Behavioral/Cognitive

In Spoken Word Recognition, the Future Predicts the Past

Laura Gwilliams,[1,3] Tal Linzen,[4] David Poeppel,[1,5] and Alec Marantz[1,2,3]

[1]Psychology Department, [2]Department of Linguistics, New York University, 10003, [3]New York University Abu Dhabi Research Institute, New York University Abu Dhabi, Saadiyat Island, [4]Department of Cognitive Science, Johns Hopkins University, Baltimore, 21218, and [5]Department of Neuroscience, Max-Planck Institute (MPIEA) 60322 Frankfurt, Germany

Speech is an inherently noisy and ambiguous signal. To fluently derive meaning, a listener must integrate contextual information to guide interpretations of the sensory input. Although many studies have demonstrated the influence of prior context on speech perception, the neural mechanisms supporting the integration of subsequent context remain unknown. Using MEG to record from human auditory cortex, we analyzed responses to spoken words with a varyingly ambiguous onset phoneme, the identity of which is later disambiguated at the lexical uniqueness point. Fifty participants (both male and female) were recruited across two MEG experiments. Our findings suggest that primary auditory cortex is sensitive to phonological ambiguity very early during processing at just 50 ms after onset. Subphonemic detail is preserved in auditory cortex over long timescales and re-evoked at subsequent phoneme positions. Commitments to phonological categories occur in parallel, resolving on the shorter timescale of ~450 ms. These findings provide evidence that future input determines the perception of earlier speech sounds by maintaining sensory features until they can be integrated with top-down lexical information.

Key words: auditory processing; lexical access; MEG; speech

Significance Statement

The perception of a speech sound is determined by its surrounding context in the form of words, sentences, and other speech sounds. Often, such contextual information becomes available later than the sensory input. The present study is the first to unveil how the brain uses this subsequent information to aid speech comprehension. Concretely, we found that the auditory system actively maintains the acoustic signal in auditory cortex while concurrently making guesses about the identity of the words being said. Such a processing strategy allows the content of the message to be accessed quickly while also permitting reanalysis of the acoustic signal to minimize parsing mistakes.

Introduction

Typically, sensory input is consistent with more than one perceptual inference and surrounding context is required to disambiguate. When this ambiguity occurs in a signal that unfolds gradually, the system is presented with a critical trade-off: either prioritize accuracy by accumulating sensory evidence over time or prioritize speed by forming interpretations based on partial information. This trade-off is particularly prevalent in speech, which is rife with noise and ambiguity. Further, because language is hierarchically structured, inference occurs both within and across levels of linguistic description: Comprehension of phonemes (e.g., /p/, /b/) is required to understand words; understanding words aids comprehension of their constituent phonemes. How does the human brain strike a balance between speed and accuracy across these different levels of representation?

When the input is an unambiguous phoneme, low-level spectrotemporal properties are first processed in primary auditory cortex ~50 ms after onset [A1/Heschl's gyrus (HG)]. Then, higher-level phonetic features are processed in superior temporal gyrus (STG) for ~100 ms (Simos et al., 1998; Ackermann et al., 1999; Obleser et al., 2003; Papanicolaou et al., 2003; Obleser et al., 2004; Chang et al., 2010; Mesgarani et al., 2014; Di Liberto et al., 2015). These are thought to be purely bottom-up computations performed on the acoustic signal. In natural language, where the acoustic signal is often consistent with more than one phoneme, the system will need to decide which categorization is the correct one. It is currently unknown where the recognition and resolution of phoneme ambiguity fits relative to this sequence of bottom-up operations.

To cope with phoneme ambiguity in speech, the brain uses neighboring information to disambiguate toward the contextually appropriate interpretation. Most prior research has focused

Received Jan. 10, 2018; revised June 6, 2018; accepted July 9, 2018.

Author contributions: L.G. wrote the first draft of the paper; T.L., D.P., and A.M. edited the paper. L.G. and A.M. designed research; L.G. performed research; L.G. analyzed data; L.G. wrote the paper.

This work was supported by European Research Council (grant ERC-2011-AdG 295810 BOOTPHON) and the Agence Nationale pour la Recherche (grants ANR-10-IDEX-0001-02 PSL and ANR-10-LABX-0087 IEC) to T.L.; the National Institutes of Health (Grant 2R01DC05660 to D.P.); and the NYU Abu Dhabi (NYUAD) Institute (Grant G1001 to A.M.). We thank Kyriaki Neophytou for her help with data collection and Lena Warnke for help with stimulus creation.

The authors declare no competing financial interests.

Correspondence should be addressed to Laura Gwilliams, Linguistics department, New York University, 10 Washington Place, New York, NY, 10003. E-mail: laura.gwilliams@nyu.edu.

DOI:10.1523/JNEUROSCI.0065-18.2018

 0270-6474/18/387585-15$15.00/0

on the use of preceding context, both in terms of the underlying computations and its neural implementation. This work suggests that previous context sets up probabilistic expectations about upcoming information and biases acoustic perception to be consistent with the predicted phonemes (Warren, 1970; Cole, 1973; Samuel, 1981). The left STG and HG appear to be involved in this process and activity in both regions correlates with the extent to which an expectation is violated (Gagnepain et al., 2012; Ettinger et al., 2014; Gwilliams and Marantz, 2015).

Here, we focus on much lesser explored *postdictive* processes, which allow subsequent context to bias perception. This phenomenon has been demonstrated behaviorally (Ganong, 1980; Connine et al., 1991; McQueen, 1991; Samuel, 1991; Gordon et al., 1993; McMurray et al., 2009; Szostak and Pitt, 2013) and has been explained in terms of commitment delay. The system waits to accumulate lexical evidence before settling on an interpretation of a phoneme and maintains subphonemic information until the commitment is made. Precisely how the brain implements subphonemic maintenance and commitment processes is currently unestablished, but previous research has indicated some regions that are likely to be involved. Activity linked to lexical processing in supramarginal gyrus affects phonetic processing in STG at a word's point of disambiguation (POD) (Gow et al., 2008). The STG and HG have also been implicated in fMRI studies of phoneme ambiguity (Blumstein et al., 2005; Myers and Blumstein, 2008; Kilian-Hütten et al., 2011), in perceptual restoration of masked phonemes (Leonard et al., 2016), and with sensitivity to post-assimilation context (Gow and Segawa, 2009).

In this study, we investigate how phoneme perception is influenced by subsequent context by addressing three questions. First, is the system sensitive to phoneme ambiguity during early perceptual processes or during higher-order postperceptual processes? Second, how is subphonemic maintenance and phonological commitment neurally instantiated? Third, what temporal constraints are placed on the system—what is the limit on how late subsequent context can be received and still be optimally integrated?

To address these questions, we recorded whole-head MEG in two experiments. Participants heard phonemes that varied in ambiguity either at the onset of syllables (Experiment 1) or at the onset of words (Experiment 2). This allowed us to address our first aim. In the second experiment, we tested for sensitivity to the ambiguity, acoustics, and two phonetic features of the onset sound (voice onset-time and place of articulation) at each phoneme along the length of the words to test for subphonemic maintenance (question two). The onset phoneme was later disambiguated once the listener could uniquely identify what word was being said. The latency of this "disambiguation point" ranged from ~150 to 700 ms, allowing us to address our third research question.

Materials and Methods

In the first experiment, participants listened to syllables that varied along an 11-step continuum from one phoneme category to another (e.g., /pa/ ↔ /ba/). Participants classified the sounds as one of the two phoneme categories (e.g., P or B). The syllables provide sensory information about onset phoneme identity but no subsequent context. This protocol is fully described in "Experiment 1."

In the second experiment, a different group of participants listened to items from word ↔ non-word continua ("parakeet" ↔ "barakeet"). This second set of stimuli thus provides both sensory evidence about the identity of the onset phoneme and subsequent contextual information. The subsequent information becomes available at the word's POD, which refers to the phoneme that uniquely identifies the word being said and therefore disambiguates the identity of the phoneme at onset. For example, in the word "parakeet," the POD is the final vowel "ee," because at that point no other English lexeme matches the sequence of phonemes. Therefore, at the POD there is sufficient information in the speech signal to uniquely identify the onset phoneme as /p/. The design of Experiment 2 was inspired by McMurray et al. (2009).

The first syllables of the words used in Experiment 2 were exactly the same as those used in Experiment 1; the only difference is that the syllable was followed by silence in the first experiment and the rest of the word in the second experiment. This allowed us to examine neural responses to the same acoustic signal in isolation and in lexical contexts.

Material creation (common to both experiments)

Word pairs were constructed using the English Lexicon Project (ELP) (Balota et al., 2007), which is a database of phonologically transcribed words and their properties. First, phonological transcriptions of all words beginning with the plosive stops *p*, *b*, *t*, *d*, *k*, and *g* were extracted. We selected this set of phonemes because it allowed us to examine responses as a function of two phonetic features. Voice onset time (VOT) refers to the amount of time between the release of the stop consonant and the onset of vocal cord vibration. If the amount of time is longer (>~40 ms), then the sound will be perceived as voiceless (e.g., *t*, *p*, or *k*); if the time is shorter (<~40 ms), then it will be perceived as voiced (e.g., *d*, *b*, or *g*). Place of articulation (PoA) refers to where in the mouth the tongue, teeth, and lips are positioned to produce a speech sound. Differences in PoA manifest as spectral differences in the acoustic signal. By measuring responses as a function of both VOT and PoA, we can examine how ambiguity is resolved when it arises from a temporal cue or from a spectral cue, respectively.

Potential word pairs were identified by grouping items that differed by just one phonetic feature in their onset phoneme. For example, the feature VOT was tested by grouping words with the onset phoneme pairs (*t*-*d*, *p*-*b*, *k*-*g*) and PoA was tested with the onset phoneme pairs (*p*-*t*, *t*-*k*). Word pairs were selected when they shared two to seven phonemes after word onset until the phonological sequence diverged. For example, the word pair parakeet/barricade was selected because it differs in voicing of the onset phoneme (*p*/*b*), shares the following four phonemes (*a*-*r*-*a*-*k*) and then diverges at the final vowel. This procedure yielded 53 word pairs: 31 differed in VOT and 22 differed in PoA. Words ranged in length from 4 to 10 phonemes (M = 6.8; SD = 1.33) and 291–780 ms (M = 528; SD = 97). Latency of disambiguation ranged from 3 to 8 phonemes (M = 5.1; SD = 0.97) and 142–708 ms (M = 351; SD = 92).

A native English speaker was recorded saying the selected 106 words in isolation. The speaker was male, aged 25, with a northeast American accent. He said each of the words in a triplet with consistent intonation (e.g., ↑ parakeet, — parakeet, ↓ parakeet). The middle token was extracted from the triplet, which promoted similar and consistent intonation and pitch across words. We have used a similar strategy in previous studies (Gwilliams and Marantz, 2015; Gwilliams et al., 2015). This extraction was done using Praat software (Boersma and Weenink, 2000).

Each item pair was exported into TANDEM-STRAIGHT for the morphing procedure (Kawahara et al., 2008; Kawahara and Morise, 2011). In short, the morphing works by taking the following steps: (1) position anchor points to mark the onset of each phoneme of the word pair, (2) place weights on each anchor point to determine the percentage contribution of each word at each phoneme, and (3) specify the number of continuum steps to generate. An explanation and tutorial of the software is available at: https://memcauliffe.com/straight_workshop/index.html.

For example, to generate the "barricade" ↔ "parricade", "barakeet" ↔ "parakeet" continua shown in Figure 1, anchor points are first placed at the onset of each phoneme in the recorded words "barricade" and "parakeet," marking the temporal correspondence between the phonemes in the word pair. Next, we decide the amount of morphing to be used at each phoneme to generate the unambiguous words/non-words at the end points of the continua. At the first phoneme, the anchor points are weighted as either 100% "barricade" to generate an unambiguous /b/ at onset, or 100% "parakeet" to generate an unambiguous /p/ at onset. All subsequent phonemes until point of disambiguation ("-arak-") are weighted with equal contributions of each word (50–50). At and after

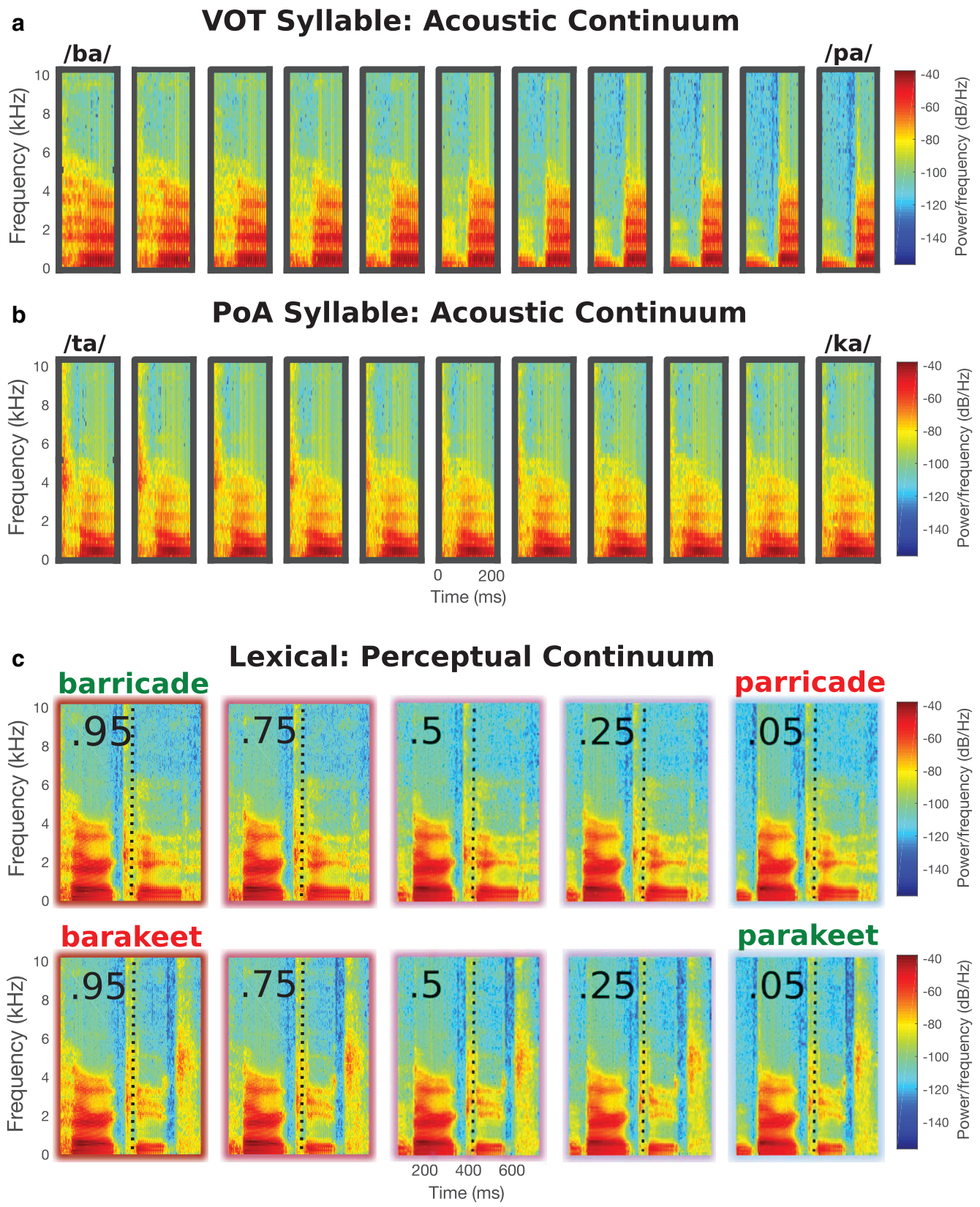

Figure 1. Stimuli examples. ***a***, An example 11-step voice onset time syllable continuum used in Experiment 1. ***b***, An example 11-step place of articulation syllable continuum used in Experiment 1. ***c***, An example five-step perceptually defined continuum pair used in Experiment 2 generated from the words "barricade" and "parakeet" (shown in green). The resultant non-words "parricade" and "barakeet" are shown in red. The point of disambiguation is represented with a dashed line.

disambiguation, anchors are again weighted at 100% either toward "parakeet" for the "parakeet–barakeet" continuum, or toward "barricade" for the "parricade–barricade" continuum.

In general, for each pair, all anchor points before the POD are placed at the 50% position and the first anchor point is positioned either in the congruent position, creating a word ("parakeet") or the incongruent (competitor) position, creating a non-word ("barakeet"). This ensures that apart from the first phoneme, the acoustic signal remains identical across the two word pairs until the disambiguation point. Eleven continua steps were created for each continuum.

The resulting 1166 auditory files were analyzed using the Penn Forced Aligner (p2fa) (Yuan and Liberman, 2008) to extract the timing of each phoneme's onset and offset along the length of the word. This created a set of annotation files, which were then visually inspected using Praat (Boersma and Weenink, 2000). The accuracy of the p2fa aligner was good overall, but a few manual adjustments were made on ~10% of the auditory files to ensure correct timing.

Experiment 1

Participants. Twenty-four right handed native English participants took part in the study (11 female; age: M = 25.44, SD = 8.44). This sample size was selected based on previous studies using the same MEG machine (Gwilliams and Marantz, 2015; Gwilliams et al., 2016; Gwilliams and Marantz, 2018). They were recruited from the New York University Abu Dhabi community and were compensated for their time. All had normal or corrected vision, normal hearing, and no history of neurological disorders.

Stimuli. From the word ↔ nonword continua described in "Material creation (common to both experiments)," we extracted just the first syllable (consonant–vowel sequence). This was done for each of the 1166 items. We then amplitude-normed the extracted files to 70 dB.

Experimental design. The syllable stimuli were separated into 11 blocks. Each block consisted of two items from each continuum, with the constraint that each item had to be at least three morphed steps away from its paired counterpart. This resulted in a total of 106 trials per block and 1166 trials total. The assignment of stimulus to block was different for each of the 24 participants and was balanced using a Latin-square design. Item order was randomized within each block.

Participants heard each syllable in turn, and had to categorize the sound as one of two options that were displayed on the screen. While participants completed the categorization task, whole-head MEG was being recorded. The screen was ~85 cm away from the participant's face while they lay in a supine position.

The experimental protocol was as follows. First, a fixation cross was presented for 1000 ms. Then, the two options appeared in upper case, flanking the fixation (e.g., "B + P"). The syllable was played 500 ms later and the participant needed to indicate which of the two options best matched the syllable they heard by responding with a button box. The options remained onscreen until a response was made. There was no limit placed on how soon participants needed to respond. At each block interval, participants had a self-terminated break. The background was always gray (RGB: 150, 150, 150). All text was in white (RGB: 0, 0, 0), size 70 Courier font. The experiment was run using Presentation software (Version 18.0; Neurobehavioral Systems, www.neurobs.com). The recording session lasted ~50 min.

Experiment 2

Participants. Twenty-five right-handed native English participants took part in the study (15 female; age: M = 24.84, SD = 7.3). Six had taken part in Experiment 1 two months earlier. All had normal or corrected vision, normal hearing, no history of neurological disorders, and were recruited from the New York University Abu Dhabi (NYUAD) community.

Stimuli. In the second study, we used items from the full word ↔ nonword continua. For these items, we wanted to make the onset phonemes across the words differ along a perceptually defined continuum rather than the 11-step acoustically defined continuum used in Experiment 1. In other words, in absence of lexical context, we wanted to make sure the phoneme would be picked out of the pair a 0.05, 0.25, 0.5, 0.75, and 0.95 proportion of the time. To set up the materials in this way, we averaged the psychometric functions over subjects for each 11-step continuum used in Experiment 1 and selected the five steps on the continuum that were closest to the desired selection proportions (Fig. 2). This converted the continuum from being defined along 11 acoustically defined steps to being defined along five perceptually defined steps. Continua were removed if the unambiguous endpoints of the continuum were not categorized with at least 80% accuracy for all subjects or if the position of the ambiguous token was not at least three points away from either endpoint of the continuum. This resulted in 49 remaining word pairs and 490 trials total. These words were amplitude normed to 70 dB.

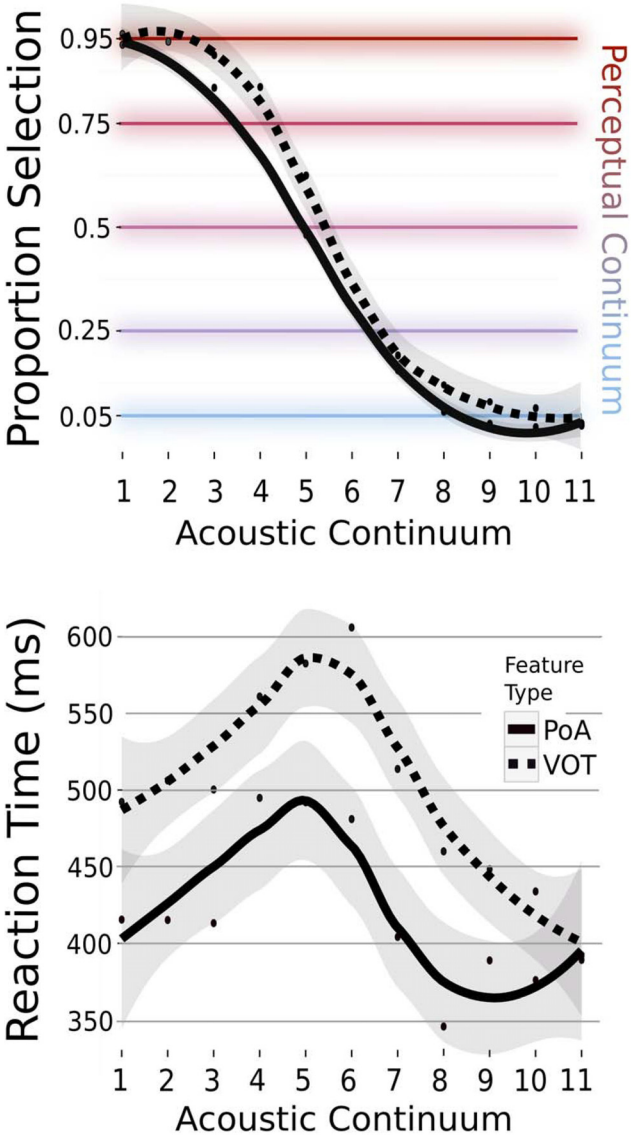

Figure 2. Behavioral results for Experiment 1. Top, Behavioral psychometric function of phoneme selection as a function of the 11-step acoustic continuum. PoA and VOT continua are plotted separately. The colored horizontal lines correspond to the five behavioral classification positions used to define the perceptual continuum used in Experiment 2. Bottom, Reaction times as a function of the 11-step continuum. Note the slow down for ambiguous tokens and slower responses to items on the VOT continuum compared with the PoA continuum.

Experimental design. Participants performed an auditory-to-visual word-matching task on two of five auditory items. They were not required to explicitly make judgements about the identity of the onset phoneme. The visual word was either the same as the auditory word (e.g., parakeet–parakeet would require a "match" response) or it was the other word of the pair (e.g., parakeet–barricade would require a "mismatch" response). One item of each 5-step continuum was made into a "match"

trial (1/5) and one other was a "mismatch" trial (1/5). These conditions were pseudorandomly assigned using a Latin-square procedure. The experiment was split into five blocks and only one token from each continuum appeared in each block. The assignment of item to block was also pseudorandomized in a Latin-square fashion. This resulted in 25 unique experimental orders, across which items were matched for block order and match–mismatch assignment.

The experimental protocol was as follows. First, a fixation cross was displayed for 500 ms. Then, while the fixation was still on the screen, the auditory word was presented. If it was a task trial, the visual word appeared 500 ms after the auditory word offset and remained on screen until participants made a match (left button) or mismatch (right button) decision with their left hand. If it was a no-task trial (3/5 of trials), a blank screen was presented and participants could move to the next trial by pressing either button. The recording lasted ~40 min. The apparatus and experiment presentation software were the same as those used in Experiment 1.

Data processing (common to both experiments)

All participants' head shapes were digitized using a hand-held FastSCAN laser scanner (Polhemus) to allow for coregistration during data preprocessing. Five points on each participant's head were also digitized: just anterior of the left and right auditory canal and three points on the forehead. Marker coils were later placed at the same five positions to localize each participant's skull relative to the sensors. These marker measurements were recorded just before and after the experiment to track the degree of movement during the recording.

Stimuli were presented binaurally to participants though tube earphones (Aero Technologies). MEG data were recorded continuously using a 208-channel axial gradiometer system (Kanazawa Institute of Technology) with a sampling rate of 1000 Hz and applying an online low-pass filter of 200 Hz.

MEG data from the two experiments underwent the same preprocessing steps. First, the continuous recording was noise reduced using the continuously adjusted least squares method (CALM) (Adachi et al., 2001) with MEG160 software (Yokohawa Electric and Eagle Technology). The noise-reduced data, digital scan and fiducials, and marker measurements were exported into MNE-Python (Gramfort et al., 2014). Bad channels were removed through visual inspection. Independent component analysis (ICA) was computed over the noise-reduced data using FastICA in MNE-Python. Components were removed from the raw recording if they contained ocular or cardiac artifacts, which were identified based on the topography of magnetic activity and time course response. The data were then epoched from 500 ms presyllable onset to 1000 ms postsyllable onset for Experiment 1 and 500 ms prephoneme onset to 1000 ms postphoneme onset for every phoneme in Experiment 2. How we determined the timing of each phoneme is described in the last paragraph of the "Material creation (common to both experiments)" section. Any trials in which amplitude exceeded a $\pm$ 2000 fT absolute or peak-to-peak threshold were removed. Baseline correction was applied to the epoch using the 200 ms preceding syllable/word onset.

To perform source localization, the location of the subject's head was coregistered with respect to the sensor array in the MEG helmet. For subjects with anatomical MRI scans ($n = 4$), this involved rotating and translating the digital scan to minimize the distance between the fiducial points of the MRI and the head scan. For participants without anatomical scans, the FreeSurfer "fsaverage" brain was used, which involved first rotation and translation and then scaling the average brain to match the size of the head scan.

Next, a source space was created consisting of 2562 potential electrical sources per hemisphere. At each source, activity was computed for the forward solution with the boundary element model method, which provides an estimate of each MEG sensor's magnetic field in response to a current dipole at that source. The inverse solution was computed from the forward solution and the grand average activity across all trials. Data were converted into noise-normalized dynamic statistical parameter map (dSPM) units (Dale et al., 2000) using an SNR value of 2. The inverse solution was applied to each trial at every source for each millisecond defined in the epoch using a fixed orientation of the dipole current that estimates the source normal to the cortical surface and retains dipole orientation.

Statistical analysis (common to both experiments)

All results reported here are based on mass univariate analyses. We focused on the following four orthogonal variables. First, "acoustics" refers to the item's position along the 11-step acoustic continuum for Experiment 1 and the five-step perceptual continuum for Experiment 2. Second, "ambiguity" refers to the absolute distance (measured in continuum steps) from the perceptual boundary between phonological categories. Here, we define the perceptual boundary as the position on the continuum where, on average, participants were equally likely to classify the phoneme as one category or the other. Third, "VOT" refers to whether the phoneme was behaviorally classified as voiced (b, d, g) or voiceless (p, t, k). Finally, "PoA" refers to whether the phoneme was behaviorally classified as being articulated as a bilabial (b, p), labiodental (t, d), or velar stop (k, g). We also included "feature type," which refers to whether the phonetic feature being manipulated along the continuum is PoA or VOT.

Sensitivity to the four stimulus variables was tested at different moments in the time course of the MEG data across the two experiments. At the onset of the syllables (see: "Experiment 1: syllable onset" section), at the onset of the words (see "Experiment 2: word onset" section), at the onset of the disambiguation point in the words (see "Experiment 2: POD onset" section), and at the onset of phonemes in the middle of the word and after disambiguation (see "Experiment 2: each phoneme onset" section).

Results

Behavioral

To analyze behavioral responses in Experiment 1, we applied a mixed-effects regression analysis using the *lme4* package (Bates et al., 2014) in *R* (R Core Team, 2014). We included the above four variables as fixed effects and by-subject slopes, as well as feature type, the interaction between feature type and ambiguity, and feature type with acoustics. The same model structure was used to fit the reaction time data and the selection data. To assess the significance of each variable, we removed each variable in turn as a fixed effect (but keeping it as a by-subject slope) and compared the fit of that model with the fit of the full model.

For reaction time, we observed a significant effect of ambiguity such that responses were significantly slower for more ambiguous items ($\chi^2 = 141.57$, $p < 0.001$). The effect of acoustics was not significant ($\chi^2 = 3.32$, $p = 0.068$). There was a significant effect of feature type such that responses were significantly slower for VOT continua than PoA continua ($\chi^2 = 99.98$, $p < 0.001$). Ambiguity and feature type revealed a significant interaction ($\chi^2 = 8.93$, $p = 0.002$). There was no interaction between feature type and acoustics.

A logistic regression was applied to behavioral selection with the same model structure and model comparison technique. Acoustics was a significant predictor ($\chi^2 = 623.26$, $p < 0.001$), as well as feature type ($\chi^2 = 21.53$, $p < 0.001$). The effect of ambiguity was not significant ($\chi^2 = 0.68$, $p = 0.41$). Neither was the interaction between feature type and ambiguity ($\chi^2 = 2.5$, $p = 0.11$) or feature type and acoustics ($\chi^2 = 2.38$, $p = 0.12$). See Figure 2 for a summary of the behavioral results.

Overall, the behavioral analysis indicates that the stimuli are being perceived as intended: we observed a typical psychometric function and a slowdown in responses for more ambiguous items.

Neural

To investigate the underlying neural correlates of retroactive perception, we ran a spatiotemporal permutation cluster analysis

over localized source estimates of the MEG data (Holmes et al., 1996; Maris and Oostenveld, 2007). This was applied across the HG and STG bilaterally, searching a time window of 0–200 ms after phoneme onset (corresponding to syllable onset, word onset, or POD onset). We implemented the test by running a multiple regression independently at each specified source and time point. Spatiotemporal clusters were formed for each variable based on adjacent β coefficients over space and time. In all analyses, we used a cluster forming threshold of $p < 0.05$ with a minimum of 10 neighboring spatial samples and 25 temporal samples. See Gwilliams et al. (2016) for more details concerning this analysis technique.

In the multiple regression, we simultaneously included the four variables described above: acoustics, ambiguity, VOT, and PoA. Trials were grouped into phoneme categories based on participants' average behavioral responses in Experiment 1. Trial number and block number were included in all models as nuisance variables. The same analysis was conducted for both Experiment 1 and 2.

Experiment 1: syllable onset

In terms of main effects, there was a significant effect of ambiguity, which formed two significant clusters: one in left HG (45–100 ms, $p < 0.005$) and one in the right STG (105–145 ms, $p = 0.029$). Acoustics formed a cluster in right HG, but it was not significant in the permutation test (40–75 ms, $p = 0.125$). VOT significantly modulated responses in right STG (85–200 ms, $p < 0.001$) and PoA in left STG (90–150 ms, $p < 0.001$).

The results for Experiment 1 are displayed in Figure 3, *A* and *B*.

Experiment 1: acoustic analysis. We were surprised to observe such early sensitivity to phonological ambiguity. Because we observe the effect at 50 ms after onset (Fig. 3*A*), it must reflect a response to not substantially more than the first 20 ms of the acoustic signal; for example, just the noise burst of the voiceless items and the initial voicing of the voiced items (based on timing estimates from the peak of evoked activity). This is because the latency between the onset of an acoustic stimulus and the first spike in primary auditory cortex (A1) can be as late as 30 ms, depending upon acoustic amplitude (for a review, see Heil, 2004). Also see Steinschneider et al. (1995) for a similar estimate for the latency response to syllables in primate A1. Therefore, a conservative estimate of how much acoustic signal has reached HG to drive the early ambiguity effect is ~20 ms.

To assess what information is available to the primary auditory cortex at this latency, we decomposed the first 20 ms of each stimulus into its frequency power spectra using fast Fourier transform (FFT). A spectral decomposition was chosen because we wanted to mirror the spectral–temporal input received by primary auditory cortex. Power at each frequency band from 0–10 kHz for all stimuli except the fully ambiguous items (leaving four continua steps) was used to train a logistic regression classifier to decode the phonological category (Fig. 4*A*). Accuracy was significantly above chance level, as determined by 1000 random permutations of phoneme labels ($p < 0.001$). The phoneme category could be decoded from each of the four continua steps, but accuracy of classification decreased as ambiguity increased (Fig. 4*B*). Importantly, continua steps themselves could not be decoded from this signal (Fig. 4*C*), suggesting that this early response indeed scales with distance from the perceptual boundary and not acoustic properties per se. This suggests that the early ambiguity effect that we observed in HG is not driven by, for example, an acoustic artifact generated during the stimuli morphing procedure.

To pursue the stimulus decoding analysis further, we applied the same logistic regression classifier to the first 60 ms of acoustic input, the likely amount of information driving the N100 m response (see the introduction). Because the N100m is thought to reflect processing of an auditory stimulus through a cortical-thalamic loop and because our estimate is that it takes 20–30 ms for the sound stimulus to reach primary auditory cortex, a reasonable estimate for the maximum auditory signal duration driving the M100 is ~80 ms. To be conservative, we chose to analyze the first 60 ms of acoustic input.

The classifier was trained either on a single 60 ms spectral segment of the signal or three sequential 20 ms spectral chunks. The former provides reasonable spectral resolution but poor temporal resolution; the latter provides the opposite. This novel analysis revealed intuitive results: the classifier more accurately distinguished VOT contrasts (a temporal cue) when trained on three 20 ms chunks and PoA contrasts (a spectral cue) when trained on a single 60 ms chunk. It may be the case that the N100m response is driven by neuronal populations that sample both at fast (~20 ms) and slower (~60 ms) frequencies to accurately identity phonemes that vary across each phonetic dimension. This analysis also provides additional support that the early timing of the ambiguity effect is not the result of an acoustic artifact, but rather reflects a valid neural response to phoneme ambiguity.

Experiment 2: word onset

In analyzing the results of Experiment 2, we were primarily interested in responses time-locked to two positions in the word. First we will present the results time-locked to word onset, which is also the onset of the phoneme that varies in ambiguity. The analysis was the same as applied for Experiment 1: spatiotemporal cluster test using multiple regression.

In terms of main effects: ambiguity formed two clusters in left HG, one significant and one not significant (150–182 ms, $p = 0.034$; 144–172 ms, $p = 0.063$). Acoustics elicited sensitivity in right HG (106–152 ms, $p = 0.019$). Sensitivity to VOT was found in right STG (92–138 ms, $p < 0.005$); sensitivity to PoA formed two clusters in left STG (86–126 ms, $p < 0.005$; 88–126 ms, $p = 0.028$).

The lateralization of effects observed in Experiment 1 was replicated: sensitivity to ambiguity and PoA in the left hemisphere and to acoustics and VOT in the right hemisphere. Because we did not apply a statistical analysis explicitly to test for the lateralization of these effects (as this was not an aim of the study), we do not want to make claims in this regard. This being said, it is plausible, based on previous studies, that the left hemisphere is more tuned to linguistically relevant features of the acoustic signal (i.e., proximity to the phonological boundary) and the right hemisphere performs more domain-general computations on the onset of an acoustic stimulus, thus tracking lower-level properties (Gage et al., 1998). Further research would need to be conducted to fully understand the hemispheric specialization associated with these effects.

The ambiguity cluster was identified at ~150 ms in the lexical context, which is later than the effect found for syllable context. However, when looking at the cluster level *t*-values across time (Fig. 5, top left), there was also a clear peak in sensitivity to ambiguity at ~50 ms. To determine whether lexical items also elicit early sensitivity to ambiguity, we ran a *post hoc* mixed-effects regression analysis averaging just in left HG (we used the whole parcellated region, not just the specific sources identified in Experiment 1) at 50 ms after word onset (the peak of the effect in

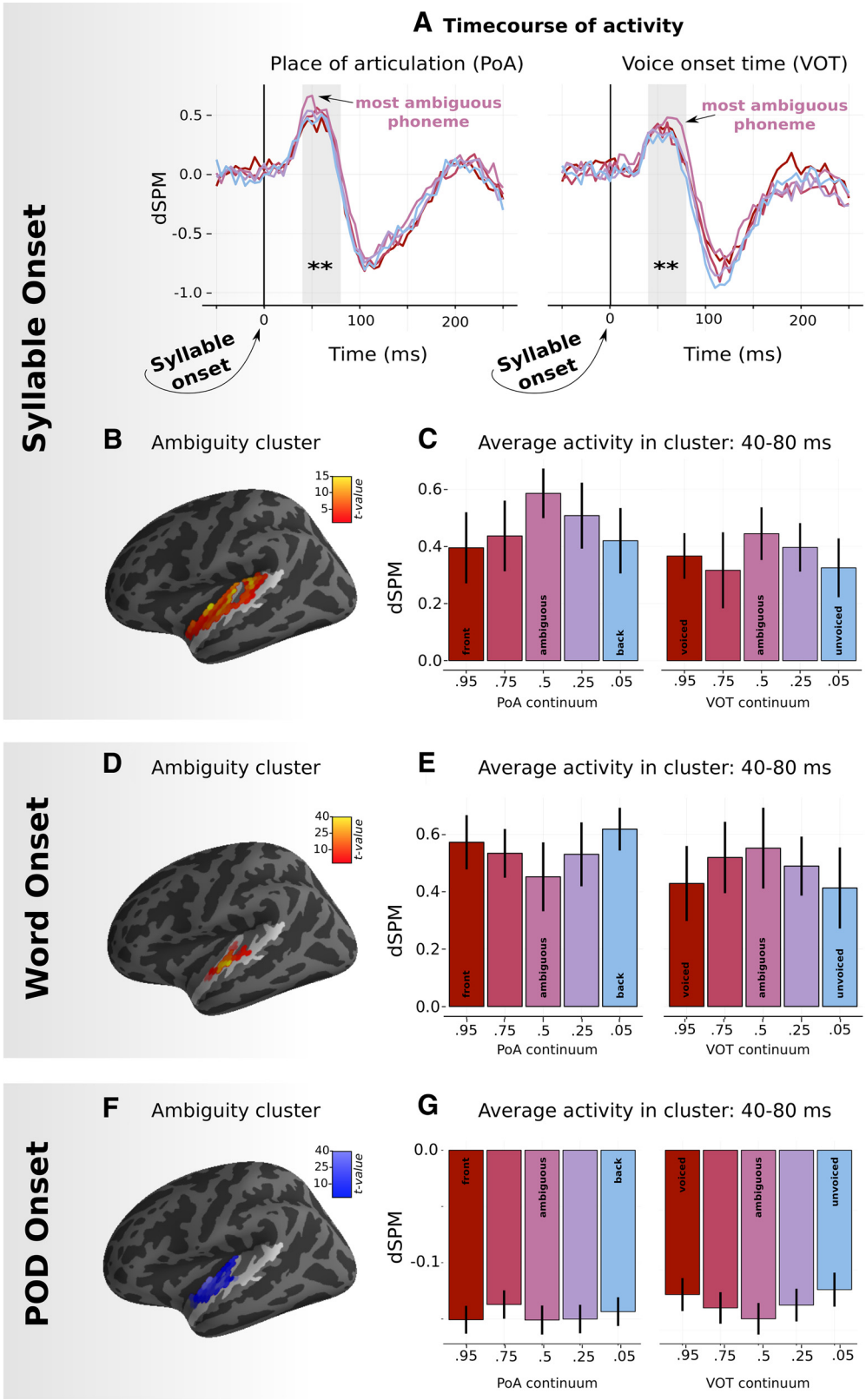

Figure 3. Early responses to ambiguity in left HG (LHG) across the two experiments. ***A***, Experiment 1: Time course of responses for each ambiguity level averaged over significant sources in LHG plotted separately for PoA and VOT continua. ***B***, Experiment 1: Location of sources found to be sensitive to ambiguity in the spatiotemporal cluster test time-locked to syllable onset. Light-shaded region of cortex represents the search volume (HG and STG). Average *t*-value over time is plotted on individual vertices. ***C***, Experiment 1: Averaged responses in significant sources in LHG over the p50m peak time-locked to syllable onset from 40 to 80 ms. Note that, for the p–t continuum, /p/ is "front" and /t/ is "back." For the t-k continuum, /t/ is "front" and /k/ is "back." ***D***, Experiment 2: Location of sources found to be sensitive to ambiguity in the spatiotemporal cluster test time-locked to word onset. ***E***, Experiment 2: Responses time-locked to word onset averaged from 40 to 80 ms over significant sources. ***F***, Experiment 2: Location of sources found to be sensitive to ambiguity in the spatiotemporal cluster test time-locked to POD onset. ***G***, Experiment 2: Response time-locked to POD onset averaged from 40 to 80 ms in significant sources. dSPM refers to a noise-normalized estimate of neural activity. $^{**}p < .01$.

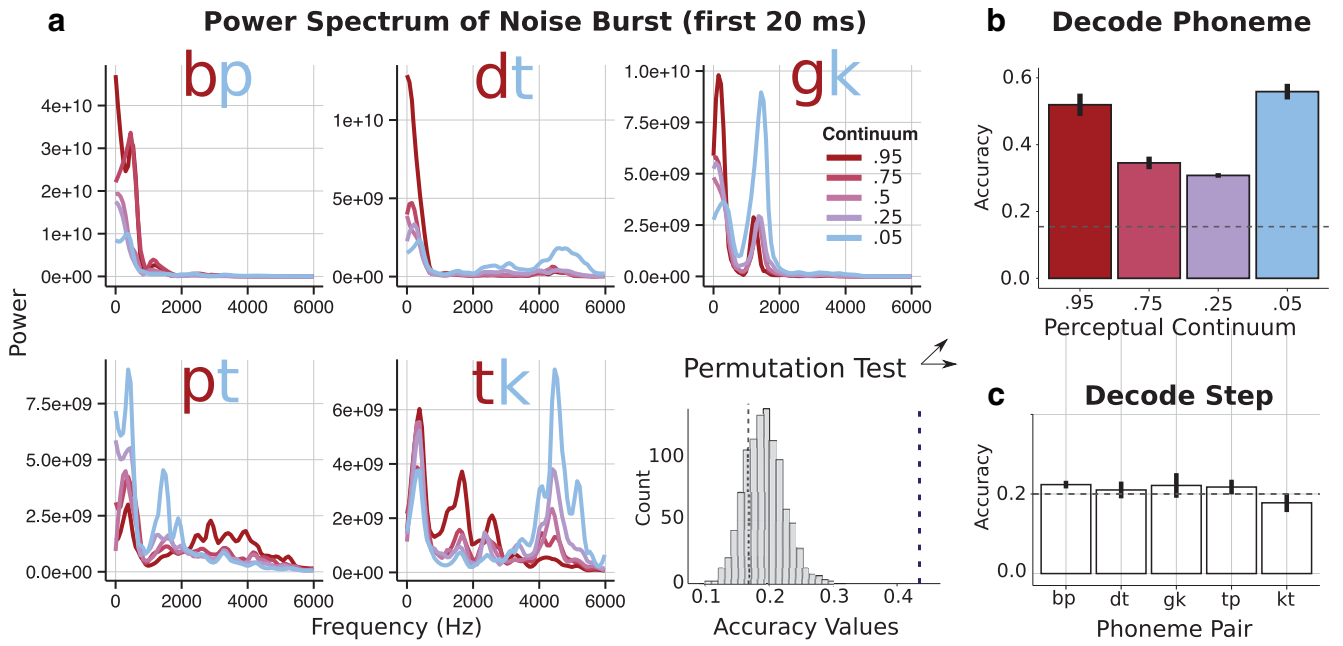

Figure 4. Decoding analysis on acoustic stimuli. ***a***, FFT decomposition of first 20 ms of the auditory stimuli plotted for each phoneme continuum. The histogram represents the 1000 permutations used to determine the significance of classification accuracy. ***b***, Accuracy of the logistic regression classifier in identifying the correct phoneme based on leave-one-out cross validation. Accuracy drops off for more ambiguous tokens. ***c***, Chance-level accuracy in classifying steps along the continuum.

Experiment 1). Ambiguity, acoustics, feature type, and their interaction were coded as fixed effects and by-subject slopes. This revealed a significant interaction between ambiguity and feature type ($\chi^2 = 5.9$, $p = 0.015$) and a significant effect of feature type ($\chi^2 = 13.14$, $p < 0.001$). When breaking the results down at each level of feature type, ambiguity was a significant factor for PoA contrasts ($\chi^2 = 4.84$, $p = 0.027$) and was trending for VOT contrasts ($\chi^2 = 3.09$, $p = 0.078$). This analysis confirms that the early ambiguity effect is replicated in lexical contexts, albeit with weaker responses. Interestingly, the direction of the effect was reversed for PoA contrasts, whereby more ambiguous tokens elicited less rather than more activity (Fig. 3*C*). This interaction may be due to differences in the task or due to processing syllables versus words but, again, more research would need to be conducted to piece these apart.

Experiment 2: POD onset

Next we ran the same analysis time-locked to the onset of the word's POD. This is the phoneme that uniquely identifies what word is being said and therefore also disambiguates the identity of the phoneme at onset. We used the same analysis technique used to assess responses at word onset.

In terms of main effects: ambiguity modulated early responses in left HG (50–84 ms, $p = 0.011$); acoustics modulated later responses in left HG (110–136 ms, $p = 0.043$). Sensitivity to VOT was found in right STG (98–140 ms, $p < 0.01$); sensitivity to PoA was found in left STG (26–96 ms, $p < 0.001$).

In sum, sensitivity to ambiguity, acoustics, PoA, and VOT of the onset phoneme is also present at point of disambiguation, with similar lateralization to that observed at onset. We can also see from the condition averages shown in Figure 5 that the overall pattern of responses is the same at word onset and POD, with a reverse of polarity. This reverse in polarity reflects a reversal in the underlying activity in auditory cortex, not a reversal in sensitivity per se.

Experiment 2: each phoneme onset

Next, we wanted to assess whether the reemergence of sensitivity to the features of the onset phoneme at POD is specific to disambiguation point or if it also reflects a general reactivation process that could be observed at other positions in the word. To test this, we analyzed responses time-locked to the first through seventh phonemes along the length of the word, the disambiguation point, as well as the first two phonemes after the disambiguation point (Fig. 6).

Spatiotemporal clustering was not the ideal analysis technique to use to test this hypothesis because statistical strength cannot be assessed if a spatiotemporal cluster is not formed, making it difficult to draw systematic comparisons about the modulation of an effect over time. Therefore, we instead applied the same multiple regression analysis reported above, but simply averaged activity over left or right auditory cortex and averaged activity within a set of temporal windows. This provided, for each trial, an average measure of neural activity for each hemisphere (2) for each time window we tested (4) for each phoneme position (10). We corrected for multiple comparisons over these 80 tests using Bonferroni correction. Because the analysis applied here is more conservative than the spatiotemporal test, we can expect some differences in the results reported above for word onset and POD.

The regression was fit to source estimates averaged over just HG for ambiguity and acoustics and averaged over both STG and HG in the analysis of PoA and VOT because this is where sensitivity to these variables was observed in the responses to syllable onset and word onset. The ambiguous items were not included in the PoA and VOT analyses because their category is ill posed by definition.

The results of the analysis are presented in Figure 6, showing the *t*-values and corresponding Bonferroni-corrected *p*-values for each multiple regression that was applied at each phoneme, time window, and region. For reference, the analysis picks up on

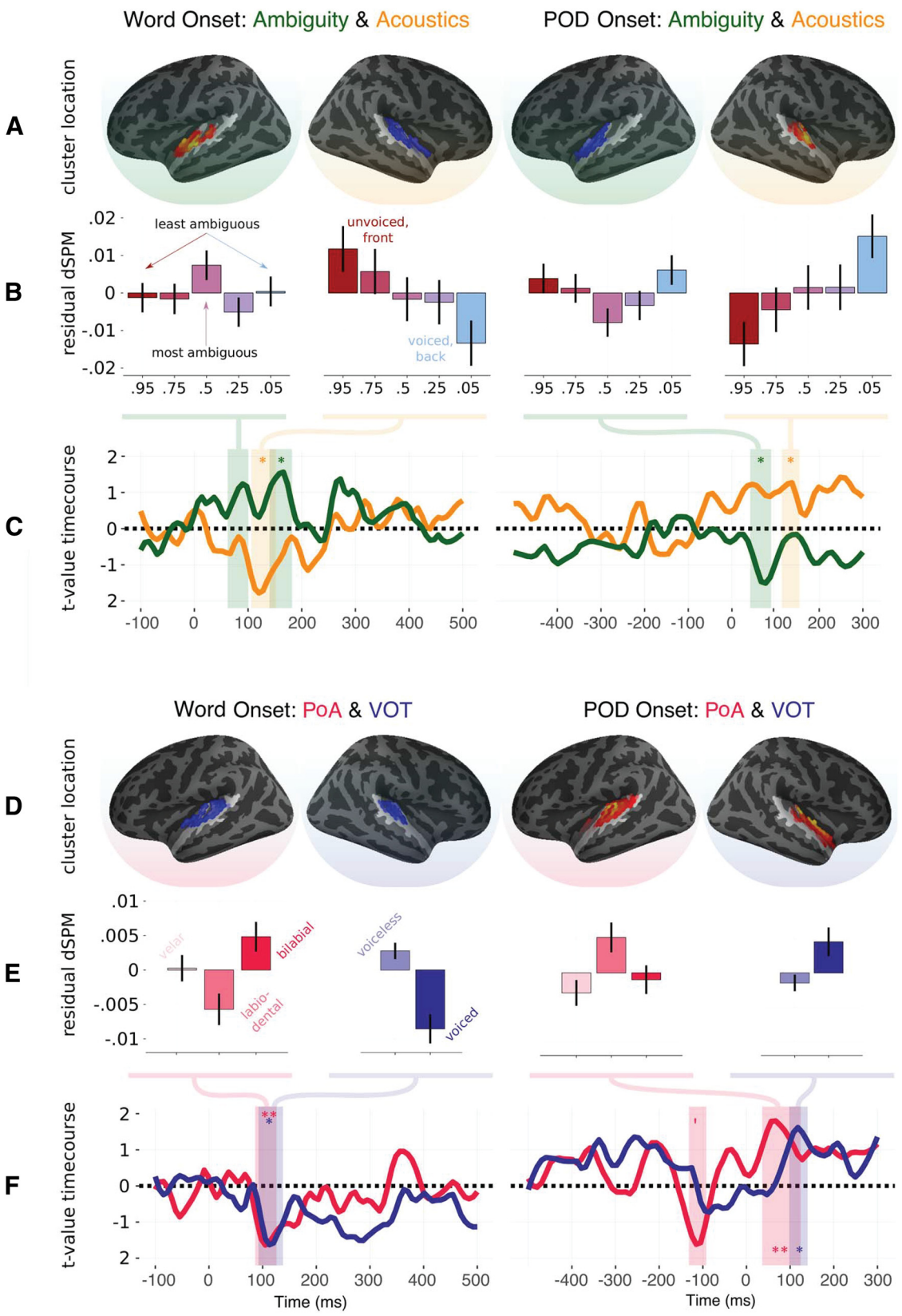

Figure 5. Time course of regression analysis for the four primary variables of interest for Experiment 2 time-locked to word onset (left column) and point of disambiguation (right column). ***A***, Location of the most significant cluster for ambiguity (green) and acoustics (orange) derived from the spatiotemporal cluster test. ***B***, Activity for each step on the continuum, averaged over the spatio-temporal extent of the cluster, after regressing out the other variables in the model: Plotting ambiguity effect after regressing out acoustics and feature type; plotting acoustic effect after regressing out ambiguity and feature type. ***C***, Mean *t*-values averaged in the corresponding cluster for ambiguity and acoustics when put into the same regression model. Note that because the cluster is formed based on the sum of adjacent *t*-values that may be either above 1.96 or below −1.96, the mean value over sources is not directly interpretable as "*t* above 1.96 = $p < 0.05$." ***D***, Location of the most significant cluster for PoA (pink) and VOT (blue). ***E***, Activity averaged for each level of the phonetic features when regressing out the other phonetic feature; for example, regressing out the effect of VOT and then plotting residual activity along the PoA dimension and vice versa. ***F***, Mean *t*-values averaged in the corresponding cluster for PoA and VOT when put into the same regression model. *$p < .05$, **$p < .01$.

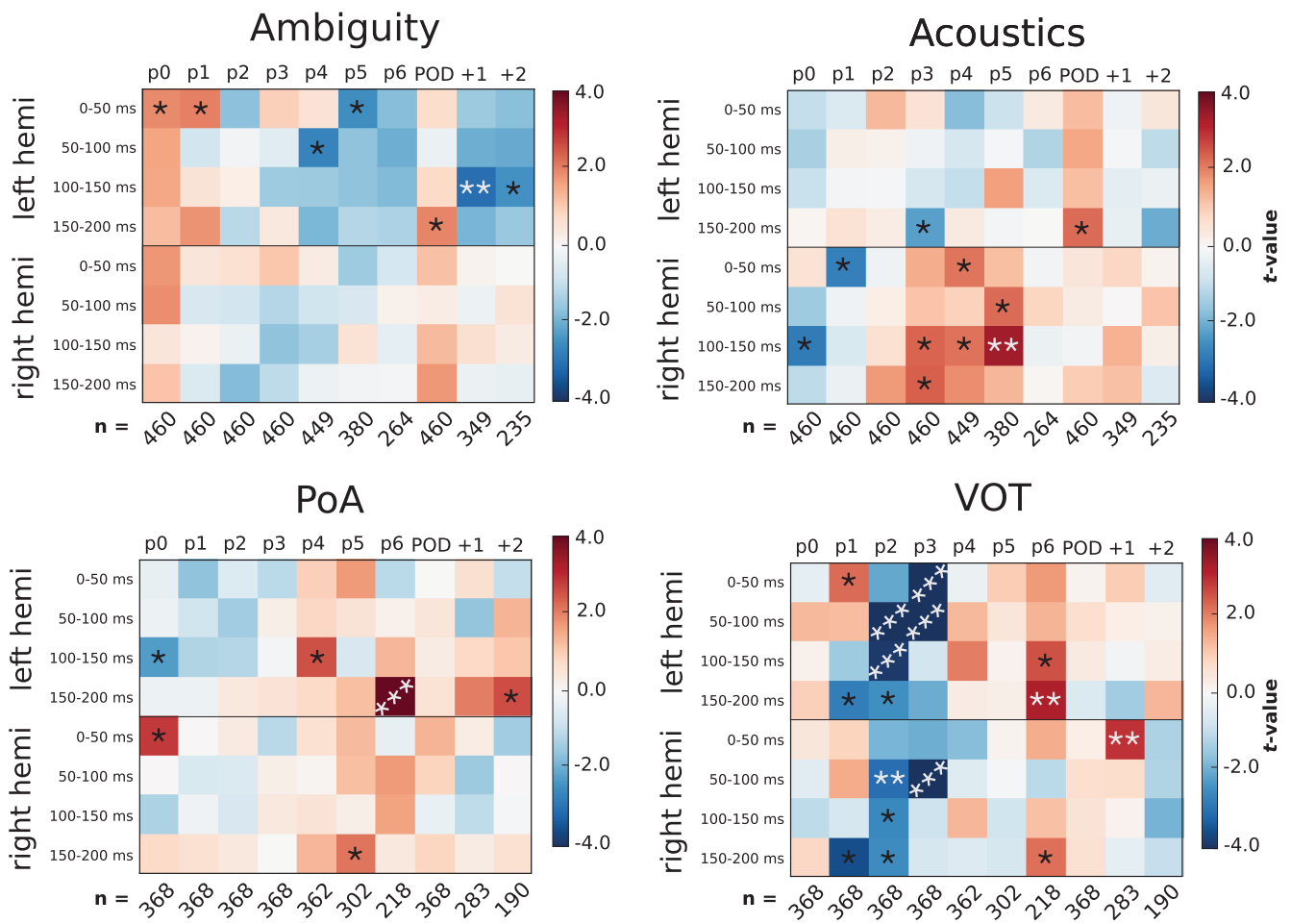

Figure 6. Results of multiple regression applied at each phoneme of the words presented in Experiment 2. Analysis was applied to average source estimates in auditory cortex at different time windows. For ambiguity and acoustics, activity was averaged over left or right HG (the results for both hemispheres are shown). For PoA and VOT, activity was averaged over left or right STG and HG. The plotted values represent the *t*-value associated with how much the regressor modulates activity in the averaged region and time window. The analysis was applied separately at the onset of a number of phonemes within the words: p0 = word onset; POD = point of disambiguation; +1 = one phoneme after disambiguation point. Bonferroni-corrected *p*-values are shown for reference: $*p < 0.05$; $**p < 0.01$; $***p < 0.001$. Average number of trials per subject is shown below the *x*-axes because the number of trials entered into the analysis decreases at longer phoneme latencies.

patterns of activity like those displayed in the condition averages in Figure 5. The results show that the reemergence of sensitivity to each of these variables is not just observed at PoD, but also at intermediate positions along the length of the word. There is not a clear relationship between the strength of the reactivation and the phoneme position; for example, the effects do not get systematically weaker with distance from word onset. Also note that fewer trials are entered into the analysis at longer phoneme latencies because of the varying length of our words. The average number of trials per subject can be seen along the *x*-axes of Figure 6.

There are also some differences depending on the feature being analyzed: VOT has a particularly strong reactivation at the third and fourth phonemes, PoA and VOT seem to be reactivated bilaterally, whereas ambiguity remains left lateralized and acoustics remains primarily right lateralized. These are interesting differences that will require further investigation.

It is worth highlighting that, although our results are indicative of hemispheric specialization, we did not explicitly test for an interaction with lateralization. Because of this, we do not want to make claims about the lateralization of our effects. We leave it to future studies to test the extent to which these processes are bilateral or specific to a particular hemisphere.

Experiment 2: phonological commitment

To determine whether the system commits to a phonological category when disambiguation occurs "too late," we tested for an interaction between disambiguation latency and whether the word resolves to the more or less likely word of the pair given acoustics at onset. The rationale is that, if the system commits to a /b/, for example, but then the word resolves to a p-onset word, more effort is required to comprehend the lexical item that was thrown away during the commitment process. However, if no commitment has occurred, there should be a minimal difference between word and non-word resolution because both the cohort of p-onset and b-onset words are still active.

The analysis had two parts. First, we analyzed responses at the disambiguation point across all trials, locating where and when there was an interaction between word/non-word resolution and POD latency as defined continuously in terms of milliseconds. The analysis was applied over the time window of 0–300 ms after point of disambiguation and over an ROI that included STG and middle temporal gyrus in the left hemisphere. An interaction was found between 196 and 266 ms after POD ($p = 0.02$; Fig. 7*A*). Activity was then averaged in this localized region and time window. Second, we used this average activity to test for an interaction between word/non-word resolution and a discretized

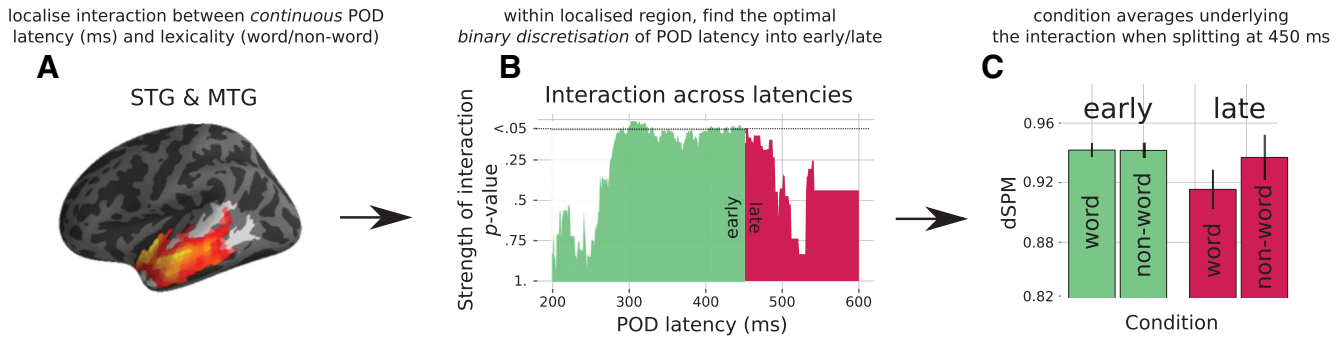

Figure 7. Testing for phonological commitment: analysis pipeline. ***A***, Location of cluster sensitive to the interaction between lexical resolution (word v. non-word) and continuous latency of POD as defined in milliseconds. ***B***, Time course of interaction between lexical resolution and "early" versus "late" disambiguation. "Early" is defined as at or before the increment from word onset shown on the *x*-axis; "late" is defined as after the millisecond on the *x*-axis. The split from green to red shows the final position that the interaction is still significant (450 ms). ***C***, Condition averages for the early/late word and non-words at POD. A significant interaction can be seen when splitting responses at 450 ms.

definition of POD latency coded as either 0 (early) or 1 (late). We perform this discretization 401 times from 200 to 600 ms, iteratively adjusting which trials are considered as having early and late disambiguation. The period of 200–600 ms was chosen because 95% of the words in the study had a POD latency within this time range. First, all words with a POD at 200 ms or earlier were coded as "early"; all those with a POD at 201 ms or later were coded as "late"; then, all words with a POD at 201 ms or earlier were coded as "early," all those with a POD at 202 ms or later were coded as "late," and so on. When testing for this interaction systematically at different latencies, we are able to map out the temporal trajectory of when the system begins to show costs associated with non-word resolution, which is suggestive of commitment processes (Fig. 7*B*).

As can be seen in the trajectory shown in Figure 7*B*, the interaction was maximized when setting the boundary between "early" and "late" between 292 and 447 ms. The effect drops off after 447 ms, suggesting that, at longer latencies, we begin incorrectly grouping together trials where commitment has not yet occurred (with POD earlier than 450 ms) with trials where commitment has occurred (POD later than 450 ms), thus resulting in a weaker interaction.

The direction of the interaction is shown in Figure 7*C*. Words and non-words elicited indistinguishable responses when disambiguation came early; however, there was a significant difference in response amplitude when disambiguation came late. It is hard to interpret the direction of this response because the overall amplitude of the MEG signal changes over time and tends to decrease as a function of distance from the initial response as evoked by the onset of the word. The amplitude at "early" and "late" disambiguation is therefore not directly comparable. Because of this, it is unclear whether the responses at late disambiguation reflect a relative increase for non-words compared with words or a relative decrease to words compared with non-words. We offer interpretations for both possibilities in the discussion below.

When running the same analysis for the ambiguity variable, no interactions were observed with latency—words that had an ambiguous onset elicited a stronger response at POD regardless of how many milliseconds or phonemes elapsed before disambiguation.

We performed the same analysis when defining latency in terms of elapsed phonemes rather than milliseconds, but no interaction was formed in the first cluster-forming stage.

Overall, it appears that non-words are more difficult to process than words when disambiguation of the onset phoneme comes later than 450 ms. This suggests that the system does indeed commit to a phonological category after approximately half a second. The interaction that we observed may reflect the system having to reinterpret the input when it has committed to the wrong category (thus perceiving a non-word) or a relative benefit in processing valid words when it has committed to the correct category.

Discussion

In this study, we aimed to address three research questions. First, does the recognition of phonological ambiguity manifest as an early perceptual process or a higher-order postperceptual process? Second, how is subphonemic maintenance and phonological commitment neurally instantiated? Third, what temporal constraints are placed on the system; in other words, what is the limit on how late subsequent context can be received and still be optimally integrated? We discuss our results in light of these three objectives.

Early sensitivity to ambiguity and acoustics

We found evidence for sensitivity to phonological ambiguity very early during processing, at just 50 ms after onset in left HG. Ambiguity was orthogonal to continuum position-that is, linear acoustic differences-sensitivity to which tended to be right lateralized and occur slightly later. Although previous studies have found the p50m to be modulated by VOT (Steinschneider et al., 1999; Hertrich et al., 2000) and PoA (Tavabi et al., 2007) and fMRI studies have found sensitivity to ambiguity in primary auditory cortex (Kilian-Hütten et al., 2011) (see the introduction), this is the first evidence of such early responses tracking proximity to perceptual boundaries. This finding supports a hierarchical over reverse-hierarchical processing model (Kilian-Hütten et al., 2011) because sensitivity is apparent before higher-level linguistic features (e.g., phonetic features, word identity) are processed. Note that we do not deny the possibility of top-down influence per se; indeed, it is likely that contextual effects vis-à-vis the experimental task, stimulus set, and task difficulty play a role. However, it appears that, relative to the processing of lexical features specifically, phoneme ambiguity is one of the first to come online. This suggests that sensitivity to ambiguity is not a byproduct of processing these higher-order properties, but rather is part of the earliest sensory stage of processing. This illustrates therefore that early stages of processing are tuned to strikingly complex features of the acoustic signal, in this case, the distance between the acoustic signal and the perceptual boundary between phonetic features.

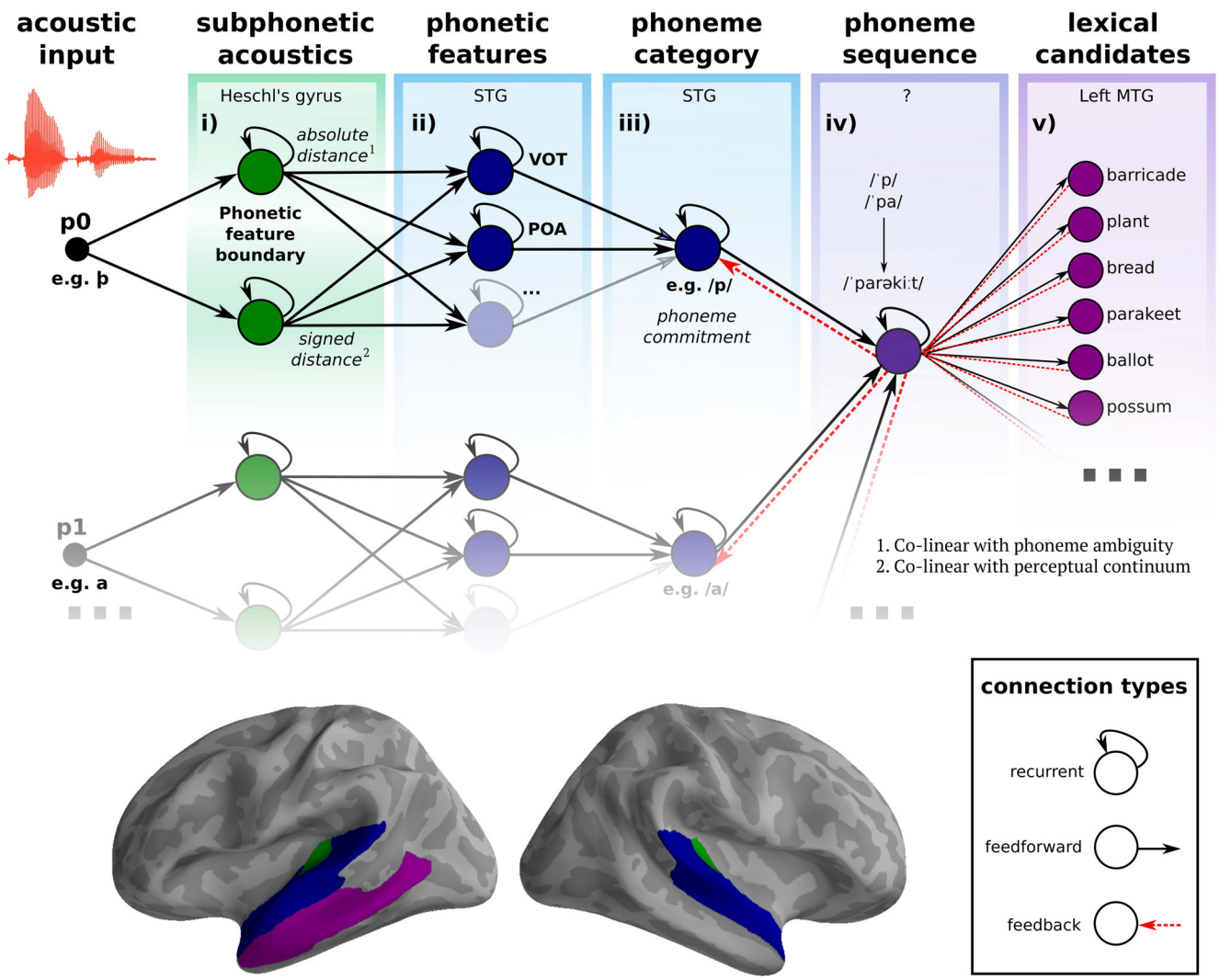

Figure 8. Schematic model of processing stages. Acoustic input in the form of spectrotemporal information is fed to primary auditory cortex (i). Here, we hypothesize that subphonetic acoustic information of the input is compared with an internal representation of the perceptual boundary between phonetic features. The absolute distance from the boundary is computed, which corresponds to phoneme ambiguity as tested in this study. The signed distance (i.e., closer to one category or another) corresponds to phoneme acoustics. This processing stage is therefore the locus of the ambiguity effect, although we do not claim that ambiguity is neurally represented per se. Next, this travels to STG (ii), where the phonetic features of a sound (e.g., VOT, PoA) are processed. Note that it is likely that other features of the sound, such as manner, are also generated at this stage, as indicated by the ellipsis. The outputs of these two stages are fed to a neural population that tries to derive a discrete phonological representation based on the features of the input (iii). This stage represents the "phoneme commitment" process, which converges over time by accumulating evidence through its own recurrent connection, as well as feedforward input from the previous stages and feedback from the subsequent stages. The output of the processes performed at each phoneme position then feeds to a node that tries to predict the phonological sequence of the word (iv) to activate potential lexical items based on partial matches with the input (*v*). Note that both /p/- and /b/-onset words are activated in the example because both cohorts are partially consistent. Below, we show the anatomical location associated with each processing stage. Stage i (processing subphonetic acoustic detail) is located in HG bilaterally (in green). Stages ii–iii (processing phonetic features) is in STG bilaterally (in blue). Stage v (activating lexical candidates) is in left middle temporal gyrus (in purple). Note the similarities with the functional organization of the dual-stream model proposed by Hickok and Poeppel (2007).

Because of the time that it takes the acoustic signal to reach primary auditory cortex, the early ambiguity effect must be reflecting a response to approximately the first 20 ms of the stimulus. Because we were able to decode phoneme category from the spectrotemporal properties of the first 20 ms of the acoustic stimuli (Fig. 4), it is clear that phoneme category information is present in the signal (also see Blumstein et al., 1977 and Stevens and Blumstein, 1978 for a similar conclusion in voiced PoA contrasts). This is consistent with an analysis by synthesis model (Halle and Stevens, 1962; Poeppel and Monahan, 2011) in which responses reflect the number of candidate phonemic representations generated by the first ~20 ms of acoustic signal. Neurons fire more when the search space over phonemic hypotheses is larger and less when there are fewer possibilities.

In addressing the first question, then, it appears that sensitivity to phonological ambiguity indeed reflects an early perceptual process and is not driven by higher-order lexical properties.

Reemergence of subphonemic detail

We observed a reemergence of sensitivity to the ambiguity, acoustics, PoA, and VOT of the word-onset phoneme at intermediate phoneme positions along the length of the word, at the disambiguation point, and at the two phonemes after disambiguation. This was specifically time-locked to the onset of each incoming phoneme and was not apparent when analyzing based on the time elapsed from word onset (cf. Figs. 5, 6). This novel finding is critically important because it supports the hypothesis that the subphonemic representation of a speech sound is main-

tained in superior temporal regions throughout the duration of a word even while subsequent phonemes are being received, perhaps suggesting that the percept of a speech sound is reassessed at each increment based on the provision of additional input. We refer to "subphonemic representations" here because there is not just prolonged sensitivity to phoneme ambiguity but rather prolonged sensitivity to all four of the orthogonal linguistic dimensions that we tested. Together, these dimensions make up what one may sensibly consider the neural representation of a speech sound. We also have evidence that ambiguity plays an additional role in modulating lexical predictions above and beyond subphonemic maintenance (Gwilliams et al., 2018). From this perspective, prolonged sensitivity to ambiguity may also arise from subsequent difficulty in processing the identity of a word with an ambiguous onset, which is consistent with a prediction error account (Rogers and Davis, 2017).

Further, it appears that phonemic reactivation is a general feature of speech comprehension rather than a specific mechanism recruited in the presence of ambiguity. Specifically, our results indicate that subphonemic information is maintained even when uncertainty about phoneme identity is low in two ways. First, reemergence of phonetic properties was not specific to the ambiguous tokens; it also occurred for the unambiguous phonemes. Second, information about phonetic features continues to be conserved after disambiguating information became available. Overall, these observations are the first to reveal that subphonemic information is maintained, not just in terms of uncertainty about categorization, but also in terms of fine-grained phonetic and acoustic detail of the phoneme under scrutiny. Both sources of information continue to be revisited over long timescales. This answers a long-standing question from the psycholinguistic literature (Bicknell et al., 2016).

In addressing our second research question, it appears that subphonemic maintenance is instantiated by maintaining phonetic, acoustic, and uncertainty information in auditory cortex and reactivating that information at the onset subsequent phonemes.

Commitment to phonological categories

Finally, we do see evidence for phonological commitment resolving on a timescale of ~300–450 ms (Fig. 7). The superiority of defining latency in terms of elapsed milliseconds rather than phonemes may indicate that commitment is based on the amount of time or number of completed processing cycles rather than intervening information. This process is supported by higher auditory processing regions in anterior STG, a location consistent with a meta-analysis of auditory word recognition (DeWitt and Rauschecker, 2012). Critically, phonological commitment seems to be computed in parallel to, and independently from, the maintenance of subphonemic detail in primary auditory regions. Before ~300 ms, there is no cost associated with resolution to a lexical item less consistent with word onset: listeners do not get temporarily misled (garden-pathed) provided resolution comes early enough (Fig. 7*C*, green bars). This suggests that the cohort of words consistent with either phonological interpretation is considered together (e.g., in the presence of b/p ambiguity, both the p-onset and b-onset words are activated). This is fully consistent with previous behavioral studies (Martin and Bunnell, 1981; Gow, 2001; Gow and McMurray, 2007) and a previous eye-tracking study (McMurray et al., 2009) that used similar materials and found look-contingent responses to be dependent upon phonetic information at lexical onset until at least ~300 ms (the longest disambiguation delay they tested). However, after ~450 ms, a difference begins to emerge when there is a mismatch between the more likely word given word onset and the resolving lexical information (Fig. 7*C*, red bars) (e.g., "barricade" is more likely if the onset phoneme was more b-like than p-like, so hearing "parakeet" is a mismatch). Because of the dynamics of the MEG response, it is hard to know whether the crux of this effect reflects a relative benefit for processing words (resulting in less activity) or a relative cost for processing non-words (resulting in more activity). If the former, committing to the correct phoneme entails receiving subsequent input that is consistent with expectations, making it easier to process. Conversely, committing to the incorrect phoneme leads to subsequent input outside of expectations, leading to something like a prediction-error response (Gagnepain et al., 2012). If the latter, increased responses to non-words may reflect the recruitment of a repair mechanism and reanalysis of the input from making an incorrect commitment.

Finding maintained sensitivity to subphonemic detail in parallel to phonological commitment is very important for the interpretation of psychophysical research, which has implicitly equated insensitivity to within-category variation with phonological commitment (Connine et al., 1991; Szostak and Pitt, 2013; Bicknell et al., 2016). This previous work has largely converged on a processing model whereby phonological commitment can be delayed for around 1 s after onset. Our results indicate that, in contrast, whereas subphonemic detail is indeed maintained over long timescales, this does not implicate that commitment is also put off for this length of time. Phonological commitment and subphonemic maintenance appear to be independent processes; it is not the case that lower-level information is discarded by the system once higher-level representations are derived.

In sum, the answer to our third research question is that subsequent context can be optimally integrated if it is received within approximately half a second, which is when the system commits to a phonological interpretation. However, subphonemic detail is maintained past the point that the system makes such a commitment.

Relationship to models of speech processing

It is unclear which model of speech processing can account for these data. Although Shortlist (Norris, 1994) and Shortlist B (Norris and McQueen, 2008) may be able to model recovery from lexical garden paths, they do not explicitly model processing of subphonemic detail. Although the MERGE model (Norris et al., 2000) is capable of modeling such detail, it proposes no feedback from the lexical to phonological levels of analysis. This is inconsistent with the observation that lexical information also serves to modulate the phonological commitment process (Gwilliams et al., 2018). Although it has been demonstrated that TRACE (McClelland and Elman, 1986) can be modified to simulate recovery by removing phoneme-level inhibition (McMurray et al., 2009), it does not provide the architecture to model initial sensitivity to phoneme ambiguity or account for how the percept of speech sounds is modulated by past and future linguistic information (see Grossberg and Kazerounian, 2011 for an overview of TRACE limitations). It is also unclear whether this modification would interfere with TRACE's success in accounting for a range of observations in spoken word recognition (for review, see Gaskell, 2007). One model proposed to deal with TRACE's shortcoming is adaptive resonance theory: each speech sound produces a resonance wave that is influenced by top-down information until it reaches equilibrium and surfaces to consciousness (Carpenter and Grossberg, 2016). Although this theory is consistent with the idea that there is a critical time limit to receive top-down information, it suggests that there is a linear decay in subphonemic

information as temporal distance from the phoneme increases. Our results do not support that conjecture. Instead, they suggest that subphonemic information is re-evoked later in processing with a similar magnitude as that experienced at onset. In light of the present results, one shortcoming of these models is their attempt to explain spoken word recognition with a single mechanism built on the assumption that acoustic-phonetic information is lost once a phonological categorization is derived. Instead, our results suggest that a multi-element processing model is more appropriate, allowing for a dynamic interaction among subphonetic, phonetic, phonological, and lexical levels of analysis.

The model depicted in Figure 8 offers a mechanistic explanation for the four main findings of the present study. First, early sensitivity to ambiguity is captured by placing the processing of subphonetic acoustics as one of the first cortical operations performed on the signal in HG. Here, we propose that the acoustic input is compared with a representation of the perceptual boundary between phonetic features. The absolute distance between the input and the boundary corresponds to our "ambiguity" variable.

Second, most nodes in the model are depicted with a recurrent connection. This is critical because it allows the system to maintain multiple representations in parallel and over long timescales.

Third, we hypothesize that each phoneme of the input goes through the same set of processes, and uses the same neural machinery (though, here we only have evidence for the first phoneme position, p0). This could explain why information about the previous phoneme reemerges at subsequent phoneme positions and the maintained signal held in the recurrent connections only becomes detectable when we "ping the brain" with additional input (Wolff et al., 2017).

Finally, phonological commitment is captured by an evidence accumulation process, which receives bottom-up input from the phonetic features, top-down input from lexical processes, and self-terminating connections to converge to a particular category. This idea of evidence accumulation has clear similarities with the drift diffusion models used to explain perceptual decision making (Gold and Shadlen, 2007) and could indeed reflect an analogous process. The consequences of incorrect commitment would be carried by feedback from the lexical stage: if commitment is wrong, no lexical items remain consistent with the input at POD and an error-like signal is fed back to the previous computations.

Conclusion

Later sounds determine the perception of earlier speech sounds through the simultaneous recruitment of prolonged acoustic-phonetic maintenance and rapid phonological commitment. In this manner, quick lexical selection is achieved by committing to phonological categories early, often before the system is completely certain that it is the correct choice. In situations where subsequent information reveals that the wrong phoneme was chosen, the maintained acoustic-phonetic information can be reanalyzed in light of subsequent context to derive the correct commitment. This facilitates rapid contact with lexical items to derive the message of the utterance, as well as continued revisitation to the phonetic level of analysis to reduce parsing errors. The human brain therefore solves the issue of processing a transient hierarchically structured signal by recruiting complementary computations in parallel rather than conceding to the trade-off between speed and accuracy.

References

Ackermann H, Lutzenberger W, Hertrich I (1999) Hemispheric lateralization of the neural encoding of temporal speech features: a whole-head magnetencephalography study. Brain Res Cogn Brain Res 7:511–518. CrossRef Medline

Adachi Y, Shimogawara M, Higuchi M, Haruta Y, Ochiai M (2001) Reduction of non-periodic environmental magnetic noise in MEG measurement by continuously adjusted least squares method. IEEE Transactions on Applied Superconductivity 11:669–672. CrossRef

Balota DA, Yap MJ, Cortese MJ, Hutchison KA, Kessler B, Loftis B, Neely JH, Nelson DL, Simpson GB, Treiman R (2007) The english lexicon project. Behavior Research Methods 39:445–459. CrossRef Medline

Bates D, Mächler M, Bolker B, Walker S (2014) Fitting linear mixed-effects models using lme4. arXiv preprint arXiv:14065823. Available at https://arxiv.org/pdf/1406.5823.pdf

Bicknell K, Jaeger TF, Tanenhaus MK (2016) Now or . . . later: Perceptual data is not immediately forgotten during language processing. Behavioral and Brain Sci 39:23–24. CrossRef

Blumstein SE, Stevens KN, Nigro GN (1977) Property detectors for bursts and transitions in speech perception. J Acoust Soc Am 61:1301–1313. CrossRef Medline

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. CrossRef Medline

Boersma P, Weenink D (2002) Praat, a system for doing phonetics by computer. Glot international 5:341–345.

Carpenter GA, Grossberg S (2016) Adaptive resonance theory, pp 1–17. Boston, MA: Springer.

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

Cole RA (1973) Listening for mispronunciations: a measure of what we hear during speech. Percept Psychophys 13:153–156. CrossRef

Connine CM, Blasko DG, Hall M (1991) Effects of subsequent sentence context in auditory word recognition: temporal and linguistic constraints. J Mem Lang 30:234–250. CrossRef

Dale AM, Liu AK, Fischl BR, Buckner RL, Belliveau JW, Lewine JD, Halgren E (2000) Dynamic statistical parametric mapping: combining fMRI and MEG for high-resolution imaging of cortical activity. Neuron 26:55–67. CrossRef Medline

DeWitt I, Rauschecker JP (2012) Phoneme and word recognition in the auditory ventral stream. Proc Natl Acad Sci U S A 109:E505–E514. CrossRef Medline

Di Liberto GM, O'Sullivan JA, Lalor EC (2015) Low-frequency cortical entrainment to speech reflects phoneme-level processing. Curr Biol 25:2457–2465. CrossRef Medline

Ettinger A, Linzen T, Marantz A (2014) The role of morphology in phoneme prediction: evidence from MEG. Brain Lang 129:14–23. CrossRef Medline

Gage N, Poeppel D, Roberts TP, Hickok G (1998) Auditory evoked M100 reflects onset acoustics of speech sounds. Brain Res 814:236–239. CrossRef Medline

Gagnepain P, Henson RN, Davis MH (2012) Temporal predictive codes for spoken words in auditory cortex. Curr Biol 22:615–621. CrossRef Medline

Ganong WF 3rd (1980) Phonetic categorization in auditory word perception. J Exp Psychol Hum Percept Perform 6:110–125. CrossRef Medline

Gaskell MG (2007) Statistical and connectionist models of speech perception and word recognition. In: The Oxford handbook of psycholinguistics (Gaskell GM, Altmann G, Altmann GTM, Bloom P, eds), pp 55–69. Oxford, UK: Oxford University Press.

Gold JI, Shadlen MN (2007) The neural basis of decision making. Annu Rev Neurosci 30:535–574. CrossRef Medline

Gordon PC, Eberhardt JL, Rueckl JG (1993) Attentional modulation of the phonetic significance of acoustic cues. Cogn Psychol 25:1–42. CrossRef Medline

Gow DW (2001) Assimilation and anticipation in continuous spoken word recognition. J Mem Lang 45:133–159. CrossRef

Gow DW, McMurray B (2007) Word recognition and phonology: the case of English coronal place assimilation. Papers in Laboratory Phonology 9:173–200.

Gow DW Jr, Segawa JA, Ahlfors SP, Lin FH (2008) Lexical influences on speech perception: a granger causality analysis of MEG and EEG source estimates. Neuroimage 43:614–623. CrossRef Medline

Gow DW Jr, Segawa JA (2009) Articulatory mediation of speech perception:

a causal analysis of multi-modal imaging data. Cognition 110:222–236. CrossRef Medline

Gramfort A, Luessi M, Larson E, Engemann DA, Strohmeier D, Brodbeck C, Parkkonen L, Hämäläinen MS (2014) MNE software for processing MEG and EEG data. Neuroimage 86:446–460. CrossRef Medline

Grossberg S, Kazerounian S (2011) Laminar cortical dynamics of conscious speech perception: neural model of phonemic restoration using subsequent context in noise. J Acoust Soc Am 130:440–460. CrossRef Medline

Gwilliams L, Marantz A (2015) Non-linear processing of a linear speech stream: the influence of morphological structure on the recognition of spoken arabic words. Brain Lang 147:1–13. CrossRef Medline

Gwilliams L, Marantz A (2018) Morphological representations are extrapolated from morpho-syntactic rules. Neuropsychologia 114:77–87. CrossRef Medline

Gwilliams LE, Monahan PJ, Samuel AG (2015) Sensitivity to morphological composition in spoken word recognition: evidence from grammatical and lexical identification tasks. J Exp Psychol Learn Mem Cogn 41:1663–1674. CrossRef Medline

Gwilliams L, Lewis GA, Marantz A (2016) Functional characterisation of letter-specific responses in time, space and current polarity using magnetoencephalography. Neuroimage 132:320–333. CrossRef Medline

Gwilliams L, Poeppel D, Marantz A, Linzen T (2018) Phonological (un) certainty weights lexical activation. Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2018), Salt Late City, Utah, January.

Halle M, Stevens K (1962) Speech recognition: a model and a program for research. IRE Transactions on Information Theory 8:155–159. CrossRef

Heil P (2004) First-spike latency of auditory neurons revisited. Curr Opin Neurobiol 14:461–467. CrossRef Medline

Hertrich I, Mathiak K, Lutzenberger W, Ackermann H (2000) Differential impact of periodic and aperiodic speech-like acoustic signals on magnetic M50/M100 fields. Neuroreport 11:4017–4020. Medline

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

Holmes AP, Blair R, Watson JD, Ford I (1996) Nonparametric analysis of statistic images from functional mapping experiments. J Cereb Blood Flow Metab 16:7–22. CrossRef Medline

Kawahara H, Morise M (2011) Technical foundations of TANDEM-STRAIGHT, a speech analysis, modification and synthesis framework. Sadhana 36:713–727. CrossRef

Kawahara H, Morise M, Takahashi T, Nisimura R, Irino T, Banno H (2008) Tandem-STRAIGHT: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0, and aperiodicity estimation. 2008 IEEE International Conference on Acoustics, Speech and Signal Processing, Las Vegas, Nevada, March.

Kilian-Hütten N, Valente G, Vroomen J, Formisano E (2011) Auditory cortex encodes the perceptual interpretation of ambiguous sound. J Neurosci 31:1715–1720. CrossRef Medline

Leonard MK, Baud MO, Sjerps MJ, Chang EF (2016) Perceptual restoration of masked speech in human cortex. Nat Commun 7:13619. CrossRef Medline

Maris E, Oostenveld R (2007) Nonparametric statistical testing of EEG- and MEG-data. J Neurosci Methods 164:177–190. CrossRef Medline

Martin JG, Bunnell HT (1981) Perception of anticipatory coarticulation effects. J Acoust Soc Am 69:559–567. CrossRef Medline

McClelland JL, Elman JL (1986) The TRACE model of speech perception. Cogn Psychol 18:1–86. CrossRef Medline

McMurray B, Tanenhaus MK, Aslin RN (2009) Within-category VOT affects recovery from "lexical" garden paths: evidence against phoneme-level inhibition. J Mem Lang 60:65–91. CrossRef Medline

McQueen JM (1991) The influence of the lexicon on phonetic categorization: stimulus quality in word-final ambiguity. J Exp Psychol Hum Percept Perform 17:433–443. CrossRef Medline

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

Myers EB, Blumstein SE (2008) The neural bases of the lexical effect: an fMRI investigation. Cereb Cortex 18:278–288. CrossRef Medline

Norris D (1994) Shortlist: a connectionist model of continuous speech recognition. Cognition 52:189–234. CrossRef

Norris D, McQueen JM (2008) Shortlist B: a bayesian model of continuous speech recognition. Psychol Rev 115:357–395. CrossRef Medline

Norris D, McQueen JM, Cutler A (2000) Merging information in speech recognition: feedback is never necessary. Behav Brain Sci 23:299–325; discussion 325–370. CrossRef Medline

Obleser J, Lahiri A, Eulitz C (2003) Auditory-evoked magnetic field codes place of articulation in timing and topography around 100 milliseconds post syllable onset. Neuroimage 20:1839–1847. CrossRef Medline

Obleser J, Lahiri A, Eulitz C (2004) Magnetic brain response mirrors extraction of phonological features from spoken vowels. J Cogn Neurosci 16:31–39. CrossRef Medline

Papanicolaou AC, Castillo E, Breier JI, Davis RN, Simos PG, Diehl RL (2003) Differential brain activation patterns during perception of voice and tone onset time series: a MEG study. Neuroimage 18:448–459. CrossRef Medline

Poeppel D, Monahan PJ (2011) Feedforward and feedback in speech perception: revisiting analysis by synthesis. Lang Cogn Proc 26:935–951. CrossRef

R Core Team (2014). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Available at http://www.R-project.org/.

Rogers JC, Davis MH (2017) Inferior frontal cortex contributions to the recognition of spoken words and their constituent speech sounds. J Cogn Neurosci 29:919–936. CrossRef Medline

Samuel AG (1981) The role of bottom-up confirmation in the phonemic restoration illusion. J Exp Psychol Hum Percept Perform 7:1124–1131. CrossRef Medline

Samuel AG (1991) A further examination of attentional effects in the phonemic restoration illusion. Q J Exp Psychol A 43:679–699. CrossRef Medline

Simos PG, Diehl RL, Breier JI, Molis MR, Zouridakis G, Papanicolaou AC (1998) MEG correlates of categorical perception of a voice onset time continuum in humans. Brain Res Cogn Brain Res 7:215–219. CrossRef Medline

Steinschneider M, Schroeder CE, Arezzo JC, Vaughan HG Jr (1995) Physiologic correlates of the voice onset time boundary in primary auditory cortex (A1) of the awake monkey: temporal response patterns. Brain Lang 48:326–340. CrossRef Medline

Steinschneider M, Volkov IO, Noh MD, Garell PC, Howard MA 3rd (1999) Temporal encoding of the voice onset time phonetic parameter by field potentials recorded directly from human auditory cortex. J Neurophysiol 82:2346–2357. CrossRef Medline

Stevens KN, Blumstein SE (1978) Invariant cues for place of articulation in stop consonants. J Acoust Soc Am 64:1358–1368. CrossRef Medline

Szostak CM, Pitt MA (2013) The prolonged influence of subsequent context on spoken word recognition. Atten Percept Psychophys 75:1533–1546. CrossRef Medline

Tavabi K, Obleser J, Dobel C, Pantev C (2007) Auditory evoked fields differentially encode speech features: an MEG investigation of the P50m and N100m time courses during syllable processing. Eur J Neurosci 25:3155–3162. CrossRef Medline

Warren RM (1970) Perceptual restoration of missing speech sounds. Science 167:392–393. CrossRef Medline

Wolff MJ, Jochim J, Akyürek EG, Stokes MG (2017) Dynamic hidden states underlying working-memory-guided behavior. Nat Neurosci 20:864–871. CrossRef Medline

Yuan J, Liberman M (2008) The Penn Phonetics Lab forced aligner. Available at http://www.ling.upenn.edu/phonetics/p2fa/.