Categorical encoding of speech sounds: beyond auditory cortices

Basil Preisig* (1,2,3,5), Lars Riecke (4) & Alexis Hervais-Adelman (3,5)

Affiliations

1) Donders Institute for Cognitive Neuroimaging, Radboud University, Nijmegen, the Netherlands

2) Max Planck Institute for Psycholinguistics, Nijmegen, the Netherlands

3) Department of Psychology, Neurolinguistics, University of Zurich, Zurich, Switzerland

4) Department of Cognitive Neuroscience, Faculty of Psychology and Neuroscience, Maastricht University, the Netherlands

5) Neuroscience Center Zurich, University of Zurich and Eidgenössische Technische Hochschule Zurich, 8057 Zurich, Switzerland

*Address for correspondence: Basil C. Preisig, PhD, basilpreisig@gmail.com

Abstract

What processes lead to categorical perception of speech sounds? Investigation of this question is hampered by the fact that categorical speech perception is normally confounded by acoustic differences in the stimulus. By using ambiguous sounds, however, it is possible to dissociate acoustic from perceptual stimulus representations. We used a binaural integration task, where the inputs to the two ears were complementary so that phonemic identity emerged from their integration into a single percept. Twenty-seven normally hearing individuals took part in an fMRI study in which they were presented with an ambiguous syllable (intermediate between /da/ and /ga/) in one ear and with a meaning-differentiating acoustic feature (third formant) in the other ear. Multi-voxel pattern searchlight analysis was used to identify brain areas that consistently differentiated between response patterns associated with different syllable reports. By comparing responses to different stimuli with identical syllable reports and identical stimuli with different syllable reports, we disambiguated whether these regions primarily differentiated the acoustics of the stimuli or the syllable report. We found that BOLD activity patterns in the left anterior insula (AI), the left supplementary motor cortex, the left ventral motor cortex and the right motor and somatosensory cortex (M1/S1) represent listeners' syllable report irrespective of stimulus acoustics. The same areas have been previously implicated in decision-making (AI), response selection (SMA), and response initiation and feedback (M1/S1). Our results indicate that the emergence of categorical speech sounds implicates decision-making mechanisms and auditory-motor transformations acting on sensory inputs.

Keywords: Speech perception; Auditory; MVPA; fMRI; Dichotic listening

Significance statement

A central question in psycholinguistic research is whether speech sounds are neutrally coded as abstract perceptual units that are distinct from the sensory cues from which they are derived. One challenge for most studies of perception is to overcome that perceptual interpretations of sensory stimuli may be confounded by physical properties of the stimuli. Here, we use functional magnetic resonance imaging (fMRI) and multi-voxel pattern analysis (MVPA) to address the question of where in the cerebral cortex syllable percepts emerge during binaural integration. By controlling for physical stimulus acoustics, we find that the perceptual report of syllables arises in higher-order non-auditory cortical areas. This opens up the possibility that these areas determine the syllables we hear.

1 Introduction

The mapping of sensory information onto common categories is a fundamental feature of human cognition. For example, we can identify a familiar person on different photographs even if taken from different angles. Likewise, we can map speech sounds onto common words even if uttered from different speakers. Categorical perception in speech was described first by Liberman and colleagues (1957) who showed that synthetic syllables along the continuum between prototypes (e.g., /ba/ vs /da/) were perceived categorically despite their linear acoustic variations. However, it remains unclear how the brain maps the large variety of sensory signals to a limited number of invariant categories.

The neural mechanisms that underlie categorical speech perception have been attributed to different brain regions (Myers et al., 2009). One view is that phonetic invariance is based on acoustic stimulus processing (Blumstein and Stevens, 1981; Diehl et al., 2004) in the auditory association cortex (e.g. the superior temporal gyrus (STG) and the superior temporal (STS)) Evidence comes from studies which identified phonetic representations in STG/STS using fMRI and intracranial recordings (Formisano et al., 2008; Chang et al., 2010; Kilian-Hütten et al., 2011; Mesgarani et al., 2014; Arsenault and Buchsbaum, 2015; Yi et al., 2019; Levy and Wilson, 2020).

Alternatively, the acoustic input may be mapped onto motor patterns or gestures used in producing speech (Liberman et al., 1967). Evidence comes from fMRI studies showing activation (Pulvermüller et al., 2006; Hervais-Adelman et al., 2012) and phoneme identity representations (Lee et al., 2012; Chevillet et al., 2013; Du et al., 2014; Evans and Davis, 2015; Cheung et al., 2016) in motor cortex. Converging evidence is provided by non-invasive brain stimulation studies which perturbed phoneme perception when applied directly over the motor cortex (D'Ausilio et al., 2009; Möttönen and Watkins, 2009; Smalle et al., 2015).

Further, invariant categorical percepts may arise from decision-making mechanisms, i.e. acoustic-to-phoneme mapping is seen as an active cognitive process wherein multiple hypotheses regarding the interpretation of the acoustic pattern are tested (Magnuson and Nusbaum, 2007). Evidence is provided by studies finding invariant neural responses in frontal areas involved in executive processing like the inferior frontal gyrus (Hasson et al., 2007; Myers et al., 2009; Lee et al., 2012) Eventually, it was proposed that invariant percepts arise from mapping the speech input onto higher-level phonological representations. In this case, invariant neural responses to speech sounds of the same category emerge in parietal areas, such as the

left angular gyrus (Blumstein et al., 2005) and the left supramarginal gyrus (Caplan et al., 1995; Zevin and McCandliss, 2005; Raizada and Poldrack, 2007).

A major methodological challenge is to identify brain regions in which neural activity tracks the perceived speech rather than its sensory properties. For instance, if physically different stimuli are used as exemplars of different categories, the perceptual representation of the stimulus is confounded with its physical properties. One approach to overcome this problem is to employ ambiguous stimuli that can elicit different percepts in order to identify brain regions that represent the perceptual report of the participant given the same acoustic stimulus (Kilian-Hütten et al., 2011; Lee et al., 2012). This approach, however, weights perceptual representations resulting from maximal sensory uncertainty more strongly, which may not generalize well to situations with higher sensory evidence.

We presented ambiguous (intermediate between /ga/ and /da/) and unambiguous stimuli (clear /ga/ vs. /da/) together with a meaning-differentiating speech feature (high vs low F3) in a binaural integration paradigm. In the ambiguous condition, the input to each ear was incomplete but complementary so that the perceived syllable identity emerged from the combination of the inputs into a single percept. In the unambiguous condition, the stimulus could be interpreted based on monaural input alone (see Fig. 1).

We used fMRI and searchlight MVPA (Kriegeskorte et al., 2006) to identify brain areas which differentiated between different syllable reports (/ga/ and /da/). In contrast to previous studies (Kilian-Hütten et al., 2011; Lee et al., 2012), we aimed to identify regions that consistently differentiate perceptual reports of both unambiguous and ambiguous stimuli. We tested whether the identified brain regions carry information about the perceived speech (/da/ vs /ga/ response) alone, or its acoustic properties (high vs low F3).

2 Material & Methods

2.1 Participants

Twenty-seven right-handed listeners with no history of hearing impairment (M=21.89 years, SD=3.14, 8 male) took part. The present analysis is based on a dataset collected as part of our previous research on the influence of transcranial brain stimulation on binaural integration (Preisig et al., 2021). All participants had normal or corrected-to-normal visual acuity. The participants reported no history of neurological, psychiatric, nor hearing disorders. All participants had normal hearing (hearing thresholds of less than 25 dB HL at 250, 500, 750,

1000, 1500, 3000, and 4000 Hz, tested on each ear using pure tone audiometry) and no threshold difference between the left ear (LE) and the right ear (RE) larger than 5dB for any of the tested frequencies. All participants gave written informed consent prior to the experiment. This study was approved by the local research ethics committee (CMO region Arnhem-Nijmegen) and was conducted in accordance with the principles of the latest Declaration of Helsinki.

2.2 Experimental Design and Task

The dataset reported in this article comprised four task fMRI runs and one fMRI run with passive listening. The data from four additional task runs during which participants underwent non-invasive brain stimulation are reported elsewhere (Preisig et al., 2021).

Each task fMRI run comprised 128 trials, 88 trials included the presentation of an auditory stimulus (4 trials included sham ramps which were thus discarded). The detailed description of auditory stimuli can be found in our previous reports (Preisig and Sjerps, 2019; Preisig et al., 2020, 2021). Each task fMRI run included 60 binaural integration trials for which the F3 frequency of the RE stimulus was set at the individual category boundary and 24 unambiguous control trials for which the F3 component of the RE stimulus supported 12 times a clear /da/ and 12 times a clear /ga/ interpretation (see Fig. 1). For binaural integration trials, the LE stimulus was 30 times the high F3 and 30 times the low F3. For control trials, LE stimulus included a F3 cue with the same F3 frequency as the RE stimulus. Control trials did not require interhemispheric integration for disambiguation because they could be readily identified based on monaural input alone, i.e., the unambiguous /da/ or /ga/ stimulus presented to the RE.

During task fMRI, each trial was 3 s long (equal to the repetition time of the fMRI sequence) and started with the acquisition of a single fMRI volume (TA = 930 ms). The auditory stimulus was presented approximately 1750 ms after trial onset (Preisig et al., 2021). The presentation of the auditory stimulus lasted 250ms. The participant's response window corresponded to the interval from auditory stimulus onset to 70 ms before the onset of the next trial.

The passive listening fMRI run consisted of 336 trials, 96 trials included auditory stimuli: 48 binaural integration trials and 48 unambiguous trials. Passive listening trials were 2 s long (equal to the repetition time of the fMRI sequence). The auditory stimulus was presented between 1450 and 1550 ms after trial onset. The presentation of the auditory stimulus lasted 250ms.

For additional details on the stimulus presentation see Preisig et al. (2021).

2.3 MRI data acquisition and preprocessing

Anatomical and functional MRI data were acquired with a 3-Tesla Siemens Prisma scanner using a 64-channel head coil. A 3-dimensional high-resolution T1-weighted anatomical volume was acquired using a 3D MPRAGE sequence with the following parameters: repetition time (TR) / inversion time (TI) / echo time (TE) = 2300/1100/3ms, 8° flip angle, FOV 256x216x176 and a 1x1x1 mm isotropic resolution. Parallel imaging (iPAT = GRAPPA) was used to accelerate the acquisition. The acquisition time (TA) of the T1-weighted images was 5 min and 21 sec.

fMRI data was acquired with sparse imaging to minimize the impact of EPI gradient noise during presentation of auditory stimuli (Hall et al., 1999). For this purpose, a delay was introduced in the TR during which the auditory stimuli were presented. This delay was 2070ms during task fMRI and 1070 ms during the passive listening run. For task fMRI each run included 128 echo planar imaging (EPI) volumes. The passive listening run included 336 EPI volumes. Each scan comprised 66 slices of 2mm thickness which were acquired using a interleaved acquisition sequence with multi-band acceleration (TR_{task}: 3000 ms, $TR_{passive\ listening}$: 2000 ms, TA: 930 ms, TE: 34 ms, flip angle: 90 deg, matrix size: 104x104x66, in plane resolution: 2x2x2mm, Multi-band accel. factor: 6x).

fMRI data were pre-processed in SPM12 (http://www.fil.ion.ucl.ac.uk/spm). Preprocessing included the following steps: (1) functional realignment and unwarping, (2) co-registration of the structural image to the mean EPI, (3) normalization of the structural image to a standard template, (4) application of the normalization parameters to all EPI volumes, and (5) spatial smoothing using a Gaussian kernel with a full-width at half maximum of 8 mm.

2.4 Univariate analyses

For the univariate analyses, voxel-wise BOLD activity was modeled by means of a single subject first-level General Linear Model (GLM) using normalized and spatially smoothed images. The model included one regressor coding the onsets of the auditory stimuli and one regressor coding the onset of the participants' button presses during task fMRI. The onsets of the button presses during task fMRI were modelled to account for the BOLD signal variability resulting from different response latencies. For each run, six realignment regressors accounting for movement-related effects and a constant term per functional imaging run were included in the model.

For each participant, T-contrasts (all auditory stimuli > implicit baseline) were computed to identify brain regions that responded significantly to auditory stimuli during passive listening and task. Contrast maps from each subject were summarized at the group level using a one-sample t-test (Fig. 2 & Fig. 3). Based on the task-evoked activation map, a binary mask was generated at a voxelwise threshold of $p<.001$.

2.5 Multi voxel pattern analysis (MVPA)

The MVPA analysis was carried out in subjects' native image space using realigned and unwrapped EPI images (Kriegeskorte et al., 2006). A first-level design matrix was specified including one regressor per condition: unambiguous /da/ report, unambiguous /ga/ report, ambiguous /da/ report, ambiguous /ga/ report. Further, the model included a regressor coding the onset of the participants' button presses. As above, six realignment regressors were further included to account for movement-related effects and a constant term per functional imaging run.

We constrained our MVPA analysis to areas that significantly responded to sound during task fMRI at the group level. For each participant, an individual task-evoked activation map was created by warping the group-level mask (for details see univariate analyses) into the subject's native space using the inverse normalization parameters.

Afterwards, we conducted an MVPA searchlight analysis (sphere radius 8mm, equivalent to 251 voxels) within sound responsive areas to identify brain regions in which different syllable reports (/da/ vs /ga/) elicited distinct spatial BOLD response patterns. To evaluate the representational consistency across unambiguous and ambiguous stimuli, an encoding model was specified using the TDT toolbox (Hebart et al., 2015) such that the fitted regression coefficients (beta values) from the unambiguous trials were used for the training and the beta values from ambiguous trials were used for the test set.

For statistical inference, we computed the crossnobis distance between the response patterns associated with /da/ and /ga/ syllable reports. The crossnobis distance is the cross-validated version of the Mahalanobis distance (multivariate noise normalized Euclidean distance) (Kriegeskorte et al., 2006; Walther et al., 2016). Here, we computed the crossnobis distance using a leave-one-run-out procedure (Allefeld and Haynes, 2014). In each of four cross-validation folds, the beta-value maps derived from unambiguous trials from three fMRI runs were cross-validated against the beta-value maps including ambiguous trials from the left-out run. Cross-validation ensures that the distance is zero if two voxel patterns are not statistically

different from each other, making crossnobis distance a suitable summary statistic for group-level inference (e.g. with the one-sample t-test). Note that because of this cross-validation, the crossnobis distance can take negative values if its true value is close to zero in the presence of noise (Sohoglu et al., 2020).

For group-level inference, individual crossnobis distance maps were normalized and smoothed (using a Gaussian kernel with a full-width at half maximum of 8 mm) and then entered into a group level random-effects analysis using permutation-based nonparametric statistics in SNPM (http://www2.warwick.ac.uk/fac/sci/statistics/staff/academic-research/nichols/software). Family-wise error correction (FWE) for multiple comparisons across voxels was applied at a threshold of $p < .05$.

3 Results

3.1 Binaural integration

Twenty-seven participants listened to ambiguous syllables (intermediate between /ga/ and /da/), whose perceived identity depends upon binaural integration, and to unambiguous syllables (clear /ga/ vs. /da/) that could be readily interpreted based on monaural input (for details see Material & Methods). During task fMRI, participants reported on every trial whether they heard a /da/ or a /ga/ syllable.

Behavioral analyses indicated that participants reliably integrate the speech feature (high vs low F3) in the binaural integration condition. Participants answered on average with 25.40 ± 2.80% (mean ± SEM) /ga/ responses and 70.60% ± 3.00% /da/ responses to ambiguous syllables combined with the high F3 and 75.00% ± 3.5% /ga/ responses and 20.70% ± 3.10% /da/ responses for ambiguous syllables combined with the low F3.

The probability of reporting an unambiguous stimulus with low or high F3 respectively as a /ga/ or /da/ syllable was high: Participants gave on average 84.4% ± 2.2% /da/ responses to unambiguous /da/ stimuli, and 87.4 ± 1.8% (mean ± SEM) /ga/ responses were registered for unambiguous /ga/ stimuli.

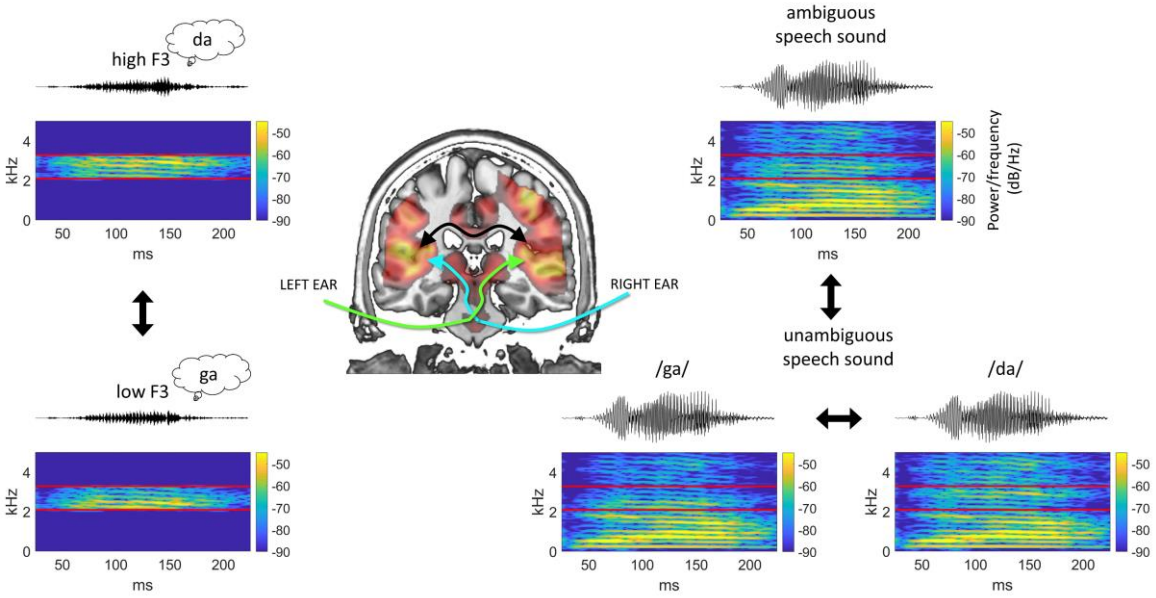

Figure 1. Adapted from (Preisig et al., 2021). (Center) Schematic illustration of the processing pathway underlying binaural integration. The green line indicates the propagation of acoustic input from the left ear to the right auditory cortex and from the right ear to the left auditory cortex. The black line illustrates the interhemispheric connection between the auditory cortices via the corpus callosum. (Left) Sound pressure waveform and corresponding sound spectrogram of the non-speech acoustic feature (F3) presented to the left ear. (Upper left) High F3 supporting a /da/ interpretation. (Lower left) Low F3 supporting a /ga/ interpretation. (Right) Sound pressure waveform and corresponding spectrogram of the syllables presented to the right ear. (Upper right) Ambiguous speech sound intermediate to the syllables /da/ and /ga/. (Lower right) Unambiguous /ga/ and /da/ syllables. The red lines in the spectrogram highlight the spectral area of stimulus manipulation.

3.2 Auditory activation during passive listening and task

In order to localize areas that respond to our syllable stimuli in the absence of any task and motor response, we mapped auditory-evoked activity during passive listening. Further, we also mapped the areas that were additionally activated by the task (see Figure 3B). For this purpose, we examined BOLD responses on all trials, contrasted with the baseline (see Methods). During passive listening and task, whole-brain analysis revealed an extended bilateral brain network including the bilateral supratemporal plane, inferior frontal areas, and motor cortical areas (see Fig. 2). In the task blocks, we observed more extensive activation in these cortical areas and

additional occipital as well as subcortical activation. Based on the task-evoked activation map, a binary mask was generated at a voxelwise threshold of p<.001 for subsequent searchlight MVPA.

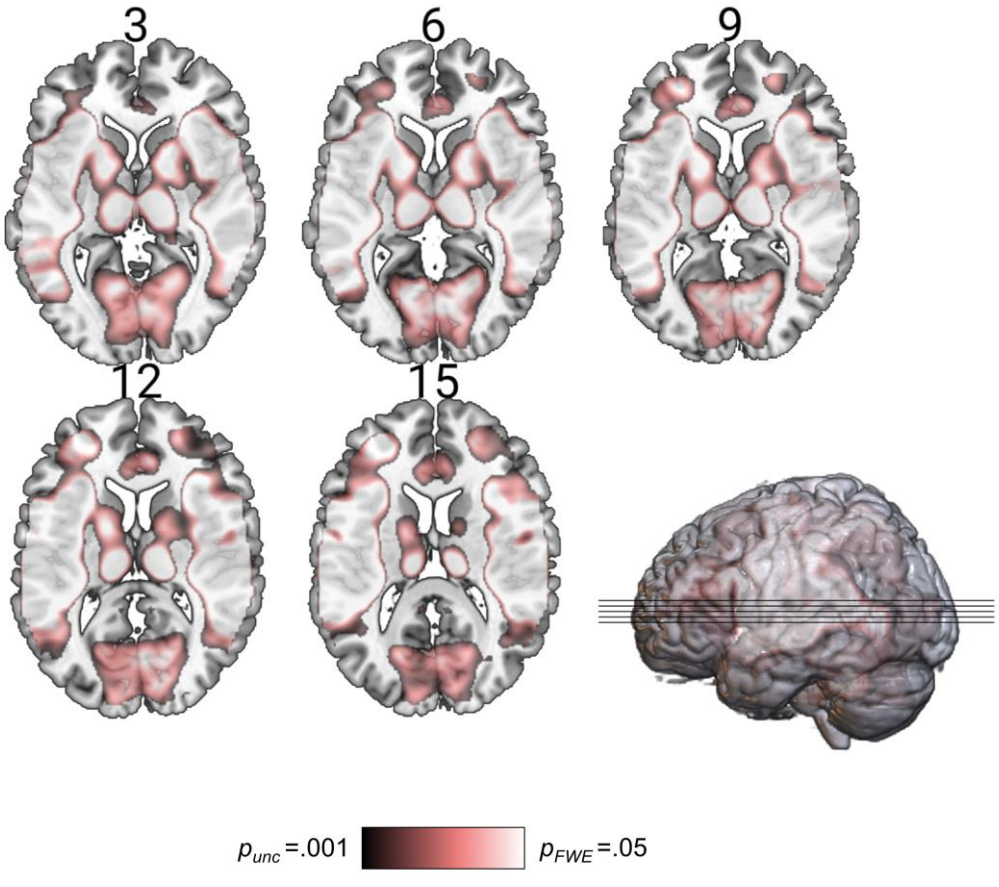

Figure 2. Results of the univariate analyses. For each participant, T-contrasts (all trials > baseline) were computed to identify brain regions that responded significantly to auditory stimuli and the task. Contrast maps from each subject were statistically assessed at the group level using a one-sample t-test. Based on this map of sound- and task-evoked responses, a binary mask was computed at a voxelwise threshold of p<.001. Numbers refer to the slice position in z direction.

3.3 Categorical representations outside of the auditory cortices

We used a searchlight procedure (Kriegeskorte et al., 2006) to compute the cross-validated Mahalanobis distance (hereafter crossnobis distance). The crossnobis distance here refers to the Euclidean distance after multivariate noise normalization (Walther et al., 2016) between the BOLD response patterns associated with different perceptual reports (/da/ vs /ga/). BOLD response patterns were extracted separately for unambiguous and ambiguous stimulus trials and

constrained to areas significantly responding to sound at the group level (see auditory task-evoked activation mask) in participants' native brain space.

We computed the crossnobis distance between /da/ and /ga/ reports as the arithmetic product of the perceptual distances in unambiguous and ambiguous stimuli.

$$(/da/\ report_{unambiguous} - /ga/\ report_{unambiguous}) \times (/da/\ report_{ambiguous} - /ga/\ report_{ambiguous})$$

In this way, we sought to identify BOLD patterns that consistently differentiated perceptual reports across unambiguous stimuli and ambiguous stimuli. Thus, BOLD response patterns which consistently differentiated /da/ vs. /ga/ reports in both unambiguous and ambiguous stimuli yielded greater distances, while inconsistent BOLD patterns yielded smaller distances.

Group-level analysis (random-effects analysis using permutation-based nonparametric statistics) of normalized and smoothed distance maps revealed significant ($p < .05$ FWE-corrected) BOLD activity patterns in the left AI, the left SMA, the left vMC, and the right M1/S1 that reliably represented the participants' syllable report (see Fig. 3 & Table S1). The results of a similar analysis in unambiguous stimuli are presented in Fig S1. It should be noted that, as physical stimulus and its perceptual interpretation were confounded in this whole-brain analysis, the observed categorical response patterns could be driven by the stimulus acoustics, the syllable percept, or both.

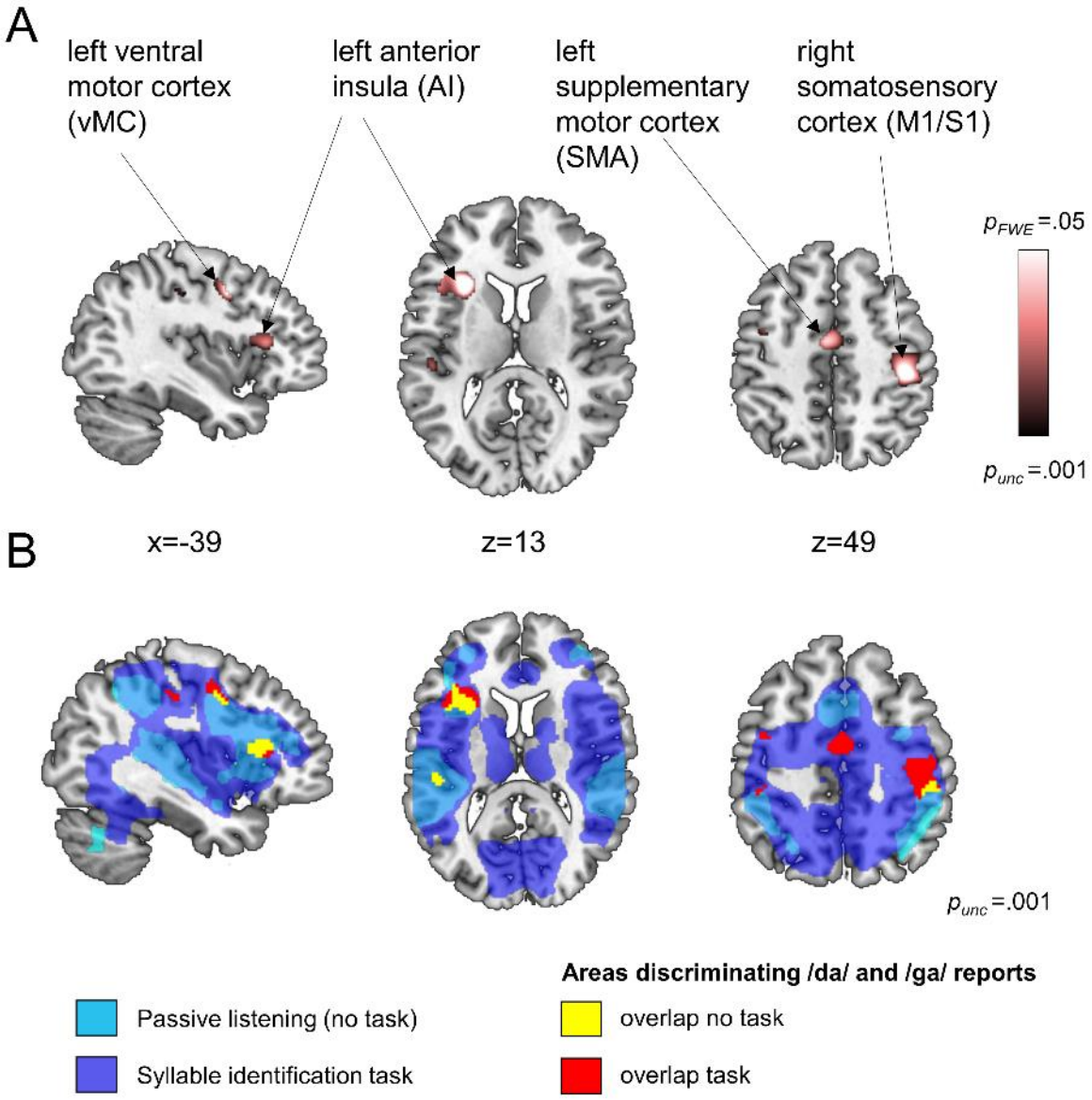

Figure 3. Results of the MVPA searchlight analyses projected onto a canonical MNI single-subject brain. (A) In the highlighted regions, the average crossnobis distance associated with different syllable reports was found to be significantly larger than zero at $p_{FWE} < .05$. (B) The clusters presented in (A) are thresholded here at $p_{unc} < .001$ and overlaid on regions activated during auditory stimulus presentation and passive listening (light blue color) and regions activated during auditory stimulus presentation and task (dark blue color).

3.4 Acoustic or phonemic representation?

In follow-up analyses, we tested whether categorical patterns derived from the unambiguous stimuli in the localized regions generalize better to the stimulus percept (/da/ vs /ga/) within the same acoustic stimulus, or to the presented acoustic stimulus (high vs low F3) within the same stimulus percept.

For this purpose, we recomputed the crossnobis distance between syllable reports within each stimulus class (high; low F3) separately. The converse was also calculated – the distance between each acoustic stimulus, within each syllable report (/da/; /ga/). In both cases the distances were cross-validated as previously, against the syllable reports in unambiguous stimuli. We found better generalization to the stimulus percept in the aforementioned areas, as reflected by larger crossnobis distances between different syllable percepts than between different stimulus acoustics (percept: $1.297 \times 10^6 \pm 3.332 \times 10^5$; acoustic: $-3.210 \times 10^5 \pm 3.635 \times 10^5$, mean ± SEM, *Shapiro-Wilk Test of Normality* $p<.001$, *Paired-samples Wilcoxon Sign Rank Test*, $Z=2.033$, $p=.042$, effect size: $r=.40$). In all brain regions but the left vMC, the distances between different syllable percepts were significantly different from zero (*One sample Wilcoxon Tests*, left AI: $Z=2.375$, $p_{holm}=.035$; left SMA: $Z=2.983$, $p_{holm}=.006$; left vMC: $Z=0.855$, $p_{holm}=.785$; right M1/S1 $Z=2.733$, $p_{holm}=.013$), whereas distances between acoustically different stimuli did not show any significant difference from zero ($ps_{uncorrected}>.305$). We found no statistically significant difference in the crossnobis distance between categories across different brain regions (left AI: $4.291 \times 10^5 \pm 1.134 \times 10^5$; left SMA: $4.273 \times 10^5 \pm 1.110 \times 10^5$; left vMC: $2.720 \times 10^5 \pm 8.421 \times 10^4$; right M1/S1: $8.228 \times 10^5 \pm 1.613 \times 10^5$, mean ± SEM, *Friedman ANOVA*. $\chi^2(3)= 6.543$, $p=.088$, effect size: $W = .10$). Pairwise post-hoc comparisons revealed that the crossnobis distance between categories was significantly larger between different syllable percepts than between different stimulus acoustics in the left SMA (*Paired-samples Wilcoxon Sign Rank Test*, $Z=2.572$, $p_{holm}=.041$, effect size: $W=.56$), but not the other regions ($ps_{uncorrected}>.023$) (see Fig. 4).

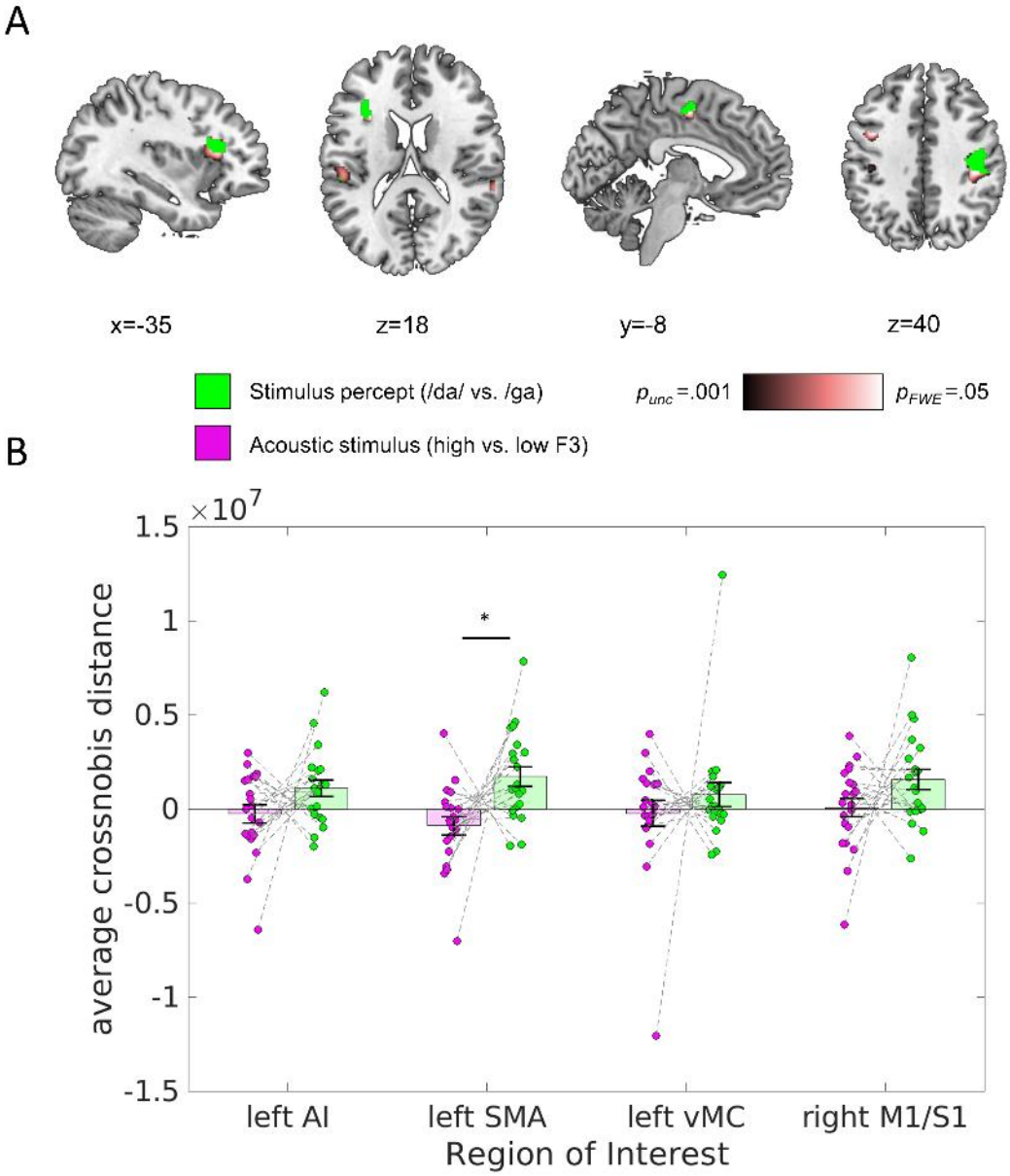

Figure 4. (A) Follow-up MVPA analysis constrained to the regions presented in Figure 3. In green, crossnobis distance between /da/ vs /ga/ percept within the same acoustic stimulus ($p_{(unc)} < .01$) is shown, overlaid on the group map presented in Fig 3. At the same threshold, we found no significant clusters for the crossnobis distance maps between high vs low F3 acoustic stimulus. (B) Average crossnobis distance extracted from the regions presented in Figure 3 form categorical patterns representing the distance between different acoustic stimuli (magenta) and different syllable percepts (green). * $p < .01$ Paired-samples Wilcoxon test, corrected for multiple comparisons (Holm-Bonferroni).

4 Discussion

In the present study, we aimed to identify a level of neural representation where speech sounds are coded as abstract categorical perceptual units that are invariant to the sensory signals from which they are derived. For this purpose, we first localized brain regions which consistently discriminated the syllable reports evoked by different unambiguous and ambiguous stimuli. We then assessed whether these regions discriminated primarily the acoustics of the stimuli or the syllable report. Our results show that the perceptual report of the syllable arises in a set of regions which include the left SMA, the left AI, the left vMC and the right M1/S1. These regions are outside of what is traditionally regarded as auditory or phonological processing areas, such as the auditory cortex. For the interpretation of these findings, it is important to keep in mind that a task (listeners had to report the syllable they perceived by button press) typically embeds perception in a context of decision-making. The neural representations that we identified in our experiment may hence reflect: (1) the auditory percept, (2) categorization of percept into syllable, i.e., decision-making, (3) motor response planning and execution, or (4) a combination of different processes along this hierarchy. In the following section we will discuss these alternative interpretations in the light of previous findings in these brain regions.

4.1 Auditory perception

Speech perception is typically seen as a hierarchical process because stronger activation is found in high-order auditory areas to complex information-bearing stimuli, such as speech and music, than to simple stimuli, such as pure tones (Schönwiesner and Zatorre, 2009; Leaver and Rauschecker, 2010; Norman-Haignere et al., 2015). Therefore, it is a commonly held belief that low-level acoustic features are extracted in early auditory areas, which are later transformed into more complex and speech-specific representations in the nonprimary cortex in the superior temporal gyrus (Rauschecker and Scott, 2009; Brodbeck et al., 2018). In this study, we found no area that represented the acoustic difference (high vs low F3) in our stimuli. This might be explained by the relatively subtle acoustic difference between our stimuli that may be difficult to resolve from the fMRI data. It is noteworthy that there is evidence that challenges this hierarchical processing structure, favoring parallel streams of auditory processing (Hackett et al., 2001; Jasmin et al., 2019; Hamilton et al., 2021).

4.2 Categorization of percept into syllable

Existing data suggest the involvement of numerous brain areas in processing of sublexical units, including pSTG/STS (Formisano et al., 2008; Chang et al., 2010; Kilian-Hütten et al., 2011;

Mesgarani et al., 2014; Arsenault and Buchsbaum, 2015; Yi et al., 2019; Levy and Wilson, 2020), IFG (Hasson et al., 2007; Myers et al., 2009; Lee et al., 2012; Chevillet et al., 2013; Du et al., 2014) as well as vMC (Du et al., 2014; Evans and Davis, 2015; Cheung et al., 2016) and premotor cortex (Chevillet et al., 2013). It has been shown that during listening the neural responses in the pSTG/STS (Mesgarani et al., 2014) and vMC (Cheung et al., 2016) show a similar spatial organization along acoustic features. This is interesting because Cheung and colleagues (2016) also found, that in contrast to listening, the neural responses in the vMC during speaking are organized along articulatory features. We found that the left vMC discriminates between syllable reports in ambiguous and unambiguous stimuli. However, when testing whether the left vMC discriminates syllable percepts while keeping the stimulus acoustics constant, we found no significant difference. Further, we found evidence for categorical representations of unambiguous stimuli in the left STS (Figure S1). These results indicate that categorical representations in the left vMC and STG/STS may dependent on the acoustic properties of the stimulus.

4.3 Syllable percepts emerge outside of the auditory cortices

The searchlight approach allowed us to identify areas discriminating syllable percepts that have rarely been associated with categorical speech perception: the left SMA and the left AI. In the midline motor areas, the strongest responses to auditory stimuli have been reported in pre-SMA and at the boundary area between pre-SMA and SMA (Lima et al., 2016). The latter overlaps with the area we identified in our study (Mayka et al., 2006; Kim et al., 2010). The SMA receives direct projections from the basal ganglia (Lehéricy et al., 2004; Akkal et al., 2007), the STG/STS (Luppino et al., 1993, 2001; Reznik et al., 2015), and the inferior-parietal (Luppino et al., 1993) and inferior-frontal cortices (Catani et al., 2012; Vergani et al., 2014). The SMA is not typically considered as a part of the speech and language network in the brain (Hickok and Poeppel, 2007; Friederici, 2011; Hagoort, 2014), despite its connections with this network. Although, SMA activity in response to speech and non-speech sounds has been reported in several studies (for reviews see (Hertrich et al., 2016; Lima et al., 2016)), the functional role of the SMA in auditory speech processing has remained elusive, possibly because SMA and pre-SMA are traditionally conceptualized as being linked to action-related processes, like speech motor control (Tourville and Guenther, 2011), rather than auditory processes (Lima et al., 2016).

The left AI cluster that we identified includes a portion of the dorsal anterior insula. Insula activity during speech perception has been reported in several studies (Benson et al., 2001;

Golestani and Zatorre, 2004; Aleman et al., 2005; Falkenberg et al., 2011) Particularly, the dorsal AI seems to be involved in speech perception, whereas other parts of the insula seem to contribute more to speech production (for a meta-analysis see (Oh et al., 2014)). Similar to the SMA, the insula has extensive connections with the auditory cortex, temporal pole, and superior temporal sulcus (Augustine, 1996; Oh et al., 2014), and is typically not considered to be involved in speech perception (Hickok and Poeppel, 2007; Friederici, 2011; Hagoort, 2014).

4.4 The putative role of the SMA and AI in auditory decision making

Recent evidence suggests that SMA may play an important role for the categorization of auditory percepts into syllables, including decision-making. In a recent study, Morán and colleagues (2021) found categorical neural responses in rhesus monkeys, who were trained to categorize numerous complex sounds, including words. The authors recorded extracellular activity directly from SMA neurons and found robust categorical responses at both the single neuron and population levels. Most importantly, they observed that neural population coding shifted from acoustic to motor representations during the task suggesting that the SMA integrates acoustic information in order to form categorical signals that drive the behavioral response. Interestingly, the population activity in error trials reflected the behavioral decision, rather than the presented physical stimulus.

The left AI has been associated with decision making in a number of studies (for a review see (Droutman et al., 2015)). Particularly, the dorsal AI seems to be functionally implicated in cognitive control processes associated with decision-making including attention re-focusing, evaluation, action, and outcome processing (Droutman et al., 2015). Further, dorsal AI activation has been related to decision ambiguity (Huettel et al., 2006) and error awareness (Ullsperger et al., 2010). Recent evidence indicates that the anterior insula could serve as a gate for conscious access to sensory information (Huang et al., 2021).

We found two interesting distinctions between the BOLD patterns in the left SMA and the left AI. First, only the cluster in the left AI overlapped with areas which were activated during passive listening to the same stimuli (Fig. 3). Second, only in the left SMA, but not in the left AI, the crossnobis distances between syllable percepts were larger between different syllable percepts than between different stimulus acoustics. This indicates the responses in the left SMA, but not left AI, were invariant to the physical stimulus properties of the stimuli. Taken together, one may speculate about the processing hierarchy and the functional differences between the processing in AI and the SMA. The AI may be more relevant for the integration of auditory

436 evidence and the preparation of a perceptual representation, which is then strengthened to
437 become an invariant representation in the SMA.

438 **4.5 Motor response planning and execution**

439 A possible alternative interpretation is that the effects reported in the SMA and the AI reflect
440 merely task-related motor or domain-general cognitive processes, such as button presses or
441 response selection (Zatorre et al., 1992; Kawashima et al., 1996). This view is supported by our
442 observation that the right M1/S1 differentiated between different syllable percepts. However, it
443 is likely that additional processes, other than button pressing, also contributed to the effect. This
444 is supported by studies showing SMA (Benson et al., 2001; Scott et al., 2004; Warren et al.,
445 2006; Jardri et al., 2007) and insula (Ackermann et al., 2001; Benson et al., 2001; Steinbrink et
446 al., 2009; Hervais-Adelman et al., 2012) activation during passive listening. Further, there are
447 studies showing activation in the SMA and the insula during auditory task on top of merely
448 motor-execution related activity (Adank et al., 2013; Bestelmeyer et al., 2014; Sammler et al.,
449 2015). Thus, our observed effect may reflect either task- and motor-execution-related activity,
450 or tactile feedback. Our interpretation is further supported by results from the passive condition,
451 which revealed activation in the anterior insula and pre-SMA during listening only.

452 **4.6 Conclusion**

453 Our finding that BOLD patterns discriminating different syllable reports occur in brain regions
454 contributing to auditory processing, categorization of percept into syllables, i.e., decision-
455 making, and motor response, suggests that the identified representations reflect a combination
456 of various processes along this hierarchy.

457 Our finding that areas whose responses to speech stimuli discriminate between phonemic
458 categories exist outside auditory cortical areas is consistent with the possibility that higher-
459 order areas are instrumental in determining the syllables we hear, and that these regions feed
460 abstract categorical representations back into the auditory association cortex (Formisano et al.,
461 2008; Chang et al., 2010; Kilian-Hütten et al., 2011) and even to earlier auditory areas (Levy
462 and Wilson, 2020).

463

5 Acknowledgement

This work was supported by the Swiss National Science Foundation [P2BEP3_168728 / PP00P1_163726] and the Janggen-Pöhn Stiftung. The authors would like to thank Benjamin Kop, Brigit Knudsen, Iris Schmits, Uriel Plones, and Paul Gaalman for their assistance as well as Martin Hebart for the methodological advice with regard to the MVPA analysis.

6 References

Ackermann H, Riecker A, Mathiak K, Erb M, Grodd W, Wildgruber D (2001) Rate-dependent activation of a prefrontal-insular-cerebellar network during passive listening to trains of click stimuli: An fMRI study. NeuroReport 12:4087–4092.

Adank P, Rueschemeyer S-A, Bekkering H (2013) The role of accent imitation in sensorimotor integration during processing of intelligible speech. Front Hum Neurosci 7 Available at: https://www.frontiersin.org/articles/10.3389/fnhum.2013.00634/full [Accessed July 8, 2021].

Akkal D, Dum RP, Strick PL (2007) Supplementary Motor Area and Presupplementary Motor Area: Targets of Basal Ganglia and Cerebellar Output. J Neurosci 27:10659–10673.

Aleman A, Formisano E, Koppenhagen H, Hagoort P, De Haan EHF, Kahn RS (2005) The functional neuroanatomy of metrical stress evaluation of perceived and imagined spoken words. Cereb Cortex 15:221–228.

Allefeld C, Haynes J-D (2014) Searchlight-based multi-voxel pattern analysis of fMRI by cross-validated MANOVA. NeuroImage 89:345–357.

Arsenault JS, Buchsbaum BR (2015) Distributed Neural Representations of Phonological Features during Speech Perception. J Neurosci 35:634–642.

Augustine JR (1996) Circuitry and functional aspects of the insular lobe in primates including humans. Brain Res Rev 22:229–244.

Benson RR, Whalen DH, Richardson M, Swainson B, Clark VP, Lai S, Liberman AM (2001) Parametrically Dissociating Speech and Nonspeech Perception in the Brain Using fMRI. Brain Lang 78:364–396.

Bestelmeyer PEG, Maurage P, Rouger J, Latinus M, Belin P (2014) Adaptation to Vocal Expressions Reveals Multistep Perception of Auditory Emotion. J Neurosci 34:8098–8105.

Blumstein SE, Myers EB, Rissman J (2005) The Perception of Voice Onset Time: An fMRI Investigation of Phonetic Category Structure. J Cogn Neurosci 17:1353–1366.

Blumstein SE, Stevens KN (1981) Phonetic features and acoustic invariance in speech. Cognition 10:25–32.

Brodbeck C, Hong LE, Simon JZ (2018) Rapid Transformation from Auditory to Linguistic Representations of Continuous Speech. Curr Biol 28:3976-3983.e5.

Caplan D, Gow D, Makris N (1995) Analysis of lesions by MRI in stroke patients with acoustic-phonetic processing deficits. Neurology 45:293–298.

Catani M, Dell'Acqua F, Vergani F, Malik F, Hodge H, Roy P, Valabregue R, Thiebaut de Schotten M (2012) Short frontal lobe connections of the human brain. Cortex 48:273–291.

506 Chang EF, Rieger JW, Johnson K, Berger MS, Barbaro NM, Knight RT (2010) Categorical
507 speech representation in human superior temporal gyrus. Nat Neurosci 13:1428–1432.

508 Cheung C, Hamilton LS, Johnson K, Chang EF (2016) The auditory representation of speech
509 sounds in human motor cortex Shinn-Cunningham BG, ed. eLife 5:e12577.

510 Chevillet MA, Jiang X, Rauschecker JP, Riesenhuber M (2013) Automatic Phoneme Category
511 Selectivity in the Dorsal Auditory Stream. J Neurosci 33:5208–5215.

512 D'Ausilio A, Pulvermüller F, Salmas P, Bufalari I, Begliomini C, Fadiga L (2009) The Motor
513 Somatotopy of Speech Perception. Curr Biol 19:381–385.

514 Diehl RL, Lotto AJ, Holt LL (2004) Speech Perception. Annu Rev Psychol 55:149–179.

515 Droutman V, Bechara A, Read SJ (2015) Roles of the Different Sub-Regions of the Insular
516 Cortex in Various Phases of the Decision-Making Process. Front Behav Neurosci 9
517 Available at: https://www.frontiersin.org/articles/10.3389/fnbeh.2015.00309/full
518 [Accessed May 21, 2019].

519 Du Y, Buchsbaum BR, Grady CL, Alain C (2014) Noise differentially impacts phoneme
520 representations in the auditory and speech motor systems. Proc Natl Acad Sci
521 111:7126–7131.

522 Evans S, Davis MH (2015) Hierarchical Organization of Auditory and Motor Representations
523 in Speech Perception: Evidence from Searchlight Similarity Analysis. Cereb Cortex
524 25:4772–4788.

525 Falkenberg LE, Specht K, Westerhausen R (2011) Attention and cognitive control networks
526 assessed in a dichotic listening fMRI study. Brain Cogn 76:276–285.

527 Formisano E, Martino FD, Bonte M, Goebel R (2008) "Who" Is Saying "What"? Brain-Based
528 Decoding of Human Voice and Speech. Science 322:970–973.

529 Friederici AD (2011) The brain basis of language processing: From structure to function.
530 Physiol Rev 91:1357–1392.

531 Golestani N, Zatorre RJ (2004) Learning new sounds of speech: reallocation of neural
532 substrates. NeuroImage 21:494–506.

533 Hackett TA, Preuss TM, Kaas JH (2001) Architectonic identification of the core region in
534 auditory cortex of macaques, chimpanzees, and humans. J Comp Neurol 441:197–222.

535 Hagoort P (2014) Nodes and networks in the neural architecture for language: Broca's region
536 and beyond. Curr Opin Neurobiol 28:136–141.

537 Hall DA, Haggard MP, Akeroyd MA, Palmer AR, Summerfield AQ, Elliott MR, Gurney EM,
538 Bowtell RW (1999) "sparse" temporal sampling in auditory fMRI. Hum Brain Mapp
539 7:213–223.

540 Hamilton LS, Oganian Y, Hall J, Chang EF (2021) Parallel and distributed encoding of speech
541 across human auditory cortex. Cell Available at:

https://www.sciencedirect.com/science/article/pii/S0092867421008783 [Accessed August 19, 2021].

Hasson U, Skipper JI, Nusbaum HC, Small SL (2007) Abstract Coding of Audiovisual Speech: Beyond Sensory Representation. Neuron 56:1116–1126.

Hebart MN, Görgen K, Haynes J-D (2015) The Decoding Toolbox (TDT): a versatile software package for multivariate analyses of functional imaging data. Front Neuroinformatics 8 Available at: https://www.frontiersin.org/articles/10.3389/fninf.2014.00088/full [Accessed February 25, 2019].

Hertrich I, Dietrich S, Ackermann H (2016) The role of the supplementary motor area for speech and language processing. Neurosci Biobehav Rev 68:602–610.

Hervais-Adelman AG, Carlyon RP, Johnsrude IS, Davis MH (2012) Brain regions recruited for the effortful comprehension of noise-vocoded words. Lang Cogn Process 27:1145–1166.

Hickok G, Poeppel D (2007) The cortical organization of speech processing. Nat Rev Neurosci 8:393–402.

Huang Z, Tarnal V, Vlisides PE, Janke EL, McKinney AM, Picton P, Mashour GA, Hudetz AG (2021) Anterior insula regulates brain network transitions that gate conscious access. Cell Rep 35:109081.

Huettel SA, Stowe CJ, Gordon EM, Warner BT, Platt ML (2006) Neural Signatures of Economic Preferences for Risk and Ambiguity. Neuron 49:765–775.

Jardri R, Pins D, Bubrovszky M, Despretz P, Pruvo J-P, Steinling M, Thomas P (2007) Self awareness and speech processing: An fMRI study. NeuroImage 35:1645–1653.

Jasmin K, Lima CF, Scott SK (2019) Understanding rostral–caudal auditory cortex contributions to auditory perception. Nat Rev Neurosci 20:425.

Kawashima R, Satoh K, Itoh H, Ono S, Furumoto S, Gotoh R, Koyama M, Yoshioka S, Takahashi T, Takahashi K, Yanagisawa T, Fukuda H (1996) Functional anatomy of GO/NO-GO discrimination and response selection — a PET study in man. Brain Res 728:79–89.

Kilian-Hütten N, Valente G, Vroomen J, Formisano E (2011) Auditory Cortex Encodes the Perceptual Interpretation of Ambiguous Sound. J Neurosci 31:1715–1720.

Kim J-H, Lee J-M, Jo HJ, Kim SH, Lee JH, Kim ST, Seo SW, Cox RW, Na DL, Kim SI, Saad ZS (2010) Defining functional SMA and pre-SMA subregions in human MFC using resting state fMRI: Functional connectivity-based parcellation method. NeuroImage 49:2375–2386.

Kriegeskorte N, Goebel R, Bandettini P (2006) Information-based functional brain mapping. Proc Natl Acad Sci 103:3863–3868.

Leaver AM, Rauschecker JP (2010) Cortical Representation of Natural Complex Sounds: Effects of Acoustic Features and Auditory Object Category. J Neurosci 30:7604–7612.

Lee Y-S, Turkeltaub P, Granger R, Raizada RDS (2012) Categorical Speech Processing in Broca's Area: An fMRI Study Using Multivariate Pattern-Based Analysis. J Neurosci 32:3942–3948.

Lehéricy S, Ducros M, Krainik A, Francois C, Van de Moortele P-F, Ugurbil K, Kim D-S (2004) 3-D Diffusion Tensor Axonal Tracking shows Distinct SMA and Pre-SMA Projections to the Human Striatum. Cereb Cortex 14:1302–1309.

Levy DF, Wilson SM (2020) Categorical Encoding of Vowels in Primary Auditory Cortex. Cereb Cortex 30:618–627.

Liberman AM, Cooper FS, Shankweiler DP, Studdert-Kennedy M (1967) Perception of the speech code. Psychol Rev 74:431.

Liberman AM, Harris KS, Hoffman HS, Griffith BC (1957) The discrimination of speech sounds within and across phoneme boundaries. J Exp Psychol 54:358–368.

Lima CF, Krishnan S, Scott SK (2016) Roles of Supplementary Motor Areas in Auditory Processing and Auditory Imagery. Trends Neurosci 39:527–542.

Luppino G, Calzavara R, Rozzi S, Matelli M (2001) Projections from the superior temporal sulcus to the agranular frontal cortex in the macaque. Eur J Neurosci 14:1035–1040.

Luppino G, Matelli M, Camarda R, Rizzolatti G (1993) Corticocortical connections of area F3 (SMA-proper) and area F6 (pre-SMA) in the macaque monkey. J Comp Neurol 338:114–140.

Magnuson JS, Nusbaum HC (2007) Acoustic differences, listener expectations, and the perceptual accommodation of talker variability. J Exp Psychol Hum Percept Perform 33:391–409.

Mayka MA, Corcos DM, Leurgans SE, Vaillancourt DE (2006) Three-dimensional locations and boundaries of motor and premotor cortices as defined by functional brain imaging: A meta-analysis. NeuroImage 31:1453–1474.

Mesgarani N, Cheung C, Johnson K, Chang EF (2014) Phonetic feature encoding in human superior temporal gyrus. Science 343:1006–1010.

Morán I, Perez-Orive J, Melchor J, Figueroa T, Lemus L (2021) Auditory decisions in the supplementary motor area. Prog Neurobiol 202:102053.

Möttönen R, Watkins KE (2009) Motor representations of articulators contribute to categorical perception of speech sounds. J Neurosci Off J Soc Neurosci 29:9819–9825.

Myers EB, Blumstein SE, Walsh E, Eliassen J (2009) Inferior Frontal Regions Underlie the Perception of Phonetic Category Invariance. Psychol Sci 20:895–903.

Norman-Haignere S, Kanwisher NG, McDermott JH (2015) Distinct Cortical Pathways for Music and Speech Revealed by Hypothesis-Free Voxel Decomposition. Neuron 88:1281–1296.

Oh A, Duerden EG, Pang EW (2014) The role of the insula in speech and language processing. Brain Lang 135:96–103.

Preisig BC, Riecke L, Sjerps MJ, Kösem A, Kop BR, Bramson B, Hagoort P, Hervais-Adelman A (2021) Selective modulation of interhemispheric connectivity by transcranial alternating current stimulation influences binaural integration. Proc Natl Acad Sci 118 Available at: https://www.pnas.org/content/118/7/e2015488118 [Accessed February 11, 2021].

Preisig BC, Sjerps MJ (2019) Hemispheric specializations affect interhemispheric speech sound integration during duplex perception. J Acoust Soc Am 145:EL190–EL196.

Preisig BC, Sjerps MJ, Hervais-Adelman A, Kösem A, Hagoort P, Riecke L (2020) Bilateral Gamma/Delta Transcranial Alternating Current Stimulation Affects Interhemispheric Speech Sound Integration. J Cogn Neurosci 32:1242–1250.

Pulvermüller F, Huss M, Kherif F, Martin FM del P, Hauk O, Shtyrov Y (2006) Motor cortex maps articulatory features of speech sounds. Proc Natl Acad Sci 103:7865–7870.

Raizada RDS, Poldrack RA (2007) Selective Amplification of Stimulus Differences during Categorical Processing of Speech. Neuron 56:726–740.

Rauschecker JP, Scott SK (2009) Maps and streams in the auditory cortex: nonhuman primates illuminate human speech processing. Nat Neurosci 12:718–724.

Reznik D, Ossmy O, Mukamel R (2015) Enhanced Auditory Evoked Activity to Self-Generated Sounds Is Mediated by Primary and Supplementary Motor Cortices. J Neurosci 35:2173–2180.

Sammler D, Grosbras M-H, Anwander A, Bestelmeyer PEG, Belin P (2015) Dorsal and Ventral Pathways for Prosody. Curr Biol 25:3079–3085.

Schönwiesner M, Zatorre RJ (2009) Spectro-temporal modulation transfer function of single voxels in the human auditory cortex measured with high-resolution fMRI. Proc Natl Acad Sci 106:14611–14616.

Scott SK, Rosen S, Wickham L, Wise RJS (2004) A positron emission tomography study of the neural basis of informational and energetic masking effects in speech perception. J Acoust Soc Am 115:813–821.

Smalle EHM, Rogers J, Möttönen R (2015) Dissociating Contributions of the Motor Cortex to Speech Perception and Response Bias by Using Transcranial Magnetic Stimulation. Cereb Cortex 25:3690–3698.

Sohoglu E, Kumar S, Chait M, Griffiths TD (2020) Multivoxel codes for representing and integrating acoustic features in human cortex. NeuroImage 217:116661.

Steinbrink C, Ackermann H, Lachmann T, Riecker A (2009) Contribution of the anterior insula to temporal auditory processing deficits in developmental dyslexia. Hum Brain Mapp 30:2401–2411.

653 Tourville JA, Guenther FH (2011) The DIVA model: A neural theory of speech acquisition and
654 production. Lang Cogn Process 26:952–981.

655 Ullsperger M, Harsay HA, Wessel JR, Ridderinkhof KR (2010) Conscious perception of errors
656 and its relation to the anterior insula. Brain Struct Funct 214:629–643.

657 Vergani F, Lacerda L, Martino J, Attems J, Morris C, Mitchell P, Schotten MT de, Dell'Acqua
658 F (2014) White matter connections of the supplementary motor area in humans. J Neurol
659 Neurosurg Psychiatry 85:1377–1385.

660 Walther A, Nili H, Ejaz N, Alink A, Kriegeskorte N, Diedrichsen J (2016) Reliability of
661 dissimilarity measures for multi-voxel pattern analysis. NeuroImage 137:188–200.

662 Warren JE, Sauter DA, Eisner F, Wiland J, Dresner MA, Wise RJS, Rosen S, Scott SK (2006)
663 Positive emotions preferentially engage an auditory-motor "mirror" system. J Neurosci
664 26:13067–13075.

665 Yi HG, Leonard MK, Chang EF (2019) The Encoding of Speech Sounds in the Superior
666 Temporal Gyrus. Neuron 102:1096–1110.

667 Zatorre RJ, Evans AC, Meyer E, Gjedde A (1992) Lateralization of phonetic and pitch
668 discrimination in speech processing. Science 256:846–849.

669 Zevin JD, McCandliss BD (2005) Dishabituation of the BOLD response to speech sounds.
670 Behav Brain Funct 1:4.

671

7 Supplementary Figures and Table

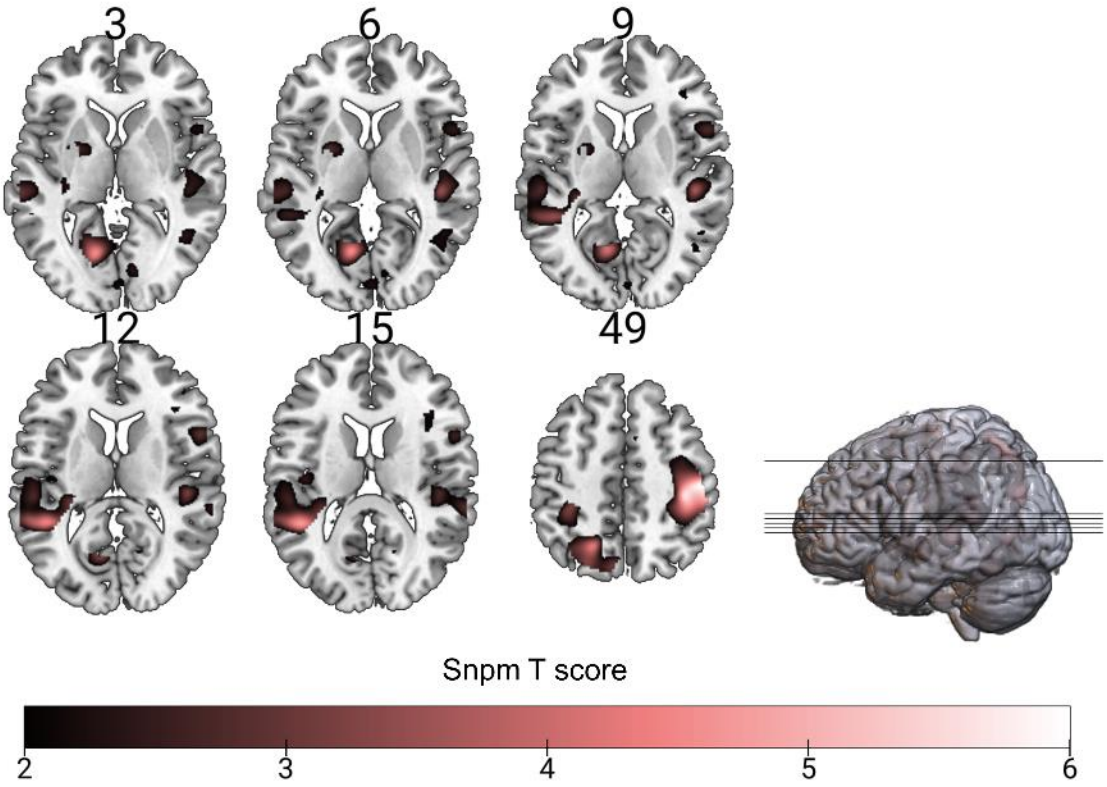

Figure S1. Results of the MVPA searchlight analyses within unambiguous stimuli, projected onto a canonical MNI single-subject brain. Numbers refer to the slice position in z direction.

678 **Table S1: The results of the group-level analysis (random-effects analysis using**
679 **permutation-based nonparametric statistics) of normalized and smoothed individual**
680 **distance maps ($p < .05_{FWE\text{-}corrected}$).**

	Coordinates			
	x	y	z	T score
right postcentral gyrus (PoG)	44	-28	44	6.45
left anterior insula (AI)	-30	20	14	6.35
left ventral motor cortex (vMC)	-38	4	42	5.67
left supplementary motor cortex (SMA)	-6	-8	52	5.22

Coordinates are in MNI space.

681
682