Published in final edited form as:

Brain Lang. 2018 September ; 184: 32–42. doi:10.1016/j.bandl.2018.06.006.

The time-course of cortical responses to speech revealed by fast optical imaging

Joseph C. Toscano^{1,2}, Nathaniel D. Anderson^{2,3}, Monica Fabiani^{2,3}, Gabriele Gratton^{2,3}, and Susan M. Garnsey^{2,3}

¹Department of Psychological & Brain Sciences Villanova University

²Beckman Institute for Advanced Science & Technology University of Illinois at Urbana-Champaign

³Department of Psychology University of Illinois at Urbana-Champaign

Abstract

Recent work has sought to describe the time-course of spoken word recognition, from initial acoustic cue encoding through lexical activation, and identify cortical areas involved in each stage of analysis. However, existing methods are limited in either temporal or spatial resolution, and as a result, have only provided partial answers to the question of how listeners encode acoustic information in speech. We present data from an experiment using a novel neuroimaging method, fast optical imaging, to directly assess the time-course of speech perception, providing non-invasive measurement of speech sound representations, localized to specific cortical areas. We find that listeners encode speech in terms of continuous acoustic cues at early stages of processing (ca. 96 ms post-stimulus onset), and begin activating phonological category representations rapidly (ca. 144 ms post-stimulus). Moreover, cue-based representations are widespread in the brain and overlap in time with graded category-based representations, suggesting that spoken word recognition involves simultaneous activation of both continuous acoustic cues and phonological categories.

Keywords

speech perception; phonological categorization; spoken language processing; optical imaging; event-related potentials

A long-standing issue in language processing concerns the nature of representations used by the brain to perceive speech. Debate continues about whether perception is based on

Corresponding author: Joseph Toscano, Department of Psychological & Brain Sciences, Villanova University, 800 E Lancaster Ave, Villanova, PA 19085, Phone: 610-519-4755, Fax: 610-519-4269, joseph.toscano@villanova.edu.

Author Contributions

All authors jointly designed the experiment and wrote the manuscript. JCT and NDA collected the data. JCT, NDA, and GG analyzed the data.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

continuous acoustic cues (Pisoni & Tash, 1974; Toscano, McMurray, Dennhardt, & Luck, 2010), discrete phonemes (Liberman et al., 1957; Chang et al., 2010), or other representations (e.g., auditory contrasts, Diehl, Lotto, & Holt, 2004; articulatory gestures, Viswanathan, Fowler, & Magnuson, 2009). For example, in English, the discrete phonemic categories of /b/ and /p/ are distinguished by several continuous acoustic cues, including voice onset time (VOT¹; Lisker & Abramson, 1964; Fig. 1A). In order to accurately recognize speech, listeners must map these cues from the speech signal onto phoneme categories. The question addressed here concerns when and where cue- and category-level representations are each used during spoken language processing.

This issue is central to an early hypothesis about speech perception, known as *categorical perception* (Liberman et al., 1957)—the idea that listeners only perceive speech in terms of discrete units, assessed behaviorally by showing that listeners' category boundaries along an acoustic continuum (obtained in an identification/labeling task) align with the peak of their discrimination function for those sounds. Categorical perception was proposed, in part, to explain human listeners' remarkable ability to recognize speech sounds, which seems to differ from their ability to recognize other sounds and from the auditory abilities of non-human animals. However, the debate over the original categorical perception hypothesis is largely settled: Listeners are indeed sensitive to graded differences within phonetic categories, contrary to early proposals (Pisoni & Tash, 1974; Massaro & Cohen, 1983; Miller, 1997; McMurray, Tanenhaus, & Aslin, 2002; Toscano et al., 2010). Moreover, sensitivity to continuous cues is critical for accurate speech perception, as it allows listeners to overcome contextual variability (McMurray & Jongman, 2011). Thus, at minimum, within-category phoneme differences are preserved in the brain in some form. However, it is unclear when during processing phonological categories play a role, and whether such representations coexist with cue-level representations. Do listeners represent speech in terms of acoustic cues early in perception? Or are sounds immediately encoded as categories (either graded or discrete)? Answering these questions is critical not only for our understanding of the mechanisms underlying language comprehension, but also for the development of computer-based speech recognition systems that mimic processes used by humans (Scharenborg, 2007) and for evaluating neurobiological models of spoken language processing (Hickok & Poeppel, 2007; Scott & Johnsrude, 2003).

Cue- vs. category-based representations

Behavioral experiments have been unable to determine whether the brain encodes acoustic cues independently of phonemes, since behavioral responses reflect both early perceptual processing and later categorization stages. More importantly, they provide only indirect information about the time-course of speech perception. If the brain encodes continuous cues, how and when are these representations used in identifying phonemes? Are both types of representations maintained in different cortical areas simultaneously? These questions relate to a larger debate about whether language processing is inherently a serial or a parallel

¹VOT is defined as the time difference between the release of consonantal closure and the onset of laryngeal voicing for word-initial stop consonants (/b,d,g,p,t,k/).

process (Trueswell, Tanenhaus, & Garnsey, 1994), and distinguishing these models requires us to examine how speech sounds are represented *during* perception.

Recently, speech researchers have turned to neurophysiological measures, primarily using functional MRI (fMRI) and event-related potential (ERP) techniques, and several neurobiological models of speech perception have been based on such data. However, both techniques have inherent limitations that prevent us from simultaneously measuring the early time-course of speech perception and localizing the brain regions involved. fMRI can identify brain areas that are engaged when processing lexical and phonological information, as well as sub-phonemic differences (Myers et al., 2009; Blumstein, Myers, & Rissman, 2005), but because it relies on hemodynamic responses, this approach has limited temporal resolution, making it impossible to distinguish early perceptual encoding from later-occurring processes.

In contrast, the ERP technique has excellent temporal resolution, allowing us to demonstrate that early sensory components such as the auditory N1 vary with continuous acoustic cues but not phonological categories (Toscano et al., 2010). However, it is unclear where these responses are generated in the brain (Picton et al., 1974; Giard et al., 1994). It is also difficult to separate contributions of multiple sources (which may encode different types of information) in scalp-recorded EEG, and therefore, to determine whether different types of representations may simultaneously co-exist in the brain.

Thus, neither method provides both the temporal and spatial resolution needed to effectively study the time-course of speech perception, and existing studies have not resolved this issue. MEG offers a potential alternative (Frye et al., 2007; Phillips et al., 2000; Obleser et al., 2003), but this technique is not widely available and still suffers from the inverse problem that prevents accurate source localization of scalp EEG (de Peralta-Menendez & Gonzalez-Andino, 1998). Intracranial EEG recording (or electrocorticography, ECOG; Chang et al., 2010) provides a promising solution, but its use is limited to studies with small groups of surgical patients. Moreover, these experiments have found inconsistent results with regard to whether or not different cortical areas are sensitive to continuous cues (Chang et al., 2010; Nourski et al., 2014; 2015; Pasley et al., 2012), partly because they have either focused on identifying categorical responses (e.g., Chang et al., 2010, which used techniques such as K-means clustering, though the data may show evidence of both continuous and categorical effects) or have relied on continuous (i.e., connected) speech (Mesgarani et al., 2014), which while more natural, lacks the level of stimulus control needed to separate cue- and category-based representations. Overall, it is unclear based on the existing ECOG literature whether listeners are sensitive to continuous cues or only higher-level categories.

Temporal and spatial localization of speech sound representations

Despite the limitations of existing methods, preliminary hypotheses may be generated based on previous studies. First, several experiments have used ERP and MEG techniques to assess early perceptual responses, such as the auditory N1, as well as measures of auditory discrimination, such as the mismatch negativity (MMN). Findings from these studies, however, have been mixed. Several have investigated within vs. between category phonetic

discrimination using the MMN (or its magnetic equivalent), with some showing differences for within-category contrasts (Joanisse, Robertson, & Newman, 2007; Sharma, Kraus, McGee, Carrell, & Nicol, 1993) and others showing effects only for between-category distinctions (Dehaene-Lambertz, 1997; Phillips et al., 2000).

Other studies have examined changes in the auditory N1 as a function of differences along specific acoustic cue continua, providing more direct measures of cue-level encoding. Early work using this approach suggested that speech sounds are encoded categorically (Sharma & Dorman, 1999), but later work found evidence that the N1 (and its magnetic equivalent) tracks continuous changes along acoustic dimensions, with no evidence of category-level effects (discrete or otherwise; Frye et al., 2007; Toscano et al., 2010). Recent work has also investigated ERP responses in connected speech. However, these studies have focused on distinguishing phoneme categories by collapsing across individual speech sounds within a category, rather than assessing encoding of specific acoustic cues (Khalighinejad et al., 2017; see also ECOG data from Mesgarani et al., 2015) or have found evidence for sensitivity to both fine-grained acoustic differences and higher-level phonological categories (Di Liberto et al., 2015). Thus, overall, results from ERP and MEG studies are mixed, with some experiments suggesting that listeners are sensitive to within-category differences, encoding this information in terms of continuous acoustic cues or as graded phonetic categories; and others suggesting that listeners are insensitive to within-category differences, encoding speech sounds in terms of discrete differences along acoustic dimensions or as discrete categories.

fMRI studies have provided information about where in the brain these responses are generated, and several have investigated the question of whether listeners are sensitive to within-category phonetic differences for sounds varying along acoustic continua. These studies suggest that listeners are sensitive to graded differences within phonetic categories in IFG (Blumstein et al., 2005) and STG (Myers et al., 2009). Moreover, there is evidence for representations that are insensitive to within-category differences in IFS (Myers et al., 2009). Thus, listeners may represent speech sounds differently in different cortical areas (as either graded or discrete categories). Note, however, that these studies have examined differences relative to phonological categories; it is unclear whether cue-level representations themselves (cf. Toscano et al., 2010) are graded or discrete in these cortical areas.

Finally, studies using intracranial EEG offer the opportunity to study the time-course of speech processing and localize this information to specific cortical areas simultaneously (albeit in a small number of surgical patients). Several of these studies have suggested evidence for categorical speech sound encoding in posterior STG (pSTG; Chang et al., 2010; Mesgarani et al., 2012). However, these studies have focused specifically on whether or not pSTG encodes between-category differences, and thus, do not establish whether effects are graded within categories. Other ECOG studies have suggested fine-grained encoding of speech sounds in pSTG, even sufficient to reconstruct acoustic representations from the intracranial data (Pasley et al., 2012). Thus, pSTG may indeed encode within-category differences, as suggested by fMRI data (Myers et al., 2009), but it is unclear whether these responses track changes along acoustic cue dimensions or only differences relative to phonological category structure.

Summary and open questions

Together, the studies above suggest that both graded and discrete representations of speech sounds may exist in different cortical areas, particularly STG (evidence for both graded and discrete categories) and IFG (graded categories). Experiments examining the early timecourse of processing using ERP and MEG techniques are mixed, with some suggesting that listeners encode discrete differences along acoustic dimensions, and others showing continuous changes in brain responses as a function of differences along continua.

What is missing from these data, however, is information about the time-course of processing for both cue- and category-based representations, localized to specific cortical areas. It is unclear when and where cue-level representations (i.e., responses tracking changes along continuous acoustic dimensions) and category-level representations (responses relative to phonological category boundaries) are found, and whether such responses are graded or discrete at early stages of processing. Previous work has not examined both types of representations with sufficient spatial and temporal resolution to address this issue.

Approach of the study

We use a novel neuroimaging technique—fast optical imaging (or, the event-related optical signal; EROS)—to overcome this problem. This method is non-invasive and provides good temporal and spatial resolution simultaneously (Gratton & Fabiani, 2001). Near-infrared light is used to measure activity-related changes in the light-scattering properties of cortical neurons. Because EROS measures neuronal activity, rather than hemodynamic responses, it provides good temporal resolution. Moreover, because the diffusion of near-infrared light across the brain is quite limited, the signal is highly localized (1 cm or less) and thus provides good spatial resolution (Gratton & Fabiani, 2003). This approach has been experimentally validated by showing that single pulse TMS results in the immediate (latency < 8 ms) elicitation of EROS responses in the stimulated region, as well as delayed (by 40–50 ms) activity in connected areas (Parks et al., 2012, 2015). As an example of the information that can be provided by this novel approach, Tse, Low, Fabiani, and Gratton (2012) used EROS to show that different kinds of contingency rules can be represented in different brain regions and accessed in rapid succession.

Here we intend to use EROS as a way to both examine the time-course of speech perception and identify cortical areas that respond to cue- and category-level differences. Furthermore, EEG can be measured simultaneously with EROS (Gratton et al., 1997), so established ERP effects can be replicated and linked with simultaneously-occurring EROS data. We used the EROS technique to ask (1) whether early representations of speech sounds are graded or discrete, (2) when during processing listeners begin to encode sounds in terms of phonological categories, and (3) whether these two types of responses overlap in time.

Method

Design

Participants heard synthesized stimuli varying in VOT between /b/ and /p/, embedded in the words *beach* and *peach*, and identified which word they heard (Figs. 1A and 1B). The continuum was centered on each subject's category boundary (determined in a prior testing session; see Procedure below), plus VOT values 5 and 10 ms above and below the boundary (a total of five VOTs). Three hundred repetitions of each stimulus were presented. Optical data were collected using 40 near-infrared (830 nm) sources and 24 detectors with a 24-ms sampling period. These data were coregistered with structural MRIs for each subject, and event-locked averages were created with a 240-ms baseline.

Subjects

Twelve young adults (six women, mean age 23) were recruited from the University of Illinois at Urbana-Champaign community in accordance with UIUC IRB protocols, provided informed consent, and were compensated \$15 per hour. Subjects reported being native English speakers, having normal hearing, normal or corrected-to-normal vision, and no known history of neurological deficits. All participants were right-handed.

Procedure

The experiment was conducted over the course of three sessions. In the first session, participants listened to synthesized versions of the words *beach* and *peach*, varying in VOT along a 9-step acoustic continuum from 0 to 40 ms VOT, in 5-ms steps (the same *beach-peach* continuum used by Toscano et al., 2010). Subjects were first presented with three repetitions of each endpoint in random order with the corresponding response option (left vs. right) displayed on the computer screen. Response order was counterbalanced across subjects, with half of the participants making a left response for /b/ stimuli and half making a left response for /p/ stimuli (the same response mappings were used during the EROS/EEG session).

Next, subjects completed a practice block where they heard the continuum endpoints (each presented randomly 10 times) and received feedback about whether they identified them correctly. Participants needed to reach a criterion of 85% correct before moving on to the test blocks (typically, they reached criterion on the first attempt). The test blocks consisted of a two-alternative forced choice (2AFC) task in which they identified whether the initial sound was a /b/ or a /p/. Each stimulus was presented 20 times in random order (180 trials total). These data were used to assess each listener's category boundary along the /b-/p/ continuum, which determined which stimuli were used in the EROS recording session (see Data Analysis section below). The session took approximately 30–60 minutes to complete.

In the second session, a structural MRI was obtained for each participant. This MRI was used during data analysis to co-register the optical data in order to more accurately determine locations of cortical activity in each participant (see EROS Recording below). This session took approximately 15–30 minutes to complete.

In the final session, subjects performed the same 2AFC task from the first session, except that they heard only five different stimuli along the /b-/p/ continuum, with VOT values centered on their individual category boundary (e.g., a listener with a category boundary of 20 ms heard stimuli with VOTs of 10, 15, 20, 25, and 30 ms). Each stimulus was repeated 300 times, for a total of 1500 trials.² Trials were divided into 20 blocks of 75 stimuli, with each block including 15 repetitions of each stimulus presented in random order. Subjects were given an opportunity to take a break after each block. This session took 4–5 hours to complete (including setup time for the EEG and EROS recording).

EEG Recording

During the last session, ERP and EROS data were recorded simultaneously while the subject completed the task. EEG responses were recorded using a Brain Products BrainAmp MR EEG amplifier with electrodes located at three midline sites: Fz, Cz, and Pz. Horizontal and vertical electrooculograms were also recorded to identify eye-movement artifacts during offline analysis. EEG was referenced to the left mastoid online, and re-referenced to the average of the left and right mastoids offline. Electrode impedance was <5 k Ω , and data were recorded at a sampling rate of 1000 Hz. Due to equipment problems, usable EEG data was not obtained during the EROS session for five participants. These participants returned for an additional session, in which they completed either 375 trials (75 repetitions of each stimulus; four subjects) or 1200 trials (240 repetitions; one subject) while EEG (but not EROS) was recorded.

EROS Recording

EROS responses were measured using an Imagent frequency-domain oximeter system (ISS, Inc., Champaign, IL). Light was delivered with a set of 40 830-nm laser sources connected via fiber optic cables to a cap worn by the subject. Light refracted through the head was carried through another set of fiber optic cables to 24 photo-multiplier tubes, modulated at 110.003125 MHz, creating a 3125-Hz cross-correlation frequency.

The array of source-detector pairs created a set of 360 channels, based on the pairing of each of the 24 detectors with 15 of the 40 infrared sources. Sources were grouped into sets of 15, multiplexed at a sampling rate of 41.67 Hz. Sources from the opposite side of the head were multiplexed together (i.e., two sources at homologous locations on the left and on the right were illuminated simultaneously), since sources on the left could not contribute to light detected on the right and vice versa. This allowed for a full sweep through the set of sources every 24 ms.³ The cap worn by the subject was designed to sample as densely as possible over the left and right temporal and frontal lobes, allowing us to examine our two main regions of interest (ROIs): left PT/pSTG and left IFG, two of the areas reported to be

²Many repetitions are needed to achieve a reliable SNR for the EROS recording. Although multiple repetitions are common in speech perception studies, it is important to consider possible perceptual learning effects. Critically, the design of the experiment helps address this issue: listeners were only given feedback during the initial practice (consisting of the continuum endpoints). In addition, stimuli in the main task were equiprobable, meaning that distributional statistics could not have led to perceptual learning or category boundary changes (cf. Clayards et al., 2008; Munson, 2011).

³Because VOT is defined as a temporal difference, it is possible that listeners compute shorter VOTs earlier than longer VOTs. However, the range of VOTs heard by a given subject during EROS recording is 20 ms, which is less than the EROS sampling interval of 24 ms. Thus, any differences in the times at which VOT values were computed would not have impacted the current results.

involved in spoken word recognition. The montage of sources and detectors, and the voxels covered by this montage, are shown in Fig. 1C.

Optical data were co-registered with a structural MRI for each subject in order to localize EROS activity. During the MRI recording session, T1-weighted anatomical (MPRAGE) MRIs were obtained using a 3T Magnetom Allegra Headscanner (Siemens, New York, NY). The preauricular points and nasion were marked with vitamin E pills during the scan. During the EROS recording session, scalp locations of sources and detectors, along with the preauricular points, nasion, and 150 scalp locations were mapped using a Polaris Vicra infrared digitization system (NDI, Waterloo, Ontario, Canada). These data were co-registered with the MRI using a fitting procedure based on the locations of the scalp and fiducial points (Whalen, Maclin, Fabiani, & Gratton, 2007). Finally, co-registered data were transformed to Talairach space (Talairach & Tournoux, 1988) for analysis.

Data analysis

The dependent variable for the optical data was the relative phase delay of the modulated light measured at the detectors. Voxel activity was computed by averaging the phase delay of light pulses for source-detector channels whose paths pass through that voxel. Data were adjusted to account for phase wrapping, normalized, corrected for hemodynamic artifacts (i.e., pulse correction; Gratton & Corballis, 1995), and bandpass filtered to remove frequencies <1 Hz and >10 Hz. Channels with highly variable phase measurements (>132 ps, indicating poor measurements from those channels) and channels with source-detector distances less than 20 mm or greater than 45 mm (i.e., distances that would not lead to cortical activity detection) were excluded from analysis.

Event-related averages with a 240-ms baseline preceding stimulus onset were created for each condition for each subject. Averages from each channel were combined with the coregistered anatomical data to determine relative phase delay in 2.5-mm³ voxels. The light path between a source and detector was estimated assuming a homogenous medium with equivalent optical properties throughout (absorption coefficient $\mu_a=0.01 \text{ mm}^{-1}$ and reduced scattering coefficient $\mu_s=1.5 \text{ mm}^{-1}$). Phase delays in voxels with overlapping paths from multiple sourcedetector pairs were computed by taking the average of the overlapping channels.

Statistical analyses were performed within each ROI by fitting functions to the z-score transformed phase delay at each voxel, with stimuli coded in terms of VOT distance from each subject's category boundary (i.e., relative VOT). Separate functions were fit to identify cue-level and category-level effects. A cue-level (i.e., continuous, non-categorical) effect would be indicated by a linear trend across the VOT continuum (with maximal phase delay at either the voiced or voiceless endpoint; directionality of the effect is not theoretically meaningful, as it presumably reflects the orientation of the neural populations encoding this information). Thus, a linear function was fit to the subject's data to identify cue-level effects.

Category-level effects could be measured in multiple ways, such as multi-dimensional scaling (e.g., Chang et al., 2010; Khalighinejad et al., 2017). The approach we adopted here corresponds closely with the method used to identify cue-level responses. Specifically, to

identify category-level effects, we examine responses relative to listeners' category boundaries, similar to estimates obtained via category goodness ratings (Miller, 1997) or reaction times (Pisoni & Tash, 1974). These responses would correspond to patterns where either the midpoint of the continuum (i.e., the subject's category boundary) had a maximal phase-delay (an *ambiguous-maximal* response), or where the two endpoints had maximal phase-delay (an *endpoint-maximal* response). These patterns were measured by fitting quadratic functions with a midpoint fixed at the VOT category boundary.

Note that this is not taken to be a measure corresponding directly to the listener's behavioral categorization responses. A sigmoid with an inflection point at the listener's category boundary would be the corresponding function for that and would measure something different from the quadratic model. However, this would be similar to a linear response, making it difficult to distinguish between cue- and category-level representations. Because of this, we did not want to rely on a sigmoidal response as a strong indicator of a category-level effect. A non-monotonic quadratic effect, however, can only be indicative of category-level information. Thus, we focus on quadratic effects centered on listeners' category boundaries as a strong test of category-based responses. Schematic functions for each type of effect are shown in the Figures 2A and 3A.

At each time point from 72 to 264 ms after stimulus onset (9 time points), the voxel with the maximum phase delay in each ROI was identified, and t-tests were performed to determine whether the fit to either function deviated from zero at that time, indicating effects of either cue encoding (a significant linear trend) or category processing (a significant quadratic trend). Corrections for multiple comparisons were made based on random field theory, using methods previously applied to EROS data (Gratton, 2000).

EEG data were analyzed using the EEGLAB (Delorme & Makeig, 2004) and ERPLAB (Lopez-Calderon & Luck, 2014) toolboxes in MATLAB. ERP waveforms were created, time-locked to stimulus onsets with a pre-stimulus baseline of 200 ms. Peak-to-peak artifact rejection (150 μ V threshold) was used to remove trials containing oculomotor and other artifacts; data from two participants were excluded from the ERP analysis due to a high number of artifacts (>50% of the trials). Mean N1 amplitude as a function of VOT was calculated as the average voltage from 100 to 150 ms for electrode Fz. Statistical analyses were performed using linear mixed-effects models (Barr et al., 2013) with N1 amplitude as the dependent measure, VOT as a centered fixed effect, and by-subject random slopes for VOT. Significance was determined via chi-square (χ^2) tests, comparing a model without a VOT term to one that included the fixed effect of VOT.

Results

We measured activity in two ROIs based on previous findings suggesting cortical areas that might track continuous cues, phoneme categories, or both. First, we examined activity in the left posterior superior temporal gyrus (pSTG), which has been argued to encode both discrete phonemic differences (Chang et al., 2010; Mesgarani et al., 2014) and within-category differences (Myers et al., 2009).⁴ Second, we examined activity in the left inferior frontal gyrus (IFG), which is involved in a variety of language-related processes (Tse et al.,

2007) and has been suggested to represent graded phonological categories (Myers et al., 2009). Within these ROIs, we looked for evidence of cue-based responses, indicated by linear trends across the VOT continuum, and category-based responses, indicated by quadratic trends with the greatest phase delay at either the listener's category boundary (ambiguous-maximal) or the endpoints (endpoint-maximal). We also examined cue- and category-based effects outside these ROIs to establish the time-course over which these processes unfold in different cortical areas.

At 96-ms post-stimulus (slightly earlier than the peak of the auditory N1 response observed in the ERP data recorded at the scalp), we found a linear trend across the VOT continuum in pSTG, specifically in the planum temporale (PT; $z_{\text{peak}}=2.3$, $p<0.05$, $\delta=0.89$). Figure 2 suggests that this trend is, in fact, log-linear as a function of VOT. However, the effect is not consistent with a category-based encoding of the sounds, which would require an inflection point centered on the listener's category boundary. Indeed, the root-mean-square error (rmse) is worse for a step function mapped to the grand average response than for a linear function ($\text{rmse}_{\text{step}}=1.053$; $\text{rmse}_{\text{linear}}=0.874$).⁵ Rather, this response suggests the neural generators encoding VOT information may be oriented in such a way that we observe differences for shorter VOTs more clearly at this time and location. An alternative explanation is that pSTG does not distinguish within-category /p/ VOTs, but such an effect would be inconsistent with both behavioral (Miller, 1997) and ERP evidence (Toscano et al., 2010) that listeners are sensitive to these differences. In either case, this effect suggests that listeners encode speech in terms of acoustic cues, rather than phoneme categories, at this early stage of processing.

A significant linear trend was also observed in IFG at 96 ms post-stimulus onset, suggesting early frontal contributions to speech perception. Furthermore, we observed cue-based responses in a number of other cortical areas from 96–192 ms, including supramarginal gyrus at 120 ms, anterior STG at 144 ms, middle temporal gyrus (MTG) at 168 ms, and pre-motor areas at 192 ms. Activity in each of these areas has potential implications for models of speech processing (see Discussion below). Overall, the data indicate that sensitivity to continuous cues is widespread during early stages of perceptual processing.

Category-level effects emerged at 144 ms in anterior IFG, with stimuli at the phoneme category boundary producing a larger response (i.e., an ambiguous-maximal trend; Fig. 3; $z_{\text{peak}}=-2.9$, $p<0.05$, $\delta=1.17$). Interestingly, this response was observed in approximately the same area of IFG as the earlier, cue-level response, suggesting the possibility that the same cortical areas process different types of information at different points in time (consistent with interactive view of spoken language processing). The fact that responses were greater for ambiguous stimuli could suggest an initial uncertainty over which category the stimulus

⁴We focus on left hemisphere ROIs, as these have been most studied in language processing. However, there is evidence that right hemisphere areas are involved as well, particularly in early speech processing, and work in vision suggests hemispheric differences in spatial processing that may be relevant (Kosslyn et al., 1989). Future extensions of this work should aim to investigate these areas as well.

⁵For the step function, model responses for the *Pol* tokens (−10 and −5 ms relative VOT) were set to the value of the listeners' data at the *Pol* endpoint, and values for the /p/ tokens (+5 and +10 ms relative VOT) were set to the value at the /p/ endpoint; model response at the category boundary set to the mean of the two endpoint responses. For the linear function, values were set to those of a straight line that passed through both endpoints. Thus, for relative VOTs of −10, 0, and +10, both model functions had the same values.

corresponds to, and mirrors similar findings showing greater activation in IFG for words with ambiguous phonemes (Rogers & Davis, 2017). In addition, we found an endpoint-maximal trend in posterior IFG at 264 ms ($z_{\text{peak}}=2.7$, $p<0.05$, $\delta=1.07$). In both cases, these effects suggest that phoneme-level (or lexical-level) information is being used by listeners, but the responses themselves are not discrete. Figure 3 suggests a graded response, similar to the P3 categorization response observed by Toscano et al. (2010) in their ERP data, and by Myers et al. (2009) in their fMRI data.

There was considerable overlap in the time windows during which we observed cue- and category-level processing across different cortical areas. Cue-level effects occurred from 96–192 ms, and phonological category effects were observed from 144–264 ms (Fig. 4). Thus, listeners start to categorize the sound very rapidly after it is encoded (within 50 ms of initial cue encoding), and maintain sensitivity to continuous cues even when category-level information is taken into account. This is consistent with continuous cascade models positing that cue-level encoding is ongoing when categorization begins (McClelland & Elman, 1986; Trueswell et al., 1994), as opposed to models in which processing proceeds in discrete stages.

In addition, since these data provide fine-grained information about the time-course of processing, they are able to reveal cortical pathways over which cue-level information is represented. The time-course of cue-level effects in different cortical areas generally fits with dual-stream models of speech perception (Hickok & Poeppel, 2007), with both a ventral and dorsal stream showing sensitivity to continuous cues (Fig. 4). Specifically, the data suggest that initial effects are observed in PT/pSTG and IFG, followed by cue-level encoding in anterior STG, then MTG, and finally pre-motor areas. These pathways differ somewhat from those posited by existing models, and moreover, they suggest that sensitivity to continuous cues is maintained in areas outside primary auditory cortex, in contrast to predictions from hierarchical processing approaches that have relied primarily on fMRI data (Davis & Johnsrude, 2003).

Finally, we compared the optical data to concurrently-recorded ERP data that were collected from midline sites (Fz, Cz, and Pz). Previous work has shown that continuous changes in VOT values are reflected in the amplitude of the fronto-central N1 ERP response (Toscano et al., 2010). We replicated this result (Fig. 1E) in the current experiment, finding a main effect of VOT on N1 amplitude, both for VOTs centered on listeners' category boundaries ($b=0.049$, $SE=0.018$, $t=2.78$, $\chi^2(1)=6.83$, $p<0.01$) and raw VOT ($b=0.047$, $SE=0.015$, $t=3.12$, $\chi^2(1)=7.23$, $p<0.01$). The timing of the N1 peak is ≈ 40 ms later than the peak observed in the EROS data, similar to responses observed in previous EROS studies (e.g., Tse et al., 2007).

Discussion

The results of this experiment reveal the time-course of speech processing in different brain areas, demonstrating that gradient cue- and category-level representations are widespread during early stages of speech perception and overlap in time. The earliest activity (latency <120 ms) is consistent with results from ERP studies (e.g., Toscano et al., 2010) showing

effects of continuous acoustic cues and localizes these responses to pSTG. Later effects showing graded category activation in IFG are consistent with fMRI data (Blumstein et al., 2005); the current results demonstrate that such representations are activated as early as 144 ms post-stimulus.

In addition, responses in middle temporal and pre-motor areas continued to show sensitivity to continuous acoustic cues for another 48 ms after category-based responses are initially observed in IFG. Thus, the results demonstrate that both continuous cue values and graded phonological categories are represented in the brain, and that they both influence the processing of speech sounds, with overlapping time courses in different brain regions. The 48 ms lag between the onset of cue- and category-level effects in our data is inconsistent with claims from intracranial EEG studies that have found categorical effects at their earliest measurement times (Chang et al., 2010; Mesgarani et al., 2014).

These results help to resolve some of the controversies in the literature around the issue of whether neural responses reflect continuous or categorical processing of speech sounds. Whereas some previous studies have used methods with limited temporal (fMRI) or spatial resolution (EEG, MEG), invasive measures (ECOG), or stimulus materials that do not allow for controlled investigation of specific acoustic cues, such as connected speech or measures of overall speech intelligibility, the current study overcomes each of these issues, allowing us to see how processing of specific speech sounds unfolds in different cortical areas over time. This reveals how listeners transform continuous acoustic cues into category-based representations, preserving sensitivity to within-category differences in the process. Moreover, the results illustrate the interactive nature of spoken language comprehension, with the same cortical areas representing different types of information (cue- vs. category-level responses) at different points during processing, and substantial overlap in the time-course during which both types of representations are maintained in the brain.

Below, we discuss how these results fit in with this broader literature on the neurobiology of spoken language processing and additional questions that would allow us to understand these processes further in future studies.

Nature of responses in different cortical areas

Is it possible that phonological category information has even earlier effects? The measure we used to assess category-based representations was a quadratic function, but as noted above, a sigmoidal function centered on the category boundary would be another way—different from the quadratic responses—to measure the influence of the listener's phonological categories.⁶ Note that the effect at 96 ms in pSTG does not meet this criterion. It is possible that such an effect occurs earlier than those observed at 144 ms, but it is difficult to separate this from continuous, cue-level representations, which also vary monotonically. Thus, the category effects observed in IFG at 144 ms provide an upper bound on the time at which listeners consider information about phonological category structure; earlier effects could be possible as well.

⁶The quadratic response corresponds to a category goodness rating or reaction time measure, whereas the sigmoidal response would correspond to the listener's categorization function. Both would be different ways of measuring effects of category-level information.

Is it possible that the effects observed in PT/pSTG do not reflect encoding in terms of acoustic cues or phonological categories? Indeed, previous work has demonstrated that these cortical areas are sensitive to lexical information (Myers et al., 2009; Gow & Olsen, 2015); however, those studies are based on fMRI responses or aim to assess the influence of top-down effects. The additional temporal resolution afforded by the EROS technique and the task used here allows us to assess the initial state of these responses. Thus, pSTG could represent both acoustic cue and lexical information, albeit at different points during processing.

Another intriguing possibility is that listeners do encode speech sounds in terms of continuous differences, but that the representations that mediate this encoding are based on phonological features (i.e., a continuum from -voice to +voice) or lexical entries (a continuum from *beach* to *peach*) rather than acoustic cue dimensions. Because the current study used minimal pair words, such a model is a possibility. However, the effects in pSTG seem to most likely be explained simply as listeners coding differences along the VOT continuum, as both phonological and lexical representations would likely lead to a response relative to a category boundary. The quadratic responses in IFG, however, could reflect either differences along a phonological feature dimension or lexical representations, both of which could conceivably maintain sensitivity to the within-category differences observed in the current study.

Models of speech processing

Overall, the present data support models of spoken language processing that are sensitive to continuous cues and process this information continuously over time. Critically, this serves to inform models of language comprehension that make predictions about whether different levels of linguistic analysis proceed in discrete stages (Frazier & Rayner, 1982) or as a continuous cascade (McClelland & Elman, 1986; Trueswell et al., 1994). The current data provide strong support for the latter approach, with clear evidence for overlap in the types of information that are processed (both continuous cues and category-level representations).

Further evidence in favor of interactive models comes from the fact that anterior IFG showed evidence of acoustic cue encoding at 96 ms and category-level responses at 144 ms. Such a result would be inconsistent with a modular system in which distinct cortical areas corresponded to specific levels of perceptual and linguistic analysis, but it is consistent with interactive models that allow for information across different levels of processing to interact. These results also fit with recent work by Xie and Myers (2018), which demonstrated effects of phonetic competition in IFG, suggesting that activation in this area is involved in multiple levels of language processing.

These data also inform models that make predictions about the underlying neurobiological processes involved in spoken language comprehension. A number of models have been proposed, relying primarily on fMRI work, lesion studies, and animal data to make predictions about cortical pathways involved in speech processing (Davis & Johnsrude, 2003; Hickok & Poeppel, 2007; Rauschecker & Scott, 2009). The current results support several aspects of these models, such as parallel activation of ventral and dorsal streams, but they are inconsistent with some of the specific pathways that have been proposed, such as a

mapping from posterior to anterior MTG (Hickok & Poeppel, 2007). We find earlier effects in anterior MTG, suggesting a different pathway. This difference is likely due to the additional temporal resolution afforded by the EROS technique, which allows us to study the time-course of processing more directly.

Moreover, previous work has suggested that posterior STG (Chang et al., 2010) and anterior temporal lobe areas (Liebenthal et al., 2000) are involved in categorical perception of speech, yet we see evidence that continuous acoustic cues are encoded in these areas. In addition, a number of models have proposed that representations become more abstract as they progress through hierarchically more complex stages of processing (Davis & Johnsruide, 2003; Okada et al., 2010; Scott & Johnsruide, 2003; Di Liberto et al., 2015). Again, those models have been supported primarily by fMRI data that lack sufficient temporal resolution to make claims about early stages of perceptual encoding, and therefore, may reflect later-occurring effects of phonological categories. Di Liberto et al. (2015) did investigate hierarchical processing using an EEG technique, offering good temporal resolution; however, their analysis of differences in hierarchical classification over time focused on category-level information (distinctive features and phonemes) rather than investigating the maintenance of continuous cue representations. Moreover, these studies have generally focused on acoustic manipulations that affect overall speech intelligibility or measured responses in connected speech, rather than differences along specific acoustic cue continua.

Our focus here on a phonologically-relevant acoustic dimension, coupled with the additional temporal detail provided by the EROS technique, leads to a different conclusion: representations based on continuous cues are widespread in the brain, and they are maintained for long periods of time in different cortical areas. The fact that these representations persist even as category-level effects begin to emerge argues against a hierarchical approach in which representations become more categorical as they spread from primary auditory areas.

The widespread cue-level activation in different cortical areas also has implications for models of speech processing, particularly in light of fMRI data suggesting the types of processing that take place in these areas. For instance, activation in left SMG may reflect lexical processing (Gow, 2012), and possible contributions from left AG at this time may reflect segmentation processes (Burton, Small, & Blumstein, 2000). It is worth noting that these processes—even if they correspond to the levels of processing observed in previous studies— maintain sensitivity to within-category phonetic differences.

Lastly, the activity in pre-motor areas at 192 ms may have implications for gestural/ articulatory (Liberman & Mattingley, 1985; Fowler, 1984) vs. auditory (Diehl, Lotto, & Holt, 2004;) models of speech perception. On the one hand, activation corresponding to differences along a continuous acoustic dimension (VOT) could be taken as evidence in favor of auditory models (or “double-weak” models that are less tied specifically to auditory representations; Nearey, 1997). On the other hand, because these responses are observed in premotor areas, it could be taken as evidence favoring gestural models. Under either type of model, however, the data suggest that sensitivity to fine-grained phonetic differences is

maintained in these representations. Thus, regardless of whether these representations are auditory or articulatory in nature, they are specified in more detail than discrete phonological features.

Advantages and disadvantages of the EROS technique

As noted above, the EROS technique affords certain advantages for addressing the specific questions posted in the current study. In particular, because it provides good spatial and temporal resolution simultaneously, we can localize responses corresponding to previously observed early ERP effects (e.g., N100 responses; Toscano et al., 2010) and investigate how processing unfolds over time in different cortical areas. We view the EROS responses observed in pSTG at approximately 100 ms as complimentary to the ERP responses observed slightly later. With respect to the latency difference between the two responses, we have consistently noted that EROS latency leads that of corresponding ERP components (cf. Tse et al., 2007, 2012, examining EROS with N400 and MMN responses). A possible explanation is that the neurons contributing to the two signals may be different, with both interneurons and pyramidal cells contributing to EROS and only pyramidal cells (activated later) contributing to ERPs. In addition, EROS is not constrained by dipole orientation but ERPs are. Other partial summation/orientation effects are also possible (e.g., activity that extends from gyri to sulci, with slightly different latencies, and possibly being weighted differently by the two measures).

EROS does have some disadvantages compared to other cognitive neuroscience techniques as well. For instance, although the temporal resolution is quite good, it trades off against the density of spatial array used. In the high-density montage used in the current study, the sampling period is 24 ms, which while sufficient for our questions, is much lower than that of EEG recording. More notably, EROS is limited to measurement of more superficial cortical areas. Thus, for example, we cannot detect activation differences deep within Heschl's gyrus. Overall, however, EROS provides distinct advantages for addressing the current set of questions and the method serves to complement findings from other cognitive neuroscience techniques.

Conclusions

Together, these results suggest that we must revisit the assumptions made in existing models of spoken language processing using data that carefully map out the time-course of processing. This study provides strong support for models that emphasize the prominent role of continuous acoustic information in speech, and it demonstrates that both acoustic cue and category-based representations are activated in parallel during spoken word recognition.

Acknowledgements

We would like to thank Jian Huang, Erika Hussey, Eric Holmes, Alex Senetar, Justin Gumina, and Heather Mauch for assistance with data collection; Edward Wlotko for assistance with the experiment design; and Ed Maclin for assistance with data analysis. Funding for this project was provided by a University of Illinois Research Board Grant to SMG and JCT, by a Beckman Postdoctoral Fellowship to JCT, and by NIH S10-RR029294.

References

- Barr DJ, Levy R, Scheepers C, & Tily HJ (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68, 255–278.
- Blumstein SE, Myers EB, & Rissman J (2005). The perception of voice onset time: An fMRI investigation of phonetic category structure. *Journal of Cognitive Neuroscience*, 17, 1353–1366. [PubMed: 16197689]
- Burton MW, Small SL, & Blumstein SE (2000). The role of segmentation in phonological processing: an fMRI investigation. *Journal of Cognitive Neuroscience*, 12(4), 679–690. [PubMed: 10936919]
- Chang EF, Rieger JW, Johnson K, Berger MS, Barbaro NM, & Knight RT (2010). Categorical speech representation in human superior temporal gyrus. *Nature Neuroscience*, 13, 1428–1432. [PubMed: 20890293]
- Clayards M, Tanenhaus MK, Aslin RN, & Jacobs RA (2008). Perception of speech reflects optimal use of probabilistic speech cues. *Cognition*, 108(3), 804–809. [PubMed: 18582855]
- Dehaene-Lambertz G (1997). Electrophysiological correlates of categorical phoneme perception in adults. *NeuroReport*, 8, 919–924. [PubMed: 9141065]
- Davis MH, & Johnsrude IS (2003). Hierarchical processing in spoken language comprehension. *Journal of Neuroscience*, 23, 3423–3431. [PubMed: 12716950]
- de Peralta-Menendez RG, & Gonzalez-Andino SL (1998). A critical analysis of linear inverse solutions to the neuroelectromagnetic inverse problem. *IEEE Transactions on Biomedical Engineering*, 45, 440–448. [PubMed: 9556961]
- Delorme A, & Makeig S (2004). EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics. *Journal of Neuroscience Methods*, 134, 9–21. [PubMed: 15102499]
- Di Liberto GM, O’Sullivan JA, & Lalor EC (2015). Low-frequency cortical entrainment to speech reflects phoneme-level processing. *Current Biology*, 25(19), 2457–2465. [PubMed: 26412129]
- Diehl RL, Lotto AJ, & Holt LL (2004). Speech perception. *Annual Review of Psychology*, 55, 149–179.
- Fowler CA (1984). Segmentation of coarticulated speech in perception. *Perception & Psychophysics*, 36(4), 359–368. [PubMed: 6522233]
- Frazier L, & Rayner K (1982). Making and correcting errors during sentence comprehension: Eye movements in the analysis of structurally ambiguous sentences. *Cognitive Psychology*, 14, 178–210.
- Frye RE, Fisher JM, Coty A, Zarella M, Liederman J, & Halgren E (2007). Linear coding of voice onset time. *Journal of Cognitive Neuroscience*, 19, 1476–1487. [PubMed: 17714009]
- Giard MH, Perrin F, Echallier JF, Thévenet M, Froment JC, & Pernier J (1994). Dissociation of temporal and frontal components in the human auditory N1 wave: A scalp current density and dipole model analysis. *Electroencephalography and Clinical Neurophysiology/Evoked Potentials Section*, 92, 238–252.
- Gow DW, Jr (2012). The cortical organization of lexical knowledge: a dual lexicon model of spoken language processing. *Brain and Language*, 121(3), 273–288. [PubMed: 22498237]
- Gow DW, Jr, & Olson BB (2016). Sentential influences on acoustic-phonetic processing: A Granger causality analysis of multimodal imaging data. *Language, Cognition & Neuroscience*, 31(7), 841–855.
- Gratton G “Opt-cont” and “Opt-3D”: A software suite for the analysis and 3D reconstruction of the event related optical signal (EROS). *Psychophysiology* 37, S44 (2000).
- Gratton G, & Corballis PM (1995). Removing the heart from the brain: Compensation for the pulse artifact in the photon migration signal. *Psychophysiology*, 32, 292–299. [PubMed: 7784538]
- Gratton G, & Fabiani M (2001). Shedding light on brain function: The event-related optical signal. *Trends in Cognitive Sciences*, 5, 357–363. [PubMed: 11477005]
- Gratton G, & Fabiani M (2003). The event-related optical signal (EROS) in visual cortex: Replicability, consistency, localization and resolution. *Psychophysiology*, 40, 561–571. [PubMed: 14570164]

- Gratton G, Fabiani M, Corballis PM, Hood DC, Goodman-Wood MR, Hirsch J, Kim K, Friedman D, & Gratton E (1997). Fast and localized event-related optical signals (EROS) in the human occipital cortex: Comparison with the visual evoked potential and fMRI. *NeuroImage*, 6, 168–180. [PubMed: 9344821]
- Hickok G, & Poeppel D (2007). The cortical organization of speech processing. *Nature Reviews Neuroscience*, 8, 393–402. [PubMed: 17431404]
- Joanisse MF, Robertson EK, & Newman RL (2007). Mismatch negativity reflects sensory and phonetic speech processing. *NeuroReport*, 18, 901–905. [PubMed: 17515798]
- Khalighinejad B, da Silva GC, & Mesgarani N (2017). Dynamic encoding of acoustic features in neural responses to continuous speech. *Journal of Neuroscience*, 37(8), 2176–2185. [PubMed: 28119400]
- Kosslyn SM, Koenig O, Barrett A, Cave CB, Tang J, & Gabrieli JD (1989). Evidence for two types of spatial representations: hemispheric specialization for categorical and coordinate relations. *Journal of Experimental Psychology: Human Perception & Performance*, 15(4), 723. [PubMed: 2531207]
- Lieberman AM, Harris KS, Hoffman HS, & Griffith BC (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology*, 54, 358–368. [PubMed: 13481283]
- Lieberman AM, & Mattingly IG (1985). The motor theory of speech perception revised. *Cognition*, 21(1), 1–36. [PubMed: 4075760]
- Liebenthal E, Binder JR, Spitzer SM, Possing ET, & Medler DA (2000). Neural substrates of phonemic perception. *Cerebral Cortex*, 15, 1621–1631.
- Lisker L, & Abramson AS (1964). A cross-language study of voicing in initial stops: Acoustical measurements. *Word*, 20, 384–422.
- Lopez-Calderon J, & Luck SJ (2014). ERPLAB: An open-source toolbox for the analysis of event-related potentials. *Frontiers in Human Neuroscience*, 8, 213. [PubMed: 24782741]
- Massaro DW, & Cohen MM (1983). Categorical or continuous speech perception: A new test. *Speech Communication*, 2, 15–35.
- McClelland JL, & Elman JL (1986). The TRACE model of speech perception. *Cognitive Psychology*, 18, 1–86. [PubMed: 3753912]
- McMurray B, & Jongman A (2011). What information is necessary for speech categorization? Harnessing variability in the speech signal by integrating cues computed relative to expectations. *Psychological Review*, 118(2), 219. [PubMed: 21417542]
- McMurray B, Tanenhaus MK, & Aslin RN (2002). Gradient effects of within-category phonetic variation on lexical access. *Cognition*, 86, B33-B42. [PubMed: 12435537]
- Mesgarani N, Cheung C, Johnson K, & Chang EF (2014). Phonetic feature encoding in human superior temporal gyrus. *Science*, 343, 1006–1010. [PubMed: 24482117]
- Miller JL (1997). Internal structure of phonetic categories. *Language and Cognitive Processes*, 12, 865–869.
- Munson CM (2011). Perceptual learning in speech reveals pathways of processing. Doctoral dissertation, University of Iowa.
- Myers EB, Blumstein SE, Walsh E, & Eliassen J (2009). Inferior frontal regions underlie perception of phonetic category invariance. *Psychological Science*, 20, 895–903. [PubMed: 19515116]
- Nourski K, Steinschneider, McMurray B, Kovach C, Oya H, Kawasaki H, and Howard M (2014). Functional organization of human auditory cortex: Investigation of response latencies through direct recordings. *NeuroImage*, 101, 598–609. [PubMed: 25019680]
- Nourski KN, Steinschneider MS, Rhone AR, Oya H, Kawasaki H, Howard MA, and McMurray B (2015). Sound identification in human auditory cortex: Differential contribution of local field potentials and high gamma power as revealed by direct intracranial recordings. *Brain and Language*, 148, 37–50. [PubMed: 25819402]
- 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. [PubMed: 14642493]

- Okada K, Rong F, Venezia J, Matchin W, Hsieh I, Saberi K, Serences JT, & Hickok G (2010). Hierarchical organization of human auditory cortex: Evidence from acoustic invariance in the response to intelligible speech. *Cerebral Cortex*, 20, 2486–2495. [PubMed: 20100898]
- Parks NA, Maclin EL, Low KA, Beck DM, Fabiani M, & Gratton G (2012). Examining Cortical Dynamics and Connectivity with Concurrent Simultaneous Single-Pulse Transcranial Magnetic Stimulation and Fast Optical Imaging. *NeuroImage*, 59, 2504–3510. [PubMed: 21925608]
- Parks NA, Mazzi C, Tapia E, Savazzi S, Fabiani M, Gratton G, & Beck DM (2015). The Influence of Posterior Parietal Cortex on Extrastriate Visual Activity: A Concurrent TMS and Fast Optical Imaging Study. *Neuropsychologia*, 78, 153–158. [PubMed: 26449990]
- Pasley BN, David SV, Mesgarani N, Flinker A, Shamma SA, Crone NE, Knight RT, & Chang EF (2012). Reconstructing speech from human auditory cortex. *PLoS Biology*, 10, e1001251. [PubMed: 22303281]
- Picton TW, Hillyard SA, Krausz HI, Galambos R (1974). Human auditory evoked potentials. I: Evaluation of components. *Electroencephalography and Clinical Neurophysiology*, 36, 179–190. [PubMed: 4129630]
- Phillips C, Pellathy T, Marantz A, Yellin E, et al. (2000). Auditory cortex accesses phonological categories: an MEG mismatch study. *Journal of Cognitive Neuroscience*, 12, 1038–1055. [PubMed: 11177423]
- Pisoni DB, & Tash J (1974). Reaction times to comparisons within and across phonetic categories. *Perception & Psychophysics*, 15, 285–290. [PubMed: 23226881]
- Rauschecker JP, & Scott SK (2009). Maps and streams in the auditory cortex: nonhuman primates illuminate human speech processing. *Nature Neuroscience*, 12, 718–724. [PubMed: 19471271]
- Rogers JC, & Davis MH (2017). Inferior frontal cortex contributions to the recognition of spoken words and their constituent speech sounds. *Journal of Cognitive Neuroscience*, 29, 919–936. [PubMed: 28129061]
- Scharenborg O (2007). Reaching over the gap: A review of efforts to link human and automatic speech recognition research. *Speech Communication*, 49, 336–347.
- Scott SK, & Johnsrude IS (2003). The neuroanatomical and functional organization of speech perception. *TRENDS in Neurosciences*, 26, 100–107. [PubMed: 12536133]
- Sharma A, & Dorman MF (1999). Cortical auditory evoked potential correlates of categorical perception of voice-onset time. *Journal of the Acoustical Society of America*, 106, 1078–1083. [PubMed: 10462812]
- Sharma A, Kraus N, McGee T, Carrell T, & Nicol T (1993). Acoustic versus phonetic representation of speech as reflected by the mismatch negativity event-related potential. *Electroencephalography and Clinical Neurophysiology/Evoked Potentials Section*, 88, 64–71.
- Talairach J, & Tournoux P (1988). Co-planar stereotaxic atlas of the human brain 3 Dimensional proportional system: An approach to cerebral imaging. Stuttgart: Thieme.
- Toscano JC, McMurray B, Dennhardt J, & Luck SJ (2010). Continuous perception and graded categorization: Electrophysiological evidence for a linear relationship between the acoustic signal and perceptual encoding of speech. *Psychological Science*, 21, 1532–1540. [PubMed: 20935168]
- Trueswell JC, Tanenhaus MK, & Garnsey SM (1994). Semantic influences on parsing: Use of thematic role information in syntactic ambiguity resolution. *Journal of Memory and Language*, 33, 285–318.
- Tse C, Lee C, Sullivan J, Garnsey SM, Dell GS, Fabiani M, & Gratton G (2007). Imaging cortical dynamics of language processing with the event-related optical signal. *Proceedings of the National Academy of Sciences*, 104, 17157–17162.
- Tse C-Y, Low KA, Fabiani M, & Gratton G (2012). Rules rule! Brain activity dissociates the representations of stimulus contingencies with varying levels of complexity. *Journal of Cognitive Neuroscience*. 24, 1941–1959. [PubMed: 22452560]
- Viswanathan N, Fowler CA, & Magnuson JS (2009). A critical examination of the spectral contrast account of compensation for coarticulation. *Psychonomic Bulletin & Review*, 16, 74–79. [PubMed: 19145013]
- Whalen C, Maclin EL, Fabiani M, & Gratton G (2007). Validation of a method for coregistering scalp recording locations with 3D structural MR images. *Human Brain Mapping*, 29, 1288–1301.

Xie X, & Myers E (2018). Left Inferior Frontal Gyrus Sensitivity to Phonetic Competition in Receptive Language Processing: A Comparison of Clear and Conversational Speech. *Journal of Cognitive Neuroscience*, 30(3), 267–280. [PubMed: 29160743]

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Highlights:

- Fast optical imaging reveals that listeners encode continuous acoustic cues in PT/pSTG at early stages of speech perception (96 ms post-stimulus)
- This sensitivity is maintained in multiple cortical areas over approximately the first 200 ms of processing
- Speech sounds are rapidly grouped into phonological categories within 50 ms of initial cue encoding, localized to IFG (144 ms post-stimulus)
- Both cue- and category-based representations overlap in time across different cortical areas during early stages of speech processing

Figure 1.

(A) Hypothetical response function that would be obtained from a human listener when presenting sounds varying in VOT. As VOT values increase, listeners are more likely to report a voiceless (/p/) percept. Spectrograms show onsets of stimuli used in the experiment, with the period of aspiration at onset (which determines the VOT) highlighted. (B) The actual behavioral response functions obtained from listeners in the experiment. Note that, for each listener, a pre-test was conducted so that stimuli were centered on their category boundary (50% point) ± 10 ms VOT. (C) Locations of infrared sources (red) and detectors (blue) on the scalp (top), and heat map showing average cortical coverage obtained from the montage across the participants in the experiment (bottom). Hot colors represent areas for which the optical montage provided greatest sensitivity (i.e., the greatest number of overlapping paths from source-detector pairs). The montage provided good coverage over the ROIs used in the study. (D) Grand average ERP waveforms at Fz as a function of VOT relative to each listeners' category boundary (relative VOT). (E) Mean N1 amplitude (mean voltage at Fz from 100–150 ms post-stimulus, chosen to capture the N1 and minimize overlap with the P2; cf. Toscano et al., 2010) as a function of relative VOT. As VOT increased, N1 amplitude decreased.

Figure 2.

(A) Sagittal surface projection showing cortical areas with neural activity reflecting a linear trend across the VOT continuum at 96-ms post-stimulus onset. Phase delay increased linearly with increasing VOT in PT/pSTG, and decreased linearly with increasing VOT in IFG. Coronal slice corresponds to location of the voxel in pSTG ROI (green box) with the largest linear trend. Colors indicate significant z-scores, indicative of either a negative (blue colors) or a positive (yellow-red colors) correlation between phase delay and VOT. Dark grey areas on brain images reflect cortical areas sampled by recording montage. (B) Sagittal surface projections for each time sample starting at 72 ms post-stimulus onset to 264 ms after onset. Significant linear trends are observed in different brain areas starting at 96 ms and continuing until 192 ms. (C) Mean phase delay as a function of VOT (relative to each listener's category boundary) at 96-ms post-stimulus onset in the left pSTG ROI shown in Panel A. 0 ms relative VOT corresponds to the listener's category boundary. Error bars indicate standard error.

Figure 3.

(A) Sagittal surface projection showing effects related to phonological category structure (i.e., a quadratic, ambiguous-maximal trend) in anterior portions of left IFG at 144 ms post-stimulus onset. Specifically, greater activity was observed for stimuli at the listener's category boundary, relative to the continuum endpoints. Coronal slice corresponds to location of peak voxel phase delay in IFG ROI (green box). Colors indicate significant z-scores for trends with greater activity for ambiguous VOTs (blue colors) or for unambiguous, endpoint VOTs (yellow-red colors). Dark grey areas indicate cortical areas sampled by recording montage. (B) Same as Panel A for the time sample at 264-ms post-stimulus onset. Here, a significant category effect is observed such that there is a greater phase delay for continuum endpoints (similar responses for unambiguous /b/ and /p/ exemplars) than for ambiguous VOTs. (C) Sagittal surface projections for each time point from 72 to 264 ms post-stimulus. (D) Mean phase delay as a function of VOT at 144 ms post-stimulus for the left IFG ROI (green box in Panel A). (E) Mean phase delay as a function of VOT at 264 ms post-stimulus for the left IFG ROI (green box in Panel B).

Figure 4.

Time-course of processing for cue- and category-level responses. Significant linear trends across the VOT continuum are observed in different cortical areas from 96 to 192 ms post-stimulus onset. This activity is first observed in PT/pSTG at 96 ms, anterior STG at 120 ms, posterior and anterior MTG at 168 ms, and pre-motor areas at 192 ms. Category-based responses are observed in anterior IFG starting at 144 ms, and in posterior IFG at 240 ms.