functional architecture of the macaque dorsal and ventral lateral frontal cortex. J. Neurophysiol. 117 (3), 1084–1099. https://doi.org/10.1152/jn.00486.2016.
Harinen, K., Rinne, T., 2013. Activations of human auditory cortex to phonemic and nonphonemic vowels during discrimination and memory tasks. NeuroImage 77, 279–287. https://doi.org/10.1016/j.neuroimage.2013.03.064.
Hickok, G., Poeppel, D., 2007. The cortical organization of speech processing. Nat. Rev. Neurosci. 8 (5), 393–402. https://doi.org/10.1038/nrn2113.
Hickok, G., Poeppel, D., 2004. Dorsal and ventral streams: a framework for understanding aspects of the functional anatomy of language. Cognition 92 (1-2), 67–99. https://doi.org/10.1016/j.cognition.2003.10.011.
Hill, K.T., Miller, L.M., 2010. Auditory attentional control and selection during cocktail party listening. Cereb. Cortex 20 (3), 583–590. https://doi.org/10.1093/cercor/bhp124.
Huth, A.G., de Heer, W.A., Griffiths, T.L., Theunissen, F.E., Gallant, J.L., 2016. Natural speech reveals the semantic maps that tile human cerebral cortex. Nature 532 (7600), 453–458. https://doi.org/10.1038/nature17637.
Ito, T., Tiede, M., Ostry, D.J., 2009. Somatosensory function in speech perception. Proc. Natl. Acad. Sci. U. S. A. 106 (4), 1245–1248. https://doi.org/10.1073/pnas.0810063106.
Jääskeläinen, I.P., 2010. The role of speech production system in audiovisual speech perception. Open Neuroimaging J. 4, 30–36. https://doi.org/10.2174/1874440001004020030.
Jefferies, E., 2013. The neural basis of semantic cognition: Converging evidence from neuropsychology, neuroimaging and TMS. Cortex 49 (3), 611–625. https://doi.org/10.1016/j.cortex.2012.10.008.
Jenkinson, M., Bannister, P., Brady, M., Smith, S., 2002. Improved optimization for the robust and accurate linear registration and motion correction of brain images. NeuroImage 17, 825–841. https://doi.org/10.1016/s1053-8119(02)91132-8.
Jenkinson, M., Smith, S., 2001. A global optimisation method for robust affine registration of brain images. Med. Image Anal. 5 (2), 143–156. https://doi.org/10.1016/S1361-8415(01)00036-6.
Kansaku, K., Carver, B., Johnson, A., Matsuda, K., Sadato, N., Hallett, M., 2007. The role of the human ventral premotor cortex in counting successive stimuli. Exp. Brain Res. 178 (3), 339–350. https://doi.org/10.1007/s00221-006-0736-8.
Kansaku, K., Johnson, A., Grillon, M.-L., Garraux, G., Sadato, N., Hallett, M., 2006. Neural correlates of counting of sequential sensory and motor events in the human brain. NeuroImage 31 (2), 649–660. https://doi.org/10.1016/j.neuroimage.2005.12.023.
Kilgard, M.P., 2012. Harnessing plasticity to understand learning and treat disease. Trends Neurosci. 35 (12), 715–722. https://doi.org/10.1016/j.tins.2012.09.002.
Leech, R., Sharp, D.J., 2014. The role of the posterior cingulate cortex in cognition and disease. Brain J. Neurol. 137, 12–32. https://doi.org/10.1093/brain/awt162.
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 https://doi.org/10.3389/fnins.2020.00436.
Leonard, M.K., Chang, E.F., 2014. Dynamic speech representations in the human temporal lobe. Trends Cogn. Sci. 18 (9), 472–479. https://doi.org/10.1016/j.tics.2014.05.001.
Liberman, A.M., Mattingly, I.G., 1985. The motor theory of speech perception revised. Cognition 21 (1), 1–36. https://doi.org/10.1016/0010-0277(85)90021-6.
Luppino, G., Murata, A., Govoni, P., Matelli, M., 1999. Largely segregated parietofrontal connections linking rostral intraparietal cortex (areas AIP and VIP) and the ventral premotor cortex (areas F5 and F4). Exp. Brain Res. 128, 181–187. https://doi.org/10.1007/s002210050833.
Manoach, D.S., Schlaug, G., Siewert, B., Darby, D.G., Bly, B.M., Benfield, A., Edelman, R. R., Warach, S., 1997. Prefrontal cortex fMRI signal changes are correlated with working memory load. NeuroReport 8 (2), 545–549.
Matelli, M., Luppino, G., 2001. Parietofrontal circuits for action and space perception in the macaque monkey. NeuroImage 14 (1), S27–S32. https://doi.org/10.1006/nimg.2001.0835.
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 (5), 762–776. https://doi.org/10.1016/j.neuropsychologia.2012.01.010.
Mesgarani, N., Chang, E.F., 2012. Selective cortical representation of attended speaker in multi-talker speech perception. Nature 485 (7397), 233–236. https://doi.org/10.1038/nature11020.
Mesgarani, N., Cheung, C., Johnson, K., Chang, E.F., 2014. Phonetic feature encoding in human superior temporal gyrus. Science 343 (6174), 1006–1010. https://doi.org/10.1126/science:1245994.
Miller, G.A., 1947. The masking of speech. Psychol. Bull. 44, 105–129. https://doi.org/10.1037/h0055960.
Mitchell, J.P., Macrae, C.N., Banaji, M.R., 2006. Dissociable medial prefrontal contributions to judgments of similar and dissimilar others. Neuron 50 (4), 655–663. https://doi.org/10.1016/j.neuron.2006.03.040.
Möttönen, R., Dutton, R., Watkins, K.E., 2013. Auditory-motor processing of speech sounds. Cereb. Cortex 23 (5), 1190–1197. https://doi.org/10.1093/cercor/bhs110.
Möttönen, R., Järveläinen, J., Sams, M., Hari, R., 2005. Viewing speech modulates activity in the left SI mouth cortex. NeuroImage 24 (3), 731–737. https://doi.org/10.1016/j.neuroimage.2004.10.011.
Möttönen, R., Watkins, K.E., 2009. Motor representations of articulators contribute to categorical perception of speech sounds. J. Neurosci. 29 (31), 9819–9825. https://doi.org/10.1523/JNEUROSCI.6018-08.2009.
Mugler, E.M., Tate, M.C., Livescu, K., Templer, J.W., Goldrick, M.A., Slutzky, M.W., 2018. Differential representation of articulatory gestures and phonemes in precentral and inferior frontal gyri. J. Neurosci. Off. J. Soc. Neurosci. 38 (46), 9803–9813. https://doi.org/10.1523/JNEUROSCI.1206-18.2018.
Mullinger, K., Debener, S., Coxon, R., Bowtell, R., 2008. Effects of simultaneous EEG recording on MRI data quality at 1.5, 3 and 7 Tesla. Int. J. Psychophysiol. Off. J. Int. Organ. Psychophysiol. 67, 178–188. https://doi.org/10.1016/j.ijpsycho.2007.06.008.
Murakami, T., Kell, C.A., Restle, J., Ugawa, Y., Ziemann, U., 2015. Left dorsal speech stream components and their contribution to phonological processing. J. Neurosci. 35 (4), 1411–1422. https://doi.org/10.1523/JNEUROSCI.0246-14.2015.
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 (2), 201–233. https://doi.org/10.1017/S0140525X00078407.
Oldfield, R.C., 1971. The assessment and analysis of handedness: The Edinburgh inventory. Neuropsychologia 9 (1), 97–113. https://doi.org/10.1016/0028-3932(71)90067-4.
Osnes, B., Hugdahl, K., Specht, K., 2011. Effective connectivity analysis demonstrates involvement of premotor cortex during speech perception. NeuroImage 54 (3), 2437–2445. https://doi.org/10.1016/j.neuroimage.2010.09.078.
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 (6), 1195–1209.e3. https://doi.org/10.1016/j.neuron.2019.09.007.
Papoutsi, M., de Zwart, J.A., Jansma, J.M., Pickering, M.J., Bednar, J.A., Horwitz, B., 2009. From phonemes to articulatory codes: an fMRI study of the role of Broca's area in speech production. Cereb. Cortex 19, 2156–2165. https://doi.org/10.1093/cercor/bhn239.
Peelle, J.E., Sommers, M.S., 2015. Prediction and constraint in audiovisual speech perception. Cortex 68, 169–181. https://doi.org/10.1016/j.cortex.2015.03.006.
Petkov, C.I., Kang, X., Alho, K., Bertrand, O., Yund, E.W., Woods, D.L., 2004. Attentional modulation of human auditory cortex. Nat. Neurosci. 7 (6), 658–663. https://doi.org/10.1038/nn1256.
Price, C.J., 2012. A review and synthesis of the first 20years of PET and fMRI studies of heard speech, spoken language and reading. NeuroImage 62 (2), 816–847. https://doi.org/10.1016/j.neuroimage.2012.04.062.

Pulvermuller, F., Huss, M., Kherif, F., Moscoso del Prado Martin, F., Hauk, O., Shtyrov, Y., 2006. Motor cortex maps articulatory features of speech sounds. Proc. Natl. Acad. Sci. U. S. A. 103 (20), 7865–7870. https://doi.org/10.1073/pnas.0509989103.

Pulvermüller, F., Moseley, R.L., Egorova, N., Shebani, Z., Boulenger, V., 2014. Motor cognition–motor semantics: action perception theory of cognition and communication. Neuropsychologia 55, 71–84. https://doi.org/10.1016/j.neuropsychologia.2013.12.002.

Rauschecker, J.P., Scott, S.K., 2009. Maps and streams in the auditory cortex: nonhuman primates illuminate human speech processing. Nat. Neurosci. 12 (6), 718–724. https://doi.org/10.1038/nn.2331.

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 (42), 13338–13343. https://doi.org/10.1523/JNEUROSCI.3012-09.2009.

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 (2), 181–194. https://doi.org/10.1007/s00429-007-0152-2.

Sanchez Panchuelo, R.M., Besle, J., Schluppeck, D., Humberstone, M., Francis, S., 2018. Somatotopy in the Human Somatosensory System. Front. Hum. Neurosci. 12, 235. https://doi.org/10.3389/fnhum.2018.00235.

Scheich, H., Brechmann, A., Brosch, M., Budinger, E., Ohl, F.W., 2007. The cognitive auditory cortex: Task-specificity of stimulus representations. Hear. Res. 229 (1-2), 213–224. https://doi.org/10.1016/j.heares.2007.01.025.

Schomers, M.R., Pulvermüller, F., 2016. Is the sensorimotor cortex relevant for speech perception and understanding? An integrative review. Front. Hum. Neurosci. 10, 435. https://doi.org/10.3389/fnhum.2016.00435.

Shannon, R.V., Zeng, F.-G., Kamath, V., Wygonski, J., Ekelid, M., 1995. Speech recognition with primarily temporal cues. Science 270 (5234), 303–304. https://doi.org/10.1126/science:270.5234.303.

Shinn-Cunningham, B.G., 2008. Object-based auditory and visual attention. Trends Cogn. Sci. 12 (5), 182–186. https://doi.org/10.1016/j.tics.2008.02.003.

Smith, S.M., 2002. Fast robust automated brain extraction. Hum. Brain Mapp. 17 (3), 143–155. https://doi.org/10.1002/hbm.10062.

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 (5), 495–502. https://doi.org/10.1016/S0730-725X(00)00128-4.

Stilp, C.E., Theodore, R.M., 2020. Talker normalization is mediated by structured indexical information. Atten. Percept. Psychophys. 82 (5), 2237–2243. https://doi.org/10.3758/s13414-020-01971-x.

Sumby, W.H., Pollack, I., 1954. Visual contribution to speech intelligibility in noise. J. Acoust. Soc. Am. 26 (2), 212–215. https://doi.org/10.1121/1.1907309.

Teder, W., Kujala, T., Näätänen, R., 1993. Selection of speech messages in free-field listening. NeuroReport 5 (3), 307–309. https://doi.org/10.1097/00001756-199312000-00032.

Treisman, A.M., 1964. Verbal cues, language, and meaning in selective attention. Am. J. Psychol. 77, 206–219. https://doi.org/10.2307/1420127.

Tschentscher, N., Hauk, O., Fischer, M.H., Pulvermüller, F., 2012. You can count on the motor cortex: finger counting habits modulate motor cortex activation evoked by numbers. NeuroImage 59 (4), 3139–3148. https://doi.org/10.1016/j.neuroimage.2011.11.037.

Turkeltaub, P.E., Branch, H.B., 2010. Localization of sublexical speech perception components. Brain Lang. 114 (1), 1–15. https://doi.org/10.1016/j.bandl.2010.03.008.

Venezia, J.H., Vaden, K.I., Rong, F., Maddox, D., Saberi, K., Hickok, G., 2017. Auditory, visual and audiovisual speech processing streams in superior temporal sulcus. Front. Hum. Neurosci. 11 https://doi.org/10.3389/fnhum.2017.00174.

Vigneau, M., Beaucousin, V., Hervé, P.Y., Duffau, H., Crivello, F., Houdé, O., Mazoyer, B., Tzourio-Mazoyer, N., 2006. Meta-analyzing left hemisphere language areas: phonology, semantics, and sentence processing. NeuroImage 30 (4), 1414–1432. https://doi.org/10.1016/j.neuroimage.2005.11.002.

Vigneau, M., Beaucousin, V., Hervé, P.-Y., Jobard, G., Petit, L., Crivello, F., Mellet, E., Zago, L., Mazoyer, B., Tzourio-Mazoyer, N., 2011. What is right-hemisphere contribution to phonological, lexico-semantic, and sentence processing?: Insights from a meta-analysis. NeuroImage 54 (1), 577–593. https://doi.org/10.1016/j.neuroimage.2010.07.036.

Wikman, P., Rinne, T., 2019. Interaction of the effects associated with auditory-motor integration and attention-engaging listening tasks. Neuropsychologia 124, 322–336. https://doi.org/10.1016/j.neuropsychologia.2018.11.006.

Wikman, P., Sahari, E., Salmela, V., Leminen, A., Leminen, M., Laine, M., Alho, K., 2021. Breaking down the cocktail party: Attentional modulation of cerebral audiovisual speech processing. NeuroImage 224, 117365. https://doi.org/10.1016/j.neuroimage.2020.117365.

Wikman, P.A., Vainio, L., Rinne, T., 2015. The effect of precision and power grips on activations in human auditory cortex. Front. Neurosci. 9 https://doi.org/10.3389/fnins.2015.00378.

Wild, C.J., Yusuf, A., Wilson, D.E., Peelle, J.E., Davis, M.H., Johnsrude, I.S., 2012. Effortful listening: the processing of degraded speech depends critically on attention. J. Neurosci. 32 (40), 14010–14021. https://doi.org/10.1523/JNEUROSCI.1528-12.2012.

Woods, D.L., Hillyard, S.A., Hansen, J.C., 1984. Event-related brain potentials reveal similar attentional mechanisms during selective listening and shadowing. J. Exp. Psychol. Hum. Percept. Perform. 10 (6), 761–777. https://doi.org/10.1037//0096-1523.10.6.761.

Yeo, B.T., Krienen, F.M., Sepulcre, J., Sabuncu, M.R., Lashkari, D., Hollinshead, M., Roffman, J.L., Smoller, J.W., Zöllei, L., Polimeni, J.R., Fischl, B., Liu, H., Buckner, R. L., 2011. The organization of the human cerebral cortex estimated by intrinsic functional connectivity. J. Neurophysiol. 106 (3), 1125–1165. https://doi.org/10.1152/jn.00338.2011.

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 (8), 1683–1688. https://doi.org/10.1097/00001756-199906030-00011.

Zatorre, R.J., Mondor, T.A., Evans, A.C., 1999. Auditory attention to space and frequency activates similar cerebral systems. NeuroImage 10 (5), 544–554. https://doi.org/10.1006/nimg.1999.0491.

Zion Golumbic, E., Ding, N., Bickel, S., Lakatos, P., Schevon, C., McKhann, G., Goodman, R., Emerson, R., Mehta, A., Simon, J., Poeppel, D., Schroeder, C., 2013. Mechanisms underlying selective neuronal tracking of attended speech at a “cocktail party”. Neuron 77 (5), 980–991. https://doi.org/10.1016/j.neuron.2012.12.037.