Contents lists available at ScienceDirect

NeuroImage

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

Multivariate fMRI responses in superior temporal cortex predict visual contributions to, and individual differences in, the intelligibility of noisy speech

Yue Zhang [a,b], Johannes Rennig [c], John F Magnotti [a,1], Michael S Beauchamp [a,*,1]

[a] Department of Neurosurgery, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, United States
[b] Department of Neurosurgery, Baylor College of Medicine, Houston, TX, United States
[c] Division of Neuropsychology, Center of Neurology, Hertie-Institute for Clinical Brain Research, University of Tübingen, Tübingen, Germany

ARTICLE INFO

Keywords:
Speech perception
Audiovisual multisensory
Bold fMRI multivariate

ABSTRACT

Humans have the unique ability to decode the rapid stream of language elements that constitute speech, even when it is contaminated by noise. Two reliable observations about noisy speech perception are that seeing the face of the talker improves intelligibility and the existence of individual differences in the ability to perceive noisy speech. We introduce a multivariate BOLD fMRI measure that explains both observations. In two independent fMRI studies, clear and noisy speech was presented in visual, auditory and audiovisual formats to thirty-seven participants who rated intelligibility. An event-related design was used to sort noisy speech trials by their intelligibility. Individual-differences multidimensional scaling was applied to fMRI response patterns in superior temporal cortex and the dissimilarity between responses to clear speech and noisy (but intelligible) speech was measured. Neural dissimilarity was less for audiovisual speech than auditory-only speech, corresponding to the greater intelligibility of noisy audiovisual speech. Dissimilarity was less in participants with better noisy speech perception, corresponding to individual differences. These relationships held for both single word and entire sentence stimuli, suggesting that they were driven by intelligibility rather than the specific stimuli tested. A neural measure of perceptual intelligibility may aid in the development of strategies for helping those with impaired speech perception.

1. Introduction

The most important form of human communication is spoken language. Speech is only useful to the extent that it is intelligible, leading to pioneering studies by scientists at Bell Labs who quantified intelligibility to improve telephone equipment (French and Steinberg, 1947). Since then, an enormous literature on speech intelligibility has developed that spans research, education, and clinical practice in the treatment of speech and hearing disorders, reviewed in Weismer (2008). Two findings in this literature are of particular interest. The first finding is that visual information from the talker's face increases speech intelligibility under noisy listening conditions (Sumby and Pollack, 1954), reviewed in Peelle and Sommers (2015a). The second finding is that, even in adults with normal hearing thresholds, there is large individual variability in the ability to understand speech under difficult listening conditions such as the presence of background noise, reviewed in Shinn-Cunningham (2017).

The development of non-invasive blood oxygen level dependent functional magnetic resonance imaging (BOLD fMRI) has spurred extensive investigations into the functional anatomy of language. An influential model holds that bilateral superior temporal cortex (STC) transforms incoming auditory information into speech representations (Hickok and Poeppel, 2004). Consistent with this model, many neuroimaging studies have found increased STC activity for intelligible speech, reviewed in Abrams et al. (2013), Davis and Johnsrude (2007), DeWitt and Rauschecker (2012), Evans (2017), Evans et al. (2014). The STC is also strongly implicated in audiovisual speech processing, reviewed in Bernstein and Liebenthal (2014), Ozker et al. (2018).

A major advance in fMRI was the development of multivariate analysis techniques to reveal information hidden from univariate

* Corresponding author.
E-mail address: beaucha@upenn.edu (M.S. Beauchamp).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.neuroimage.2023.120271
Received 7 March 2023; Received in revised form 20 June 2023; Accepted 6 July 2023

analyses (Haynes and Rees, 2006; Kriegeskorte et al., 2006; Norman et al., 2006). Previous multivariate studies of auditory-only speech demonstrated that response patterns in STC are related to speech intelligibility (Abrams et al., 2013; Evans et al., 2014; Evans and Davis, 2015; McGettigan et al., 2012a; Okada et al., 2010). A limitation of these studies was that intelligibility was usually manipulated in an all-or-none fashion by comparing across two conditions, one in which speech was always intelligible (such as clear speech) and one in which speech was never intelligible (such as spectrally-rotated speech). In a recent study, we took a different approach by *post hoc* sorting fMRI trials based on participant responses (Rennig et al., 2020) an approach first used to study memory (Brewer et al., 1998; Wagner et al., 1998). This allowed for a comparison of the multivariate pattern of fMRI activity evoked by intelligible compared with unintelligible trials, within a single noisy speech condition.

Multivariate dissimilarity between fMRI response patterns has a close relationship with perception (Kriegeskorte and Kievit, 2013), suggesting that it could be a useful tool for interrogating visual contributions to, and individual differences in, noisy speech perception. However, previous studies on individual differences tested only univariate analyses, with conflicting results. An fMRI study of young adults conducted at 1.5 T identified three single voxels, in frontal cortex and temporal cortex, with a significant univariate BOLD signal correlation with intelligibility, as measured with a post-test conducted after the MRI session (McGettigan et al., 2012b). An fMRI study of young and old adults conducted at 3 T found no voxels with a significant univariate BOLD signal correlation with intelligibility in young adults, while old adults showed significant correlations in two brain locations in frontal cortex and sensorimotor cortex (Du et al., 2016). A scalp EEG study found that the amplitude of the event-related potential in supramarginal gyrus showed a correlation with individual differences (Kim et al., 2021).

To examine whether multivariate analysis could resolve these conflicting results, we developed a multivariate neural dissimilarity measure. The measure used *post hoc* trial sorting to measure the fMRI response to noisy but intelligible speech. The dissimilarity between the response patterns to clear speech and noisy but intelligible speech was measured separately for each participant, using an individual-differences multidimensional scaling approach applied to fMRI data collected from 37 young adults. To establish generalizability, the neural-perceptual relationship was assessed for both noisy auditory-only and audiovisual speech and for single words and complete sentences.

2. Methods

2.1. Overview

Two datasets were analyzed, both consisting of fMRI and behavioral data collected from healthy human participants presented with speech in five different formats: clear auditory speech paired with a video of a talking face (AcV), noisy auditory speech paired with a video of a talking face (AnV), clear auditory speech without a talking face/black screen (Ac), noisy auditory speech without a talking face/black screen (An), and a video of a talking face without audio (V). As described below, stimuli in the AnV and An conditions were sorted by perceptual ratings into trials that were intelligible ("-Y") or not intelligible ("-N"), producing a total of seven conditions (AcV, AnV-Y, AnV-N, Ac, An-Y, An-N, V). The speech stimuli were either single words or sentences. The fMRI dataset for the word study has not been previously published. The fMRI dataset for the sentence study is a re-analysis of published data (Rennig and Beauchamp, 2022).

2.2. Reliability and reproducibility

To promote reliability and reproducibility, the analysis code and data and additional stimulus information may be found in the attached files, *Supplementary_Analysis.html* and *SummaryData.xlsx*. The files are organized by manuscript order.

2.3. Human subjects

All experiments were approved by the Committee for the Protection of Human Subjects at Baylor College of Medicine, Houston, TX. For the word study, fifteen healthy right-handed participants (5 males, mean age 22 yrs, range 18 – 37 yrs) with normal or corrected-to-normal vision and normal hearing provided written informed consent. For the sentence study, twenty-two healthy right-handed participants (14 females, mean age 25, range 18 - 34) with normal or corrected to normal vision and normal hearing were consented. There was no overlap between the participant groups, for a total *n* of 37 participants.

2.4. Speech stimuli

For the word study, the stimuli consisted of 297 single English words recorded from 12 talkers (6 males). For the sentence study, the stimuli consisted of 80 sentences recorded from a single male talker and generously shared by Van Engen et al. (2017). All stimuli are listed in *SummaryData.xlsx*. The visual angle subtended by the face videos in the MR scanner was approximately 20° and the sound pressure level was approximately 80 dB SPL.

To create auditory noisy speech, the original audio recordings were combined with pink noise. Pink noise is commonly used in studies of auditory function because it contains decreasing energy at increasing frequency, making it less aversive than white noise. Pink noise and the stimulus audio track were normalized by the absolute value of the respective maximum, audio $\text{track}_{\text{normalized}}$ = audio $\text{track}_{\text{native}}$/max(abs(audio $\text{track}_{\text{native}}$)). The power of the signal in the sentence audio track and the pink noise were determined and the signal-to-noise ratio (SNR) calculated as $\log 10(\text{power}_{\text{signal}}/\text{power}_{\text{noise}})$. The volume of the pink noise was increased or decreased iteratively to reach an SNR of −8 dB SPL for words and −16 dB SPL for sentences. The audio track and pink noise were then summed and then re-normalized to equalize the volume across stimuli. The original clear auditory recordings were normalized to equate root-mean-square amplitude across stimuli, but no compression or normalization was done within each stimulus.

Videos were edited using Adobe Premiere Pro. For auditory-only stimuli (Ac, An) the visual component of the stimulus consisted of a fixation crosshair in the center of a gray screen of the approximate same luminance as the face stimulus.

Acoustic noise induces different consequences on intelligibility not only according to the signal-to-noise ratio (SNR) but also according to the linguistic element (syllable, word or sentence). For a given SNR, the intelligibility of a word varies according to its phonetic and phonotactic structure. For a sentence, the misperception of some linguistic elements does not prevent a substantial understanding of the sentence.

2.5. Perceptual task

Perceptual data was collected in the MR scanner with no feedback. Participants rated their understanding of each stimulus with a button press. For word stimuli, the task instruction was “are you sure what the word was” and participants pressed one of two button to signal “I am sure of the word” or “I am not sure of the word”. For sentence stimuli, participants pressed one of three buttons to signal “understood everything” (all words in the sentence understood); “understood something” (at least one word in the sentence); “understood nothing” (no words in the sentence). There were very few "understood everything" responses, so "understood everything" and "understood something" responses were grouped for analysis.

This resulted in two types of *post hoc* perceptually-sorted trials for both studies: "intelligible" (participant sure of word, or understood some or everything for sentences) and "unintelligible" (unsure of word,

nothing understood for sentences). Here we use "intelligible" in the colloquial sense of "comprehensible" or "able to be understood".

To minimize perceptual learning, stimuli were never repeated within participants. The order and format of each stimulus was randomized across participants to counterbalance any stimulus effects. For the word stimuli, this was accomplished by dividing the words into five groups. Participants were randomized into one of five batches. Each batch was presented with each word group in a different format (Ac, An, AnV, AcV, V) so that, across participants, each word was presented in every different format an equal number of times.

2.6. Experimental design

For both studies, a rapid event-related fMRI design was used. The trial (event) duration was 3.04 s for the word study and 6.0 s for the sentence study. In the word study, there was a silent period of 2 s within each trial. Words were presented during this silent period to avoid auditory contamination by the loud echo-planar pulse sequence. Following word presentation, participants responded when MR acquisition commenced, resulting in one MR acquisition per trial. In the sentence study, stimuli were presented simultaneous with the MR acquisition without a silent period, resulting in two MR acquisitions per trial. Participants in the word study were presented with a total of 297 word trials and sentence participants were presented with a total of 160 sentence trials. However, the amount of fMRI data collected and total time in the scanner was comparable between studies since sentence trials were twice as long but contained two MR acquisitions instead of one.

Trials were ordered in a pseudo-random optimal sequence generated by the program optseq2 (Dale et al., 1999; https://surfer.nmr.mgh.harvard.edu/optseq). For the word study, six scan series with 65 trials each (390 total trials) were collected from each participant. To present each of the 297 words exactly once, three series contained 49 trials and three contained 50. Each scan series also contained 15 or 16 trials of fixation baseline to permit estimation of the amplitude of the stimulus evoked hemodynamic response relative to fixation. For the sentence study, four scan series (300 s long each) were acquired, each contained 40 trials interspersed with a total of 60 s of fixation baseline. For the word study, the efficiency after post-hoc sorting of trials was 0.0015.

2.7. fMRI acquisition and stimulus presentation

Both word and sentence studies used a Siemens 3 tesla MR scanner in Baylor College of Medicine's Core for Advanced MRI (CAMRI). The sentence study used a TRIO with a 32-channel head coil. Prior to initiation of the word study, the TRIO was upgraded to a PRISMA FIT with a 64-channel head coil. The sentence study used a TR of 1.5 s (TE = 30 ms, flip angle = 72°, in-plane resolution of 2 × 2 mm, 69 2 mm axial slices, multiband factor: 3, GRAPPA factor: 2). The word study used a TR of 3.04 s (TE = 38 ms, flip angle = 78°, in-plane resolution of 2 × 2 mm, 65 2 mm axial slices, multiband acceleration factor: 6). The multiband acquisition permitted data collection in 1.04 s. In combination with a TR of 3.04 s and clustered/sparse acquisition (Edmister et al., 1999; Hall et al., 1999) this provided a silent window of 2 s between acquisitions.

For both studies, stimuli were presented and synchronized with MR data acquisition in Matlab (The Mathworks, Inc., Natick, MA, USA) using the Psychophysics Toolbox extensions (Brainard, 1997). Visual stimuli were presented on a 32-inch (1920 pixels by 1080 pixels) MR-compatible BOLDview LCD screen placed behind the bore of the MR scanner and viewed through a mirror attached to the head coil. Auditory stimuli were played via MR-compatible noisy reduction headphones. Behavioral responses were collected using a fiber-optic button response pad (Current Designs, Haverford, PA, USA).

2.8. Structural MRI acquisition and analysis

The anatomical scan series consisted of two T1-weighted MPRAGE anatomical volumes. Functional data was collected using a multi-slice echo planar imaging sequence (Perrachione and Ghosh, 2013): TR = 3040 ms, TE = 38 ms, flip angle = 78°, in-plane resolution of 2 × 2 mm, 65 2 mm axial slices, multiband accelerate factor: 6, phase encoding direction: Anterior-to-Posterior. Alternate EPI scans were collected using the opposing phase encoding direction (Posterior-to-Anterior).

The second anatomical volume was aligned to the first using a 6-parameter affine transformation with a mutual information cost function using the AFNI program *3dAllineate*. The aligned volumes were averaged to improve gray-white contrast and FreeSurfer was used to construct a cortical surface model (Dale et al., 1999a) which was visualized with the AFNI program *SUMA* (Argall et al., 2006).

To minimize patient fatigue, total scan time including all functional and structural acquisitions was approximately 30 min.

2.9. fMRI analysis

fMRI analysis was conducted using the Analysis of Functional NeuroImages (AFNI) package (Cox, 1996). Preprocessing consisted of susceptibility distortion correction; slice-time correction; and motion correction by aligning the EPIs to the average anatomical image. The time series of each voxel was scaled to have a mean of 100 so that all signal changes were automatically in units of percent difference from the mean.

A generalized linear model (GLM) was applied to the MR time series in each voxel using 3dDeconvolve. To estimate the amplitude of the activation in each voxel, the time of each stimulus event was convolved with a gamma-variate hemodynamic response function. Stimuli in the AnV and An conditions were *post-hoc* sorted by perceptual ratings into trials that were intelligible ("-Y") or not intelligible ("-N"), producing a GLM with seven regressors of interest (AcV, AnV-Y, AnV-N, Ac, An-Y, An-N, V). Regressors of no interest consisted of a polynomial to model baseline fluctuations and six mean-subtracted motion estimates from motion correction.

2.10. ROI construction

Regions of interest (ROIs) were defined individually for each hemisphere based on the automated parcellation of the cortical surface (Destrieux et al., 2010). For the main analysis, five relevant FreeSurfer atlas labels per hemisphere (10 total labels per participant) were combined into a single temporal cortex ROI: superior temporal gyrus (STG); superior temporal sulcus (STS); transverse superior temporal gyrus (also known as Heschl's Gyrus, HG); transverse superior temporal sulcus (also known as Heschl's sulcus, HS); planum temporale (PT). For one hemisphere in one participant (right hemisphere of participant QP) FreeSurfer parcellation failed to identify two atlas labels in the right hemisphere, resulting in a total of 8 total labels for this participant. For the secondary analysis, six subregions of the superior temporal cortex ROI were analyzed separately, consisting of three subregions in the left hemisphere and three in the right hemisphere. The subregions were the STG; the STS; and the combination of HG, HS, and PT (HG+).

The STC (and STC subregions) were defined solely using the anatomical FreeSurfer parcellation, without any additional functional thresholding. Therefore, for each participant, the fMRI pattern comparisons were always conducted within ROIs of exactly the same size.

Voxels with absolute percent signal change exceeding 2.5% over the course of an MR scan series were excluded from the mask (<1% of voxels); most of these extreme-valued voxels were at the very edge of the brain, making their high signal change likely a result of motion or vascular artifacts.

2.11. Multivariate fMRI analysis

For multivariate fMRI analysis, the mean percentage signal change across conditions in each surface node was calculated and subtracted from the response to each individual condition. This increases the dynamic range of the fMRI pattern correlation (Haxby et al., 2001) and is especially important in superior temporal cortex where many nodes show a positive response to all speech stimuli (Rennig and Beauchamp, 2022). To compute the fMRI pattern similarity between each pair of conditions, the mean-centered percentage signal change across the ROI for the first condition was correlated with the mean-centered percentage signal change in the second condition, resulting in a single correlation value for each pair of conditions for each hemisphere. There were 7 conditions, resulting in 21 pairwise correlation per hemisphere. The correlations were converted into dissimilarities using the formula sqrt (1-*r*). Next, individual-differences multidimensional scaling (MDS) was used to decompose the dissimilarity matrices from each participant into two dimensions. This was performed using the Carroll-Chang decomposition (Carroll and Chang, 1970) also known as IDIOSCAL (Individual DIfferences in Orientation SCALing) as implemented in the *smacof* package (Leeuw and Mair, 2009). IDIOSCAL provides an advantage over simple averaging correlation matrices across participants because it simultaneously estimates individual representational spaces and the group space (analogous to treating participants as a random factor rather than a fixed factor). The group MDS revealed two separate drivers of the fMRI response patterns. For visualization in Fig. 1, the MDS space was rotated so that these two drivers lay along the x-axis and the y-axis (distances are invariant to rotation).

2.12. Neural-perceptual correlations

To create a neural measure of intelligibility, the Euclidean distance in MDS space between clear and noisy speech was measured in each participant, separately for audiovisual speech (distance between AcV and AnV-Y) and auditory-only speech (distance between Ac and An-Y).

The neural-perceptual relationship was modeled using generalized

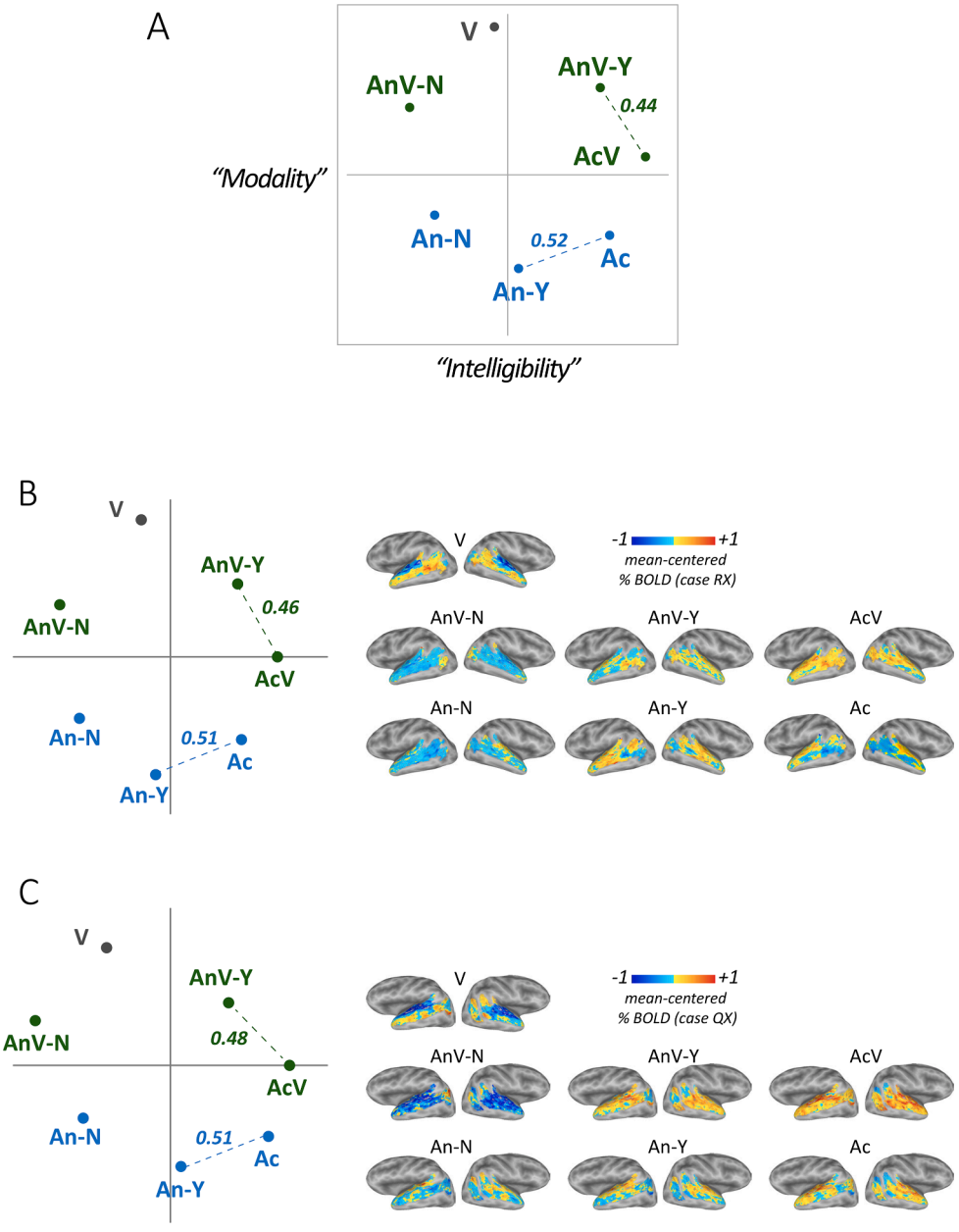

Fig. 1. A. Brain responses to seven kinds of speech were measured with BOLD fMRI: visual-only speech (V); audiovisual speech with added auditory noise, sorted by whether participants rated it as unintelligible (AnV-N) or intelligible (AnV-Y); clear audiovisual speech (AcV); noisy auditory-only speech sorted by intelligibility (An-N, An-Y); clear auditory-only speech (Ac). Individual-differences multidimensional scaling (MDS) was used to decompose the fMRI response patterns in superior temporal cortex (STC). Each colored symbol shows the location of one speech condition in MDS space. The MDS axis labels are descriptive, highlighting the separation by stimulus modality (colors: blue for auditory-only, green for audiovisual, black for visual-only) and intelligibility (more intelligible stimuli at the right of the space). Green dashed line shows the dissimilarity between clear and noisy but intelligible audiovisual speech, the two most similar patterns (number shows distance in MDS space). Blue dashed line shows the dissimilarity between clear and noisy but intelligible auditory-only speech.
B. fMRI response patterns to the different speech conditions in individual participant RX (stimulus material consisted of single words). Left panel shows the MDS decomposition, right panel shows the fMRI response patterns. At each STC location, the mean response across speech conditions was calculated and subtracted from the response, so that cool colors indicate responses below the mean speech response (rather than below fixation baseline).
C. Individual-differences MDS decomposition of fMRI response patterns to the different speech conditions in individual participant QX (stimulus material consisted of entire sentences). .

mixed-effects models (Yu et al., 2022), equivalent to logistic regression with a random effect of participant as implemented in *lme4* (Bates et al., 2015). The dependent measure was the count of intelligible trials *vs.* unintelligible trials. The fixed effect was the neural distance between clear and intelligible noisy speech. The random factor was participant. To assess the predictive accuracy of the model, we correlated the intelligibility rating predicted by the model (without the random effect term) with each participant's actual intelligibility rating.

Effect of hemisphere. To determine the statistical significance of the differences between left and right hemisphere STC, we fit separate representational spaces for each hemisphere in each subject and estimated generalized mixed-effects models for each hemisphere. We used a paired *t*-test on the squared residuals from each model (actual intelligibility – predicted intelligibility) to assess their relative goodness of fit.

Effect of STC subregion. To compare neural-perceptual correlations across STC subregions, in each hemisphere of each participant a control ROI of the same size as the average subregion was created from random nodes within the STC. An MDS space was created across participants, the dissimilarity measured, and the neural-perceptual correlation calculated. This process was repeated 1000 times to create an empirical null distribution for each hemisphere. A *p*-value was calculated by finding the number of simulation runs in which the correlation from a subregion exceeded the simulated values from the same hemisphere.

Effect of study. The main analysis combined data from a study which used words as stimulus material and a study which used sentences as a stimulus material. To compare the studies, separate MDS spaces were constructed for each study. The accuracy of the neural-perceptual models from each study were compared using a resampling procedure. In each simulation, participants were randomly assigned to one study or another (maintaining the different number of participants in each study). Neural-perceptual correlations were calculated for each shuffle and the model accuracies compared (difference of neural-perceptual correlations). The actual words *vs.* sentences correlation difference was compared to the distribution of correlation differences from the shuffled data to assess statistical significance.

2.13. Univariate analysis

For univariate analyses, beta coefficients were averaged across all voxels in an ROI to produce a single value per stimulus condition.

3. Results

Participants rated the intelligibility of speech presented in the MR scanner. As expected, seeing the face of the talker provided a perceptual benefit: for every participant, the intelligibility of noisy audiovisual speech was equal to or greater than the intelligibility of noisy auditory-only speech (76% *vs.* 47% trials rated intelligible, paired $t_{36} = 12.8$, $p = 10^{-14}$).

The fMRI response patterns evoked in superior temporal cortex by the different speech conditions were analyzed using individual-differences multidimensional scaling (MDS). This produced both a group MDS space (showing consistencies in the response patterns across participants) and an MDS space for each participant (showing individual differences).

The group MDS revealed two separate drivers of the fMRI response patterns (Fig. 1A). The first axis corresponded to stimulus modality: along this axis, auditory-only speech conditions clustered in one half of MDS space while conditions that included the face of the talker clustered in the other. The second axis corresponded to speech intelligibility: along this axis, conditions with intelligible speech were found in one half of the MDS space while unintelligible speech conditions were in the other.

The most similar fMRI response patterns were evoked by clear audiovisual speech and noisy (but intelligible) audiovisual speech (consistent with the response drivers identified in MDS space, these conditions had the same similar stimulus modality and intelligibility). The patterns evoked by clear auditory-only speech and noisy (but intelligible) auditory-only speech were also similar, but less so than the clear and noisy audiovisual patterns. We tested the idea that the greater fMRI pattern similarity between clear and noisy audiovisual speech (compared with clear and noisy auditory-only speech) could underlie the perceptual advantage of audiovisual over auditory-only speech. In each individual participant's MDS space, the pattern similarity between clear and noisy speech was measured, separately for audiovisual and auditory-only speech; results for two sample participants are shown in Fig. 1B and C. For 36/37 participants, the audiovisual patterns were more similar than the auditory-only patterns (paired $t_{36} = -14.9$, $p < 10^{-16}$; Fig. 2A).

3.1. Individual differences

Within participants, there was a consistent pattern of greater intelligibility and more similar fMRI response patterns for audiovisual compared with auditory-only speech, prompting an exploration of whether this relationship also held true for individual differences in noisy speech perception. fMRI dissimilarity was plotted against perceptual intelligibility (Fig. 2C). As the similarity decreased, predicted intelligibility also decreased (generalized mixed-effect model, $\chi^2_1 = 293$, $p < 10^{-16}$). The goodness of fit was quantified by correlating the

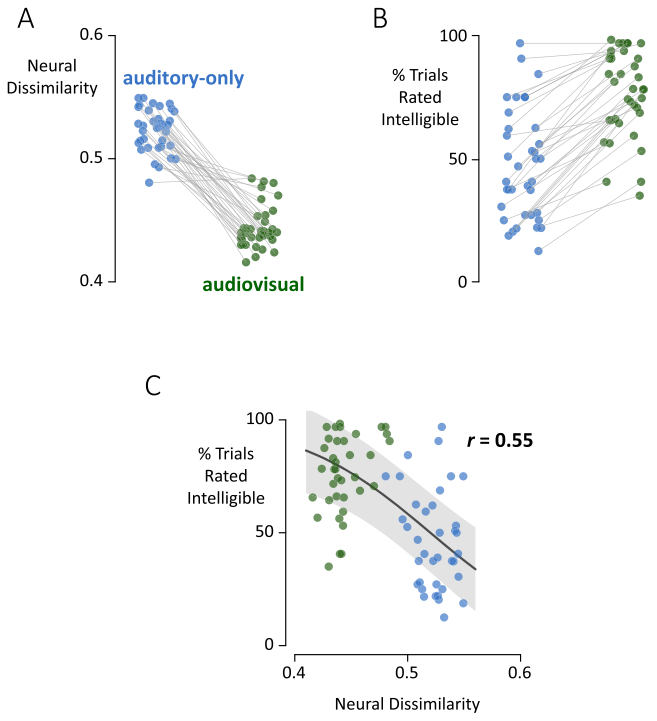

Fig. 2. A. The neural dissimilarity (distance in MDS space; dashed line in *Fig. 1*) was calculated between the fMRI response patterns in bilateral superior temporal cortex to clear speech and noisy (but intelligible) speech, separately for auditory-only speech (blue symbols) and audiovisual speech (green symbols). One pair of connected symbols for each participant.
B. The percent of noisy speech trials rated intelligible for noisy auditory-only speech (blue symbols) and noisy audiovisual speech (green symbols). One pair of connected symbols for each participant.
C. A generalized mixed-effects model was constructed to predict perceptual intelligibility (Fig. 2B) from neural dissimilarity (Fig. 2A). As dissimilarity increased, predicted intelligibility decreased (generalized mixed-effect model, $\chi^2_1 = 293$, $p < 10^{-16}$). The goodness of fit was quantified by correlating the predicted intelligibility for each participant with reported intelligibility ($r = 0.55$). One blue symbol and one green symbol per participant. Black line shows mean fit, shaded error shows SEM.

predicted intelligibility for each participant with reported intelligibility. The predictive accuracy of the model was $r = 0.55$.

3.2. Effect of hemisphere

Our initial analysis showed that fMRI response patterns, measured in an ROI consisting of left and right superior temporal cortex, predicted individual differences in noisy speech perception. Predictive accuracy was similar when MDS spaces and predictive models were created separately for each hemisphere ($r = 0.52$ for left hemisphere *vs.* $r = 0.56$ for right hemisphere, $t_{73} = 0.18$, $p = 0.85$).

3.3. Effect of subregion

Next, we examined whether different subregions of superior temporal cortex differed in their predictive accuracy. Six subregions were examined, three in each hemisphere: superior temporal sulcus (STS); superior temporal gyrus (STG); and the combination of Heschl's gyrus and sulcus, and planum temporale (HG+). Individual-differences MDS was conducted separately for each subregion in each hemisphere and the neural-perceptual correlation calculated (Fig. 3). Within the left hemisphere, the STG had the highest neural-perceptual correlation, significantly different from the left hemisphere null distribution (STG: $r = 0.65$, $p < 0.001$; HG+: $r = 0.53$, $p = 0.06$; STS: $r = 0.51$, $p = 0.29$). Within the right hemisphere, all subregions had correlations that were not significantly different from the right hemisphere null distribution (HG+: $r = 0.65$, $p = 0.28$; STG: $r = 0.62$, $p = 0.41$; STS: $r = 0.56$, $p = 0.63$).

3.4. Effect of study

The main analysis combined data from two studies that used similar experimental designs, except that one used sentences as stimulus material and another that used single words, with no overlapping participants (Fig. 1B and C show single participants from each study). Given the differences in lexical-semantic processing between sentences and words, we examined differences between the studies.

For the perceptual data, a generalized linear-mixed effects model was constructed with participant as random factor; dependent variable of count of trials rated intelligible *vs.* unintelligible; and fixed factors of study (words *vs.* sentences), modality (auditory *vs.* audiovisual), and noise (clear *vs.* noisy). For noisy speech, there was a significant benefit of seeing the face of the talker for words (odds ratio of 4.6, AnV *vs.* An) and for sentences (odds ratio of 4.0) but the two odds ratios were not significantly different ($p = 0.40$; complete model results in *Supplementary_Analysis.html.*)

Separate MDS spaces were constructed for the word and sentence studies, and generalized mixed-effect models of the neural-perceptual relationship were constructed. Both models explained significant variance ($\chi^2_1 = 183$, $p < 10^{-16}$ for words and $\chi^2_1 = 94$, $p < 10^{-16}$ for sentences). The predictive accuracy for the word study was greater but not significantly so ($r = 0.72$ for words *vs.* $r = 0.52$ for sentence, $p = 0.09$).

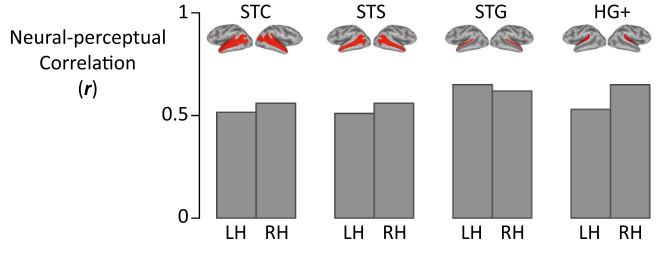

Fig. 3. The neural-perceptual correlation (shown in *Fig. 2C*) was calculated separately for left hemisphere (LH) and right hemisphere (RH) STC and for subregions of the STC consisting of superior temporal sulcus (STS); superior temporal gyrus (STG); and HG+, the combination of planum temporale (PT), transverse superior temporal gyrus (also known as Heschl's Gyrus, HG), and transverse superior temporal sulcus (also known as Heschl's sulcus, HS). Subregions ordered from largest to smallest.

3.5. Other analyses

The response pattern to intelligible noisy speech was measured by *post hoc* sorting the fMRI data, with the unavoidable consequence that different participants had different numbers of intelligible trials. This raises the concern that measurement of the response pattern could be unreliable in participants with fewer trials. To assess this possibility, we removed the 10% of participants ($n = 4$) with the fewest intelligible trials. This changed the neural-perceptual correlation only slightly (from $r = 0.55$ to $r = 0.56$) demonstrating that the correlation was not driven by participants with low trial counts.

3.6. Univariate fMRI results

The preponderance of previous fMRI studies of speech perception have applied univariate fMRI analyses, prompting us to examine our data through a univariate lens. The univariate response was calculated by averaging the response across each ROI, instead of the multivariate approach of correlating patterns of activity.

Speech stimuli evoked a robust hemodynamic response in the superior temporal cortex, peaking 4 to 6 s after stimulus onset (Fig. 4). To quantify the effects of intelligibility on the univariate response, the amplitude of the BOLD signal change was entered into an LME with

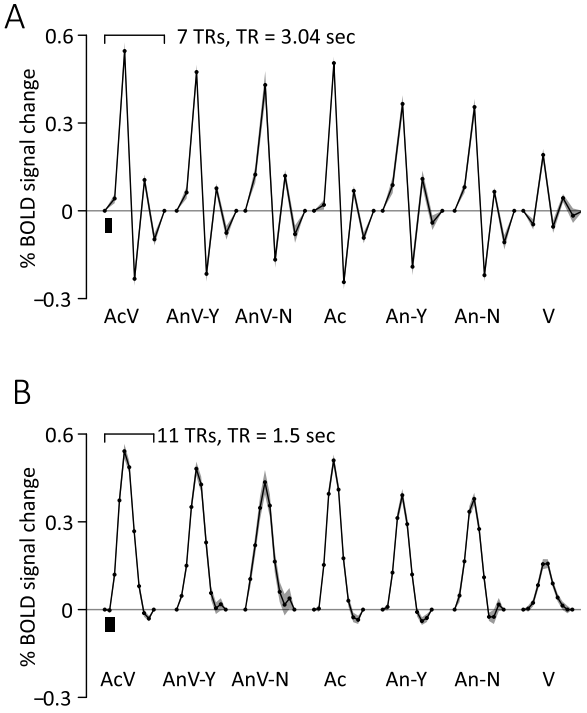

Fig. 4. A. In the word study, single words were presented at the beginning of each trial (black square). The impulse response function across the superior temporal cortex ROI was estimated in each of seven conditions and averaged across participants (lines separated by gaps). A clustered acquisition was used so that the words were always presented in a 2 s silent period between MR acquisitions. The effective TR was 3.04 s.
B. In the sentence study, entire sentences were presented at the beginning of each trial (black square). Continuous acquisition was used, so that the impulse response function was estimated at the TR of 1.5 s.

stimulus modality (noisy auditory *vs.* noisy audiovisual), intelligibility (Y *vs.* N), study (words *vs.* sentences) and STC subregion (STS, STG, HG+ in each hemisphere) as fixed factors and participant as a random factor.

There was a significant main effect of STC subregion ($\chi^2_5 = 368$, $p < 10^{-16}$) driven by a smaller overall response in the STS than in the other ROIs. The main effect of modality ($\chi^2_1 = 20$, $p = 10^{-5}$) was driven by a larger response to noisy audiovisual speech compared to auditory-only speech. The main effect of study ($\chi^2_1 = 10$, $p = 0.001$) was driven by a larger response to sentences than words. There was no main effect of intelligibility ($\chi^2_1 = 0$, $p = 0.92$) and none of the interactions were significant (Fig. 5; complete model output in *Supplementary_Analysis.html*).

4. Discussion

4.1. Summary of main finding

The main finding was of a relationship between the multivariate pattern of fMRI activity in superior temporal cortex (STC) and the intelligibility of noisy speech. The neural dissimilarity between the response patterns to clear and noisy (but intelligible) speech was calculated. When the neural patterns were more similar, noisy speech was more intelligible. This relationship was consistent across several manipulations. Perceptually, noisy audiovisual speech was more intelligible than noisy auditory-only speech, while neurally, fMRI response patterns were more similar for audiovisual than for auditory-only speech. Across individuals, participants with more similar neural patterns for clear and noisy speech were better able to understand noisy speech. Across stimulus material (words and sentences), the same relationship was observed. This consistency suggests that the neural-perceptual relationship was driven by intelligibility, rather than the precise sensory content of the speech stimulus.

4.2. Possible neural mechanisms

Measures of neural dissimilarity made with fMRI have a close correspondence with perception, especially at higher levels of sensory processing (Kriegeskorte and Kievit, 2013). In STC, pattern dissimilarity between perceptually different speech sounds emerges rapidly following speech onset (Chang et al., 2010). The neural similarity between clear speech and noisy (but intelligible) speech in the present study was measured across the average fMRI response pattern evoked by many different words. One possible mechanism for this could be a neural ignition process, in which neural responses to perceived stimuli spread throughout association cortex, while responses to meaningless stimuli remain confined to early sensory areas (Beauchamp et al., 2012; Fisch et al., 2009).

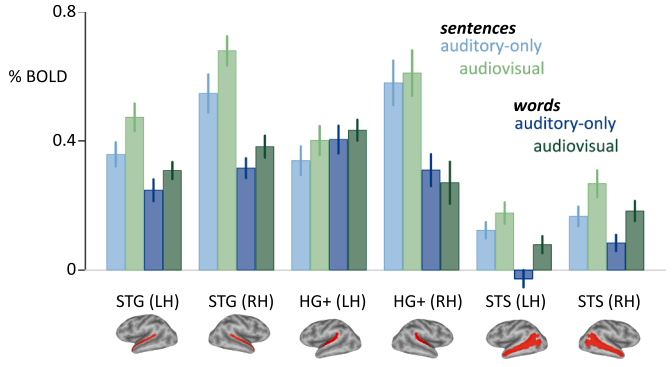

Fig. 5. The mean BOLD signal change was calculated for four conditions: noisy auditory-only sentences (light blue); noisy audiovisual sentences (light green); noisy auditory-only words (dark blue) and noisy audiovisual words (dark green). All nodes within each ROI were averaged, and then averaged across participants (error bar shows SEM across participants).

4.3. Limitations of the present study

Two major limitations of the present study are methodological. As in previous fMRI studies of noisy speech perception, participants responded with a button press. While spoken responses are possible in the MR scanner, they introduce large imaging artifacts (Birn et al., 1998). In addition, BOLD fMRI does not directly measure neural activity, but rather its downstream effects on the cerebral vasculature (Aubert et al., 2007). In future studies, it would be important to replicate these results using other methods (such as intracranial EEG) that do not share these limitations.

4.4. Comparison with previous studies

At least three previous studies have reported individual differences in the neural processing of noisy speech, with conflicting results (Du et al., 2016; Kim et al., 2021; McGettigan et al., 2012b). Our study differed from these studies in several respects, most importantly in *post hoc* sorting of trials and multivariate instead of univariate analysis.

4.5. Post hoc sorting

Previous neuroimaging studies of individual differences in noisy speech perception compared conditions in which intelligibility was low with conditions in which intelligibility was high. McGettigan and colleagues compared conditions with different levels of auditory vocoding (McGettigan et al., 2012b); Du and colleagues compared conditions with different levels of white noise added (Du et al., 2016); and Kim and colleagues added multi-talker babble at different signal-to-noise ratios (Kim et al., 2021). However, this design introduces a confound: in addition to the intelligibility differences between conditions, there are also physical (low-level sensory) differences in the stimulus. An alternative approach was first described in the memory literature (Brewer et al., 1998; Wagner et al., 1998). Event-related fMRI was used to measure the response to individual items at encoding. A subsequent test determined which items were accurately remembered, and the response to correct *vs.* incorrectly encoded items (that were otherwise very similar) were *post hoc* sorted into separate bins. This experimental design has been successfully applied in studies of noisy speech perception (Bishop and Miller, 2009; Rennig and Beauchamp, 2022) and was used in the present study. Noisy speech stimuli were presented and participants reported their perception. For each participant, the BOLD fMRI responses to the noisy stimuli that were intelligible or not were sorted and analyzed separately.

4.6. Multivariate analysis

Using univariate analyses in which one brain location was analyzed at a time, previous neuroimaging studies of individual differences in noisy speech perception found conflicting results (Du et al., 2016; Kim et al., 2021; McGettigan et al., 2012b). One reason may be that univariate measures have limited power; in a direct comparison within individuals, Rennig and Beauchamp (2022) found that neural differences between intelligible and unintelligible speech were much larger for multivariate ($\chi^2 = 102$) than univariate ($\chi^2 = 11$) analyses.

A key property of the multivariate analysis is that it compares the response to clear speech with the response to noisy (but intelligible) speech. Conceptually, the pattern of response to clear speech can be thought of as a reference, corresponding to a neural template or prototype; the more closely the response pattern evoked by noisy speech approximates the clear speech pattern, the more intelligible the noisy speech will be.

Previous multivariate studies contrasted clear speech (always intelligible) with spectrally rotated speech (never intelligible); or noise-vocoded speech (sometimes intelligible) with spectrally-rotated noise vocoded speech (never intelligible) (Abrams et al., 2013; Evans et al.,

2014; Evans and Davis, 2015; McGettigan et al., 2012a; Okada et al., 2010). Spectral rotation is very different from the noise encountered in natural listening conditions, so that differences in brain responses between clear speech and spectrally rotated speech would not necessarily be expected to correlate with the ability to understand speech-in-noise across participants. In contrast, neural responses to noisy (but intelligible) speech may be more likely (with the proper measurement and analysis techniques) to be related to perceptual differences in the ability to perceive noisy speech.

The sentence and word studies were each conducted at a single, fixed noise level (−8 dB SPL for words and −16 dB SPL for sentences). It would be interesting to test the neural intelligibility index on fMRI datasets containing a variety of noise levels (Davis and Johnsrude, 2003; Du et al., 2014; Golestani et al., 2013). The prediction is that at increasing noise levels, the pattern difference between noisy speech and clear speech would become greater and greater, corresponding to the decrease in perceptual intelligibility. The relationship between physical signal to noise ratio and perceptual intelligibility follows a typically S-shaped psychometric function (Ross et al., 2021) and it would be interesting to ascertain if the multivariate metric took a similar form with increasing noise levels.

4.7. Region of interest

There are a number of statistical pitfalls in identifying brain-behavior relationships using MRI (Marek et al., 2022; Vul et al., 2009). The most serious problem involves the dimensional mismatch inherent in most MRI studies between the number of brain measurements (up to a million voxels) and the number of participants (typically ten to twenty). Across a multitude of brain voxels, some will show a high value for the brain-behavior correlation. Selecting for significant correlations results in false positives and effect-size inflation (Palmer and Pe'er, 2017; Vasishth et al., 2018). To avoid this "winner's curse", in the current study the neural-perceptual relationship *across* individuals was examined within STC and component subregions, the brain area identified as the most consistent neural correlate of speech intelligibility *within* individuals across many previous neuroimaging studies (Abrams et al., 2013; Davis and Johnsrude, 2007; DeWitt and Rauschecker, 2012; Eckert et al., 2016; Evans, 2017; Evans et al., 2014; Hakonen et al., 2017; Holmes et al., 2021; Johnsrude et al., 2002; Wong et al., 2008).

4.8. Individual differences in speech perception

For any measurable trait, physical or psychological, there is enormous variability across individuals, reflecting an interaction between the environment and the genetic and epigenetic makeup of each individual. This interaction is particularly pronounced for language: while humans are genetically endowed with the ability the acquire language, the specific language(s) learned by a person is determined by their environment (Wong et al., 2022). Individual differences are pervasive across the language system and are related to environmental variables (such as the quantity and quality of the language that an individual is exposed to) and to individual differences in other cognitive functions, such as working memory, executive function, and statistical learning, reviewed in (Kidd et al., 2018; Shinn-Cunningham, 2017).

4.9. Anatomical subregions

Superior temporal cortex (STC) was used as the region of interest in our study because of the consensus in the neuroimaging literature that it is a key site for within-subject intelligibility effects, reviewed in Abrams et al. (2013), Davis and Johnsrude (2007), DeWitt and Rauschecker (2012), Evans (2017), Evans et al. (2014). Invasive recordings in patients with epilepsy have found that larger responses to noisy speech in STC in patients with better task performance (Nourski et al., 2019); that STC restores missing acoustic content from the incoming speech signal (Leonard et al., 2016); and that STC responses to visual speech correspond to its perceptual benefit for noisy auditory speech (Karas et al., 2019). Magnetoencephalographic recordings track the maintenance of online speech representations in STC for the resolution of ambiguous speech tokens (Gwilliams et al., 2018) and electrical stimulation of STC results in improved speech-in-noise perception (Patel et al., 2022).

In the multivariate analysis, only modest differences between STC subregions were observed in our study. It may be that with the improved temporal resolution of iEEG, MEG or EEG, it would be possible to identify the early emergence of neural differences between intelligible speech in one anatomical subregion, which then spread to other subregions. Minimal differences in neural-perceptual correlation were observed between hemisphere, consistent with a recent neuroimaging study (Wehbe et al., 2021).

The neural intelligibility metric in the present study was the Euclidean distance between clear and noisy speech response patterns. This approach is relatively simple, with no free parameters: each location in STC contributes equally to the pattern difference. A more complex approach would be to differentially weight different locations in STC using a machine learning approach (Kaniuth and Hebart, 2022). This would likely increase neural-perceptual correlations, at the risk of overfitting a dataset with many more brain locations (neural measures) than participants (perceptual measures).

With the appropriate statistical precautions, the neural dissimilarity measure could be adapted for use in a brain-behavior searchlight analysis (Emmerling et al., 2016). This would permit identification of all brain areas with significant neural-perceptual correlations. In cortex, frontal, parietal, cingular-opercular, and insular regions have all been implicated in speech perception under difficult conditions and would be important to investigate for their contributions to individual differences (Alain et al., 2018; Chevillet et al., 2013; Du et al., 2014; Vaden et al., 2015, 2013). Subcortically, the medial geniculate nucleus has been identified as a site for group differences in speech perception (Mihai et al., 2021; Schelinski et al., 2022). Stimulus-response mapping on whole-head scalp EEG data shows that there are widespread cortical signals related to perception of noisy auditory and audiovisual speech (Di Liberto et al., 2018; O'Sullivan et al., 2021).

Univariate and multivariate analyses provide complementary windows into BOLD datasets (Davis et al., 2014). In the univariate analysis, there was a significant main effect of STC subregion, driven by a low mean response in the STS. In large ROIs such as the STS, some voxels are activated (response above fixation baseline) and others that show little activity or are deactivated (response below fixation baseline). Averaging them in a univariate analysis results in low mean signal change. In contrast, in the multivariate analysis, both activations and deactivations at individual brain locations convey information.

The univariate analysis showed a significant main effect for stimulus material, driven by a larger response in trials containing entire sentences than trials containing single words. This is to be expected, as a sentence containing many words (as well as additional linguistic and semantic information) should drive more neural activity than an isolated, out-of-context word. In addition to the stimulus material difference, the sentence and word studies were each conducted at a single, fixed noise level that differed between studies (−8 SPL for words and −16 dB SPL for sentences); and sparse sampling MR acquisition was used in the word study but not the sentence study.

Finally, there was a main effect for the modality contrast, driven by a larger response for audiovisual than auditory trials. This replicates many previous findings of a larger STC response to multisensory audiovisual stimuli compared with unisensory stimuli, such as (Beauchamp et al., 2004; van Atteveldt et al., 2004).

4.10. Benefits of visual speech

The perceptual results of the present study replicate one of the most reliable findings in the language literature, that visual information from

the talker's face increases the intelligibility of noisy speech (Peelle and Sommers, 2015b; Sumby and Pollack, 1954). Multisensory integration is beneficial because the external sources of noise in different sensory modalities are largely independent: the presence of loud background music in a restaurant does not interfere with seeing our tablemate's face (Stein and Meredith, 1993). The face of the talker provides multiple sources of information about speech. First, it provides temporal information about the occurrence of speech, because of the high correlation between the visual speech envelope and the auditory speech envelope (Ghazanfar and Takahashi, 2014). Seeing the talker's mouth open is a reliable cue that auditory speech is expected soon. Secondly, it provides information about the content of speech. While there is not a one-to-one correspondence between particular auditory and visual speech tokens, the facial configuration of the talker is highly informative (Cappelletta and Harte, 2012). Most auditory phonemes are incompatible with any given mouth shape adopted by the talker. Since the visual mouth shape is adopted before vocalization begins, visual speech can provide a head start on processing the content of forthcoming auditory speech (Karas et al., 2019). The powerful influence of the face of the talker on auditory speech perception is illustrated by the McGurk effect, a well-known speech illusion (McGurk and MacDonald, 1976).

4.11. Contributors to the neural dissimilarity measure

Activity patterns in superior temporal cortex represent a readout of many factors that contribute to speech perception. Peripheral differences could contribute to individual differences in brain activity. In the present study, the healthy adulty participants reported normal (or corrected-to-normal) vision and normal hearing, but no objective tests of hearing were reported. Similarly large individual differences in noisy audiovisual and auditory-only speech perception have also been observed in participants with normal hearing thresholds, as measured with a standard audiometric exam (Kim et al., 2021; Van Engen et al., 2017). However, college students at high risk for hearing damage (due to exposure to loud sounds without hearing protection) were found to have impaired cochlear function and worse speech-in-noise understanding than students at low risk, even though both groups had normal hearing thresholds when tested at standard audiometric frequencies (Liberman et al., 2016). These peripheral differences could reduce the fidelity of inputs into superior temporal cortex, changing the neural intelligibility index.

Kim et al. (2021) used EEG to measure two different processes contributing to individual differences in speech in noise, a lower-level process that separates the speech signal from the noise, and a higher-level process that converts the speech into meaningful tokens. Both processes would be expected to alter patterns of activity in superior temporal cortex. Individual differences in audiovisual speech perception are linked to face viewing behavior. Human naturally fixate the mouth of the talker when noise is added to auditory speech. However, participants who fixate the mouth of the talker even when it is not required, during presentation of clear speech, receive more benefit from seeing the face during presentation of noisy speech, possibly due to their greater experience and expertise with face-voice correspondence (Rennig et al., 2020; Wegner-Clemens et al., 2020). Through this route, the complex networks of brain areas linked to eye movement control and face perception could indirectly influence in the intelligibility index computed from superior temporal cortex.

4.12. Applications

The results of the present study suggest that if the pattern of brain response evoked in the superior temporal cortex by noise speech could somehow be made more similar to the pattern evoked by clear speech, the result could be enhanced intelligibility. One obvious possibility for this would be real-time neurofeedback. It has been shown that providing participants with feedback about their brain response, especially positive feedback when the brain pattern becomes more similar to a desired pattern, can learn to modulate their own brain responses, resulting in enhanced perception (Ramot et al., 2017; Watanabe et al., 2017). To improve noisy speech perception, the response pattern to clear speech would be measured; then noisy speech would be presented, and participants would receive feedback about the similarity of the evoked fMRI pattern in superior temporal cortex with the reference clear speech pattern.

More invasively, there is great interest in implanted brain-computer interfaces as remedies for communication deficits and to compensate for lost peripheral sensory function (Beauchamp et al., 2020; Moses et al., 2021; Vansteensel and Jarosiewicz, 2020). Electrical stimulation of intracranial electrodes implanted in superior temporal cortex of a single patient resulted in improved speech-in-noise perception (Patel et al., 2022). A neural prosthetic implanted in superior temporal cortex could help normalize the pattern of activity to make it easier for patients to understand speech. In the event that the evoked response to clear speech could not be measured in an individual, a template "clear speech" pattern derived from the group average of many participants could be used (Kragel et al., 2018).

Data code availability

To ensure reliability and reproducibility, the analysis code and data used for all analyses and figures presented in the manuscript may be found in the attached files, *Supplementary_Analysis.html* and *SummaryData.xlsx*. The files are organized by manuscript section and figure number.

CRediT authorship contribution statement

Yue Zhang: Conceptualization, Investigation. **Johannes Rennig:** Conceptualization, Investigation. **John F Magnotti:** Conceptualization, Investigation. **Michael S Beauchamp:** Conceptualization, Investigation.

Declaration of Competing Interest

None.

Data availability

The data is attached as a supplementary material file.

Acknowledgments

We thank Lacey Delay and Krista Runge for CAMRI support. The research was funded by NIH NS065395 and NS113339. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.neuroimage.2023.120271.

References

Abrams, D.A., Ryali, S., Chen, T., Balaban, E., Levitin, D.J., Menon, V., 2013. Multivariate activation and connectivity patterns discriminate speech intelligibility in Wernicke's, Broca's, and Geschwind's areas. Cereb. Cortex N. Y. N 23, 1703–1714. https://doi.org/10.1093/cercor/bhs165, 1991.

Alain, C., Du, Y., Bernstein, L.J., Barten, T., Banai, K., 2018. Listening under difficult conditions: an activation likelihood estimation meta-analysis. Hum. Brain Mapp. 39, 2695–2709. https://doi.org/10.1002/hbm.24031.

Argall, B.D., Saad, Z.S., Beauchamp, M.S., 2006. Simplified intersubject averaging on the cortical surface using SUMA. Hum Brain Mapp 27, 14–27. https://doi.org/10.1002/hbm.20158.

Aubert, A., Pellerin, L., Magistretti, P.J., Costalat, R., 2007. A coherent neurobiological framework for functional neuroimaging provided by a model integrating compartmentalized energy metabolism. Proc. Natl. Acad. Sci. U A 104, 4188–4193.
Bates, D., Mächler, M., Bolker, B., Walker, S., 2015. Fitting linear mixed-effects models using lme4. J. Stat. Softw. 1, 1–48.
Beauchamp, M.S., Lee, K.E., Argall, B.D., Martin, A., 2004. Integration of auditory and visual information about objects in superior temporal sulcus. Neuron 41, 809–823.
Beauchamp, M.S., Oswalt, D., Sun, P., Foster, B.L., Magnotti, J.F., Niketeghad, S., Pouratian, N., Bosking, W.H., Yoshor, D., 2020. Dynamic stimulation of visual cortex produces form vision in sighted and blind humans. Cell 181, 774–783. https://doi.org/10.1016/j.cell.2020.04.033 e5.
Beauchamp, M.S., Sun, P., Baum, S.H., Tolias, A.S., Yoshor, D., 2012. Electrocorticography links human temporoparietal junction to visual perception. Nat. Neurosci. 15, 957–959. https://doi.org/10.1038/nn.3131.
Bernstein, L.E., Liebenthal, E., 2014. Neural pathways for visual speech perception. Front. Neurosci. 8, 386. https://doi.org/10.3389/fnins.2014.00386.
Birn, R.M., Bandettini, P.A., Cox, R.W., Jesmanowicz, A., Shaker, R., 1998. Magnetic field changes in the human brain due to swallowing or speaking. Magn. Reson. Med. 40, 55–60.
Bishop, C.W., Miller, L.M., 2009. A multisensory cortical network for understanding speech in noise. J. Cogn. Neurosci. 21, 1790–1805. https://doi.org/10.1162/jocn.2009.21118.
Brainard, D.H., 1997. The psychophysics toolbox. Spat. Vis. 10, 433–436. https://doi.org/10.1163/156856897X00357.
Brewer, J.B., Zhao, Z., Desmond, J.E., Glover, G.H., Gabrieli, J.D., 1998. Making memories: brain activity that predicts how well visual experience will be remembered. Science 281, 1185–1187. https://doi.org/10.1126/science.281.5380.1185.
Cappelletta, L., Harte, N., 2012. Phoneme-to-viseme mapping for visual speech recognition. Presented at the. In: Proceedings of the 1st International Conference on Pattern Recognition Applications and Methods. SciTePress, pp. 322–329. https://doi.org/10.5220/0003731903220329.
Carroll, J.D., Chang, J.J., 1970. Analysis of individual differences in multidimensional scaling via an n-way generalization of "Eckart-Young" decomposition. Psychometrika 35, 283–319. https://doi.org/10.1007/BF02310791.
Chang, E.F., Rieger, J.W., Johnson, K., Berger, M.S., Barbaro, N.M., Knight, R.T., 2010. Categorical speech representation in human superior temporal gyrus. Nat. Neurosci. 13, 1428–1432. https://doi.org/10.1038/nn.2641.
Chevillet, M.A., Jiang, X., Rauschecker, J.P., Riesenhuber, M., 2013. Automatic phoneme category selectivity in the dorsal auditory stream. J. Neurosci. Off. J. Soc. Neurosci. 33, 5208–5215. https://doi.org/10.1523/JNEUROSCI.1870-12.2013.
Cox, R.W., 1996. AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. Comput. Biomed. Res. Int. J. 29, 162–173. https://doi.org/10.1006/cbmr.1996.0014.
Davis, M.H., Johnsrude, I.S., 2007. Hearing speech sounds: top-down influences on the interface between audition and speech perception. Hear. Res. 229, 132–147. https://doi.org/10.1016/j.heares.2007.01.014.
Dale, A.M., Fischl, B., Sereno, M.I., 1999. Cortical surface-based analysis. I. Segmentation and surface reconstruction. Neuroimage 9, 179–194.
Davis, M.H., Johnsrude, I.S., 2003. Hierarchical processing in spoken language comprehension. J. Neurosci. 23, 3423–3431.
Davis, T., LaRocque, K.F., Mumford, J.A., Norman, K.A., Wagner, A.D., Poldrack, R.A., 2014. What do differences between multi-voxel and univariate analysis mean? How subject-, voxel-, and trial-level variance impact fMRI analysis. Neuroimage 97, 271–283. https://doi.org/10.1016/j.neuroimage.2014.04.037.
Destrieux, C., Fischl, B., Dale, A., Halgren, E., 2010. Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature. Neuroimage 53, 1–15. https://doi.org/10.1016/j.neuroimage.2010.06.010.
DeWitt, I., Rauschecker, J.P., 2012. Phoneme and word recognition in the auditory ventral stream. Proc. Natl. Acad. Sci. U. S. A. 109, E505–E514. https://doi.org/10.1073/pnas.1113427109.
Di Liberto, G.M., Crosse, M.J., Lalor, E.C., 2018. Cortical measures of phoneme-level speech encoding correlate with the perceived clarity of natural speech. eNeuro 5, 2018. https://doi.org/10.1523/ENEURO.0084-18.2018. ENEURO.0084-18.
Du, Y., Buchsbaum, B.R., Grady, C.L., Alain, C., 2016. Increased activity in frontal motor cortex compensates impaired speech perception in older adults. Nat. Commun. 7, 12241. https://doi.org/10.1038/ncomms12241.
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, 7126–7131. https://doi.org/10.1073/pnas.1318738111.
Eckert, M.A., Teubner-Rhodes, S., Vaden, K.I.J., 2016. Is listening in noise worth it? The neurobiology of speech recognition in challenging listening conditions. Ear Hear. 37, 101S. https://doi.org/10.1097/AUD.0000000000000300.
Edmister, W.B., Talavage, T.M., Ledden, P.J., Weisskoff, R.M., 1999. Improved auditory cortex imaging using clustered volume acquisitions. Hum. Brain Mapp. 7, 89–97.
Emmerling, T.C., Zimmermann, J., Sorger, B., Frost, M.A., Goebel, R., 2016. Decoding the direction of imagined visual motion using 7T ultra-high field fMRI. Neuroimage 125, 61–73. https://doi.org/10.1016/j.neuroimage.2015.10.022.
Evans, S., 2017. What has replication ever done for us? Insights from neuroimaging of speech perception. Front. Hum. Neurosci. 11, 1–5. https://doi.org/10.3389/fnhum.2017.00041.
Evans, S., Davis, M.H., 2015. Hierarchical organization of auditory and motor representations in speech perception: evidence from searchlight similarity analysis. Cereb. Cortex 25, 4772–4788. https://doi.org/10.1093/cercor/bhv136.
Evans, S., Kyong, J.S., Rosen, S., Golestani, N., Warren, J.E., McGettigan, C., Mourão-Miranda, J., Wise, R.J.S., Scott, S.K., 2014. The pathways for intelligible speech: multivariate and univariate perspectives. Cereb. Cortex N. Y. N 24, 2350–2361. https://doi.org/10.1093/cercor/bht083, 1991.
Fisch, L., Privman, E., Ramot, M., Harel, M., Nir, Y., Kipervasser, S., Andelman, F., Neufeld, M.Y., Kramer, U., Fried, I., Malach, R., 2009. Neural "Ignition": enhanced activation linked to perceptual awareness in human ventral stream visual cortex. Neuron 64, 562–574. https://doi.org/10.1016/j.neuron.2009.11.001.
French, N.R., Steinberg, J.C., 1947. Factors governing the intelligibility of speech sounds. J. Acoust. Soc. Am. 19, 90–119. https://doi.org/10.1121/1.1916407.
Ghazanfar, A.A., Takahashi, D.Y., 2014. Facial expressions and the evolution of the speech rhythm. J. Cogn. Neurosci. 26, 1196–1207. https://doi.org/10.1162/jocn_a_00575.
Golestani, N., Hervais-Adelman, A., Obleser, J., Scott, S.K., 2013. Semantic *versus* perceptual interactions in neural processing of speech-in-noise. Neuroimage 79, 52–61. https://doi.org/10.1016/j.neuroimage.2013.04.049.
Gwilliams, L., Linzen, T., Poeppel, D., Marantz, A., 2018. In spoken word recognition, the future predicts the past. J. Neurosci. Off. J. Soc. Neurosci. 38, 7585–7599. https://doi.org/10.1523/JNEUROSCI.0065-18.2018.
Hakonen, M., May, P.J.C., Jääskeläinen, I.P., Jokinen, E., Sams, M., Tiitinen, H., 2017. Predictive processing increases intelligibility of acoustically distorted speech: behavioral and neural correlates. Brain Behav. 7, e00789. https://doi.org/10.1002/brb3.789.
Hall, D.A., Haggard, M.P., Akeroyd, M.A., Palmer, A.R., Summerfield, A.Q., Elliott, M.R., Gurney, E.M., Bowtell, R.W., 1999. Sparse" temporal sampling in auditory fMRI. Hum. Brain Mapp. 7, 213–223.
Haxby, J.V., Gobbini, M.I., Furey, M.L., Ishai, A., Schouten, J.L., Pietrini, P., 2001. Distributed and overlapping representations of faces and objects in ventral temporal cortex. Science 293, 2425–2430. https://doi.org/10.1126/science.1063736.
Haynes, J.D., Rees, G., 2006. Decoding mental states from brain activity in humans. Nat. Rev. Neurosci. 7, 523–534.
Hickok, G., Poeppel, D., 2004. Dorsal and ventral streams: a framework for understanding aspects of the functional anatomy of language. Cognition 92, 67–99. https://doi.org/10.1016/j.cognition.2003.10.011.
Holmes, E., Zeidman, P., Friston, K.J., Griffiths, T.D., 2021. Difficulties with speech-in-noise perception related to fundamental grouping processes in auditory cortex. Cereb. Cortex 31, 1582–1596. https://doi.org/10.1093/cercor/bhaa311.
Johnsrude, I.S., Giraud, A.L., Frackowiak, R.S.J., 2002. Functional imaging of the auditory system: the use of positron emission tomography. Audiol. Neurootol. 7, 251–276. https://doi.org/10.1159/000064446.
Kaniuth, P., Hebart, M.N., 2022. Feature-reweighted representational similarity analysis: a method for improving the fit between computational models, brains, and behavior. Neuroimage 257, 119294. https://doi.org/10.1016/j.neuroimage.2022.119294.
Karas, P.J., Magnotti, J.F., Metzger, B.A., Zhu, L.L., Smith, K.B., Yoshor, D., Beauchamp, M.S., 2019. The visual speech head start improves perception and reduces superior temporal cortex responses to auditory speech. Elife 8, 1–19. https://doi.org/10.7554/eLife.48116.
Kidd, E., Donnelly, S., Christiansen, M.H., 2018. Individual differences in language acquisition and processing. Trends Cogn. Sci. 22, 154–169. https://doi.org/10.1016/j.tics.2017.11.006.
Kim, S., Schwalje, A.T., Liu, A.S., Gander, P.E., McMurray, B., Griffiths, T.D., Choi, I., 2021. Pre- and post-target cortical processes predict speech-in-noise performance. Neuroimage 228, 117699. https://doi.org/10.1016/j.neuroimage.2020.117699.
Kragel, P.A., Koban, L., Barrett, L.F., Wager, T.D., 2018. Representation, pattern information, and brain signatures: from neurons to neuroimaging. Neuron 99, 257–273. https://doi.org/10.1016/j.neuron.2018.06.009.
Kriegeskorte, N., Goebel, R., Bandettini, P., 2006. Information-based functional brain mapping. Proc. Natl. Acad. Sci. U A 103, 3863–3868.
Kriegeskorte, N., Kievit, R.A., 2013. Representational geometry: integrating cognition, computation, and the brain. Trends Cogn. Sci. 17, 401–412. https://doi.org/10.1016/j.tics.2013.06.007.
Leeuw, J.de, Mair, P., 2009. Multidimensional Scaling Using Majorization: SMACOF in R. J. Stat. Softw. 31, 1–30. https://doi.org/10.18637/jss.v031.i03.
Leonard, M.K., Baud, M.O., Sjerps, M.J., Chang, E.F., 2016. Perceptual restoration of masked speech in human cortex. Nat. Commun. 7, 13619. https://doi.org/10.1038/ncomms13619.
Liberman, M.C., Epstein, M.J., Cleveland, S.S., Wang, H., Maison, S.F., 2016. Toward a differential diagnosis of hidden hearing loss in humans. PLoS One 11, e0162726. https://doi.org/10.1371/journal.pone.0162726.
Marek, S., Tervo-Clemmens, B., Calabro, F.J., Montez, D.F., Kay, B.P., Hatoum, A.S., Donohue, M.R., Foran, W., Miller, R.L., Hendrickson, T.J., Malone, S.M., Kandala, S., Feczko, E., Miranda-Dominguez, O., Graham, A.M., Earl, E.A., Perrone, A.J., Cordova, M., Doyle, O., Moore, L.A., Conan, G.M., Uriarte, J., Snider, K., Lynch, B.J., Wilgenbusch, J.C., Pengo, T., Tam, A., Chen, J., Newbold, D.J., Zheng, A., Seider, N. A., Van, A.N., Metoki, A., Chauvin, R.J., Laumann, T.O., Greene, D.J., Petersen, S.E., Garavan, H., Thompson, W.K., Nichols, T.E., Yeo, B.T.T., Barch, D.M., Luna, B., Fair, D.A., Dosenbach, N.U.F., 2022. Reproducible brain-wide association studies require thousands of individuals. Nature 603, 654–660. https://doi.org/10.1038/s41586-022-04492-9.
McGettigan, C., Evans, S., Rosen, S., Agnew, Z.K., Shah, P., Scott, S.K., 2012a. An application of univariate and multivariate approaches in FMRI to quantifying the hemispheric lateralization of acoustic and linguistic processes. J. Cogn. Neurosci. 24, 636–652. https://doi.org/10.1162/jocn_a_00161.
McGettigan, C., Faulkner, A., Altarelli, I., Obleser, J., Baverstock, H., Scott, S.K., 2012b. Speech comprehension aided by multiple modalities: behavioural and neural interactions. Neuropsychologia 50, 762–776. https://doi.org/10.1016/j.neuropsychologia.2012.01.010.
McGurk, H., MacDonald, J., 1976. Hearing lips and seeing voices. Nature 264, 746–748.

Mihai, P.G., Tschentscher, N., von Kriegstein, K., 2021. Modulation of the primary auditory thalamus when recognizing speech with background noise. J. Neurosci. Off. J. Soc. Neurosci. 41, 7136–7147. https://doi.org/10.1523/JNEUROSCI.2902-20.2021.

Moses, D.A., Metzger, S.L., Liu, J.R., Anumanchipalli, G.K., Makin, J.G., Sun, P.F., Chartier, J., Dougherty, M.E., Liu, P.M., Abrams, G.M., Tu-Chan, A., Ganguly, K., Chang, E.F., 2021. Neuroprosthesis for decoding speech in a paralyzed person with anarthria. N. Engl. J. Med. 385, 217–227. https://doi.org/10.1056/NEJMoa2027540.

Norman, K.A., Polyn, S.M., Detre, G.J., Haxby, J.V., 2006. Beyond mind-reading: multi-voxel pattern analysis of fMRI data. Trends Cogn. Sci. 10, 424–430.

Nourski, K.V., Steinschneider, M., Rhone, A.E., Kovach, C.K., Kawasaki, H., Howard, M. A., 2019. Differential responses to spectrally degraded speech within human auditory cortex: an intracranial electrophysiology study. Hear. Res. 371, 53–65. https://doi.org/10.1016/j.heares.2018.11.009.

Okada, K., Rong, F., Venezia, J., Matchin, W., Hsieh, I.H., Saberi, K., Serences, J.T., Hickok, G., 2010. Hierarchical organization of human auditory cortex: evidence from acoustic invariance in the response to intelligible speech. Cereb. Cortex 20, 2486–2495. https://doi.org/10.1093/cercor/bhp318.

O'Sullivan, A.E., Crosse, M.J., Liberto, G.M.D., de Cheveigné, A., Lalor, E.C., 2021. Neurophysiological indices of audiovisual speech processing reveal a hierarchy of multisensory integration effects. J. Neurosci. Off. J. Soc. Neurosci. 41, 4991–5003. https://doi.org/10.1523/JNEUROSCI.0906-20.2021.

Ozker, M., Yoshor, D., Beauchamp, M.S., 2018. Converging evidence from electrocorticography and BOLD fMRI for a sharp functional boundary in superior temporal gyrus related to multisensory speech processing. Front. Hum. Neurosci. 12, 141. https://doi.org/10.3389/fnhum.2018.00141.

Palmer, C., Pe'er, I., 2017. Statistical correction of the Winner's Curse explains replication variability in quantitative trait genome-wide association studies. PLos Genet. 13, e1006916 https://doi.org/10.1371/journal.pgen.1006916.

Patel, P., Khalijhinejad, B., Herrero, J.L., Bickel, S., Mehta, A.D., Mesgarani, N., 2022. Improved speech hearing in noise with invasive electrical brain stimulation. J. Neurosci. 42, 3648–3658. https://doi.org/10.1523/JNEUROSCI.1468-21.2022.

Peelle, J.E., Sommers, M.S., 2015a. Prediction and constraint in audiovisual speech perception. Cortex 68, 169–181. https://doi.org/10.1016/j.cortex.2015.03.006.

Peelle, J.E., Sommers, M.S., 2015b. Prediction and constraint in audiovisual speech perception. Cortex J. Devoted Study Nerv. Syst. Behav. 68, 169–181. https://doi.org/10.1016/j.cortex.2015.03.006.

Perrachione, T.K., Ghosh, S.S., 2013. Optimized design and analysis of sparse-sampling fMRI experiments. Front. Neurosci. 7, 55. https://doi.org/10.3389/fnins.2013.00055.

Ramot, M., Kimmich, S., Gonzalez-Castillo, J., Roopchansingh, V., Popal, H., White, E., Gotts, S.J., Martin, A., 2017. Direct modulation of aberrant brain network connectivity through real-time NeuroFeedback. eLife 6, e28974. https://doi.org/10.7554/eLife.28974.

Rennig, J., Beauchamp, M.S., 2022. Intelligibility of audiovisual sentences drives multivoxel response patterns in human superior temporal cortex. Neuroimage 247, 118796. https://doi.org/10.1016/j.neuroimage.2021.118796.

Rennig, J., Wegner-Clemens, K., Beauchamp, M.S., 2020. Face viewing behavior predicts multisensory gain during speech perception. Psychon. Bull. Rev. 27, 70–77. https://doi.org/10.3758/s13423-019-01665-y.

Ross, B., Dobri, S., Schumann, A., 2021. Psychometric function for speech-in-noise tests accounts for word-recognition deficits in older listeners. J. Acoust. Soc. Am. 149, 2337. https://doi.org/10.1121/10.0003956.

Schelinski, S., Tabas, A., von Kriegstein, K., 2022. Altered processing of communication signals in the subcortical auditory sensory pathway in autism. Hum. Brain Mapp. 43, 1955–1972. https://doi.org/10.1002/hbm.25766.

Shinn-Cunningham, B., 2017. Cortical and sensory causes of individual differences in selective attention ability among listeners with normal hearing thresholds. J. Speech Lang. Hear. Res. 60, 2976–2988. https://doi.org/10.1044/2017_JSLHR-H-17-0080.

Stein, B.E., Meredith, M.A., 1993. The Merging of the Senses. MIT Press.

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

Vaden, K.I., Kuchinsky, S.E., Ahlstrom, J.B., Dubno, J.R., Eckert, M.A., 2015. Cortical activity predicts which older adults recognize speech in noise and when. J. Neurosci. 35, 3929–3937. https://doi.org/10.1523/JNEUROSCI.2908-14.2015.

Vaden, K.I., Kuchinsky, S.E., Cute, S.L., Ahlstrom, J.B., Dubno, J.R., Eckert, M.A., 2013. The Cingulo-Opercular network provides word-recognition benefit. J. Neurosci. 33, 18979–18986. https://doi.org/10.1523/JNEUROSCI.1417-13.2013.

van Atteveldt, N., Formisano, E., Goebel, R., Blomert, L., 2004. Integration of letters and speech sounds in the human brain. Neuron 43, 271–282. https://doi.org/10.1016/j.neuron.2004.06.025.

Van Engen, K.J., Xie, Z., Chandrasekaran, B., 2017. Audiovisual sentence recognition not predicted by susceptibility to the McGurk effect. Atten. Percept. Psychophys. 79, 396–403. https://doi.org/10.3758/s13414-016-1238-9.

Vansteensel, M.J., Jarosiewicz, B., 2020. Brain-computer interfaces for communication. Handb. Clin. Neurol. 168, 67–85. https://doi.org/10.1016/B978-0-444-63934-9.00007-X.

Vasishth, S., Mertzen, D., Jäger, L.A., Gelman, A., 2018. The statistical significance filter leads to overoptimistic expectations of replicability. J. Mem. Lang. 103, 151–175. https://doi.org/10.1016/j.jml.2018.07.004.

Vul, E., Harris, C., Winkielman, P., Pashler, H., 2009. Puzzlingly high correlations in fMRI studies of emotion, personality and social cognition. Perspect. Psychol. Sci. 4, 274–290.

Wagner, A.D., Schacter, D.L., Rotte, M., Koutstaal, W., Maril, A., Dale, A.M., Rosen, B.R., Buckner, R.L., 1998. Building memories: remembering and forgetting of verbal experiences as predicted by brain activity. Science 281, 1188–1191.

Watanabe, T., Sasaki, Y., Shibata, K., Kawato, M., 2017. Advances in fMRI real-time neurofeedback. Trends Cogn. Sci. 21, 997–1010. https://doi.org/10.1016/j.tics.2017.09.010.

Wegner-Clemens, K., Rennig, J., Beauchamp, M.S., 2020. A relationship between autism-spectrum quotient and face viewing behavior in 98 participants. PLoS One 15, e0230866. https://doi.org/10.1371/journal.pone.0230866.

Wehbe, L., Blank, I.A., Shain, C., Futrell, R., Levy, R., von der Malsburg, T., Smith, N., Gibson, E., Fedorenko, E., 2021. Incremental language comprehension difficulty predicts activity in the language network but not the multiple demand network. Cereb. Cortex 31, 4006–4023. https://doi.org/10.1093/cercor/bhab065.

Weismer, G., 2008. Speech Intelligibility. the Handbook of Clinical Linguistics. John Wiley & Sons, Ltd, pp. 568–582. https://doi.org/10.1002/9781444301007.ch35.

Wong, P.C.M., Kang, X., So, H.C., Choy, K.W., 2022. Contributions of common genetic variants to specific languages and to when a language is learned. Sci. Rep. 12, 580. https://doi.org/10.1038/s41598-021-04163-1.

Wong, P.C.M., Uppunda, A.K., Parrish, T.B., Dhar, S., 2008. Cortical mechanisms of speech perception in noise. J. Speech Lang. Hear. Res. 51, 1026–1041. https://doi.org/10.1044/1092-4388(2008/075.

Yu, Z., Guindani, M., Grieco, S.F., Chen, L., Holmes, T.C., Xu, X., 2022. Beyond t-test and ANOVA: applications of mixed-effects models for more rigorous statistical analysis in neuroscience research. Neuron 110, 21–35. https://doi.org/10.1016/j.neuron.2021.10.030.