The acoustics of speech are characterized by immense variability. Individual speakers differ in how they produce words, and even the same speaker will produce different acoustic patterns across repetitions of a word. Despite this variability, listeners can accurately recognize speech. Thus, a central question in spoken-language comprehension is how listeners transform variable acoustic signals into less variable, linguistically meaningful categories. This process is fundamental for basic language processing, but is also relevant to other areas, such as language and reading impairment (Thibodeau & Sussman, 1979; Werker & Tees, 1987) and automatic speech recognition.

Speech perception has been framed in terms of two levels of processing (Pisoni, 1973): the perceptual encoding of continuous acoustic cues and the subsequent mapping of this information onto categories such as phonemes or words. Theories of speech perception differ in the nature of representations at both levels and in the transformations that mediate them (Goldinger, 1998; Liberman & Mattingly, 1985; Oden & Massaro, 1978). Historically, a dominant question was whether perception is graded or categorical (discrete, nonlinear) with respect to the continuous input (Liberman, Harris, Hoffman, & Griffith, 1957; Schouten, Gerrits, & van Hessen, 2003). Discreteness could arise from several sources: an inherent, nonlinear encoding of speech into articulatory gestures (Liberman & Mattingly, 1985), the learned influence of phonological categories (Anderson, Silverstein, Ritz, & Jones, 1977), or discontinuities in low-level auditory processing (Kuhl & Miller, 1975; Sinex, MacDonald, & Mott, 1991).

If perception is nonlinear in one of these ways, listeners will be less sensitive to differences within a category than to differences between categories (or completely insensitive to differences within a category). Consider voice-onset time
Continuous Perception and Graded Categorization

(VOT), the time difference between the release of constriction and the onset of voicing. VOT leaves an acoustic trace that serves as a continuous cue distinguishing voiced (\(/b/, /d/, /g/\)) from voiceless (\(/p/, /t/, /k/\)) stops. If perception of VOT is categorical, then VOTs between 0 and 20 ms (e.g., /b/) may be encoded as more similar to each other than to VOTs greater than 20 ms (e.g., /p/), even if the acoustic distance between them is the same.

Early behavioral work suggested that perception is categorical: Listeners are poor at discriminating acoustic differences within the same category and good at discriminating equivalent distances spanning a boundary (Liberman et al., 1957; Repp, 1984). Such findings supported claims that early perceptual processes encode speech in terms of categories and abstract away from fine-grained detail in the signal. Subsequent research challenged this latter claim, demonstrating that within-category differences are discriminable (Carney, Widin, & Vieville, 1977; Massaro & Cohen, 1983; Pisoni & Tash, 1974) and that such differences are meaningful: Phonemic categories (McMurray, Aslin, Tanenhaus, Spivey, & Subik, 2008; Miller, 1997) and lexical categories (Andruski, Blumstein, & Burton, 1994; McMurray, Tanenhaus, & Aslin, 2002) show a graded structure that is sensitive to within-category distinctions.

Thus, the issue of within-category sensitivity has been settled: Listeners are sensitive to fine-grained acoustic information. However, it is not known whether perceptual encoding itself is linear or nonlinear, a critical distinction for determining what information listeners have access to when dealing with variability in the speech signal. To assess the linearity of perceptual encoding, one must examine the perceptual representations of acoustic cues. Perhaps these vary linearly with the acoustic input, and category boundaries are established at a later stage. Alternatively, perceptual encoding may be nonlinear for one of the reasons mentioned earlier.

This question is difficult to answer with behavioral techniques because they reflect the combined influence of perceptual and categorization processes and are sensitive to task characteristics (Carney et al., 1977; Gerrits & Schouten, 2004; Massaro & Cohen, 1983). Some studies have shown that discrimination is independent of categories, but this has been observed only with unnatural tasks, not in situations that reflect real-world language processing (Gerrits & Schouten, 2004; Massaro & Cohen, 1983; Schouten et al., 2003).

Measurements of neural activity may allow researchers to observe perceptual representations more directly. Blumstein, Myers, and Rissman (2003), for example, used functional magnetic resonance imaging to assess within-category sensitivity, but differences were examined with respect to the phonological category, rather than raw VOT. Thus, this study does not address the question of perceptual encoding.

Event-related potentials (ERPs) offer a useful tool for isolating perceptual activity from categorization during real-time processing. Numerous ERP experiments have used the mismatch negativity (MMN) as a measure of either change detection or discrimination (Näätänen & Picton, 1987; Pratt, in press), but they have obtained conflicting results. Several studies have suggested that phonological categories influence the MMN (Dehaene-Lambertz, 1997; Sams, Aulanko, Aaltonen, & Näätänen, 1990; Sharma & Dorman, 1999), whereas other studies have suggested that they do not (Joanisse, Robertson, & Newman, 2007; Sharma, Kraus, McGee, Carrell, & Nicol, 1993). More important, the MMN is defined as a difference between responses to a rare stimulus and a frequent stimulus and therefore requires the use of contrived tasks that make it difficult to assess how each stimulus is represented independently.

Neurophysiological work has also examined responses to individual stimuli, allowing a better comparison to natural language processes. Many of these studies suggest discontinuous encoding of continuous cues (Sharma & Dorman, 1999; Sharma, Marsh, & Dorman, 2000; Steinschneider, Volkov, Noh, Garell, & Howard, 1999). Studies measuring the auditory N1 ERP component have found a single peak for short VOTs and a double peak for long VOTs. However, if the first peak is driven by the release burst and the second peak by voicing onset, the two peaks may merge when they occur close in time (i.e., at short VOTs; Sharma et al., 2000). Frye et al. (2007) reported a single peak for the M100 magnetoencephalogram component (an analogue of the N1) for both short and long VOTs, a result consistent with continuous encoding. However, they examined only four stimulus conditions, making it difficult to assess whether the response is linear across the entire VOT continuum and to rule out variation between participants’ categories as a source of this result. Thus, the N1 may be sensitive to within-category differences, but observing this sensitivity may be difficult in stimuli with a high-amplitude burst.

Therefore, it is not known whether there is a level of processing at which speech cues are encoded linearly. Data showing sensitivity to fine-grained detail at later stages do not address this issue per se, and the neural evidence regarding early processing has been inconclusive.

We assessed sensitivity at perceptual levels to determine if we could find a linear relationship between an acoustic cue (VOT) and brain responses. We used the fronto-central auditory N1, which has been shown to respond to a wide range of stimulus types (Näätänen & Picton, 1987), as a measure of perceptual-level processing. This component is generated in auditory cortex within Heschl’s gyrus approximately 50 ms after the initial response of primary auditory cortex (Pratt, in press) and, thus, originates late in perceptual processing but early in language processing. Our stimuli did not contain high-amplitude bursts, thereby minimizing the problem of overlapping N1s described by Sharma et al. (2000).

We also assessed gradience at the level of phonological categories using the P3 component, which has been shown to reflect categorization in a number of domains, including speech (Maïste, Wiens, Hunt, Scherg, & Picton, 1995), and should reflect phonological categorization of the stimuli. This measurement was intended to confirm behavioral experiments’
suggestion that fine-grained detail is preserved until postperceptual stages in language processing.

We expected that if listeners are sensitive to within-category acoustic variation, we would observe this sensitivity in both the N1 and the P3 components. More important, if listeners encode perceptual information linearly at early stages of processing, the N1 would not show effects of phonological category information or auditory discontinuities.

Method

Design

Participants performed an auditory oddball task in which they heard four equiprobable words (beach, peach, dart, tart) over Sennheiser (Old Lyme, CT) 570 headphones while we recorded ERPs from scalp electrodes. VOT was manipulated between 0 ms (prototypical for beach and dart) and 40 ms (prototypical for peach and tart) in nine steps. Each word was designated the target in a different block. Participants were instructed to press one button for the target word and a different button for any other stimulus. Thus, approximately 25% of the stimuli were categorized by participants as targets (sufficient to produce a P3 wave), but the actual probability of the target category depended on the participant’s VOT boundary and generally varied between 17% and 33% across participants and continua (see Fig. S1 in the Supplemental Material available online).

Behavioral task

In each of the four blocks, one of the two continua was task relevant, and the other was task irrelevant, depending on which word was designated the target for that block (e.g., when dart was the target, the dart-tart continuum was task relevant, and the beach-peach continuum was task irrelevant). Block order varied between participants with the requirement that the same continuum could not be task relevant on successive blocks.

A gamepad recorded behavioral responses. Participants pressed one button with either their left or their right hand (alternated as participants were run) to make a “target” response and another button with the opposite hand to make a “nontarget” response. Eighteen practice trials were presented at the beginning of each block. Participants were given the opportunity to take a short break every 35 trials, and there was a longer break halfway through the experiment. In total, 630 trials (not including practice trials) were presented in each block, and each of the nine steps of the two continua was presented approximately 35 times.

After the main experiment, participants performed a two-alternative, forced-choice labeling task in which they categorized each token of each continuum as starting with “b” or “p” (for the beach-peach continuum) or as starting with “d” or “t” (for the dart-tart continuum). Each continuum was presented in a separate block, and stimuli were randomly presented six times within each block. We computed participants’ boundaries by fitting logistic functions to these data.

Participants

Participants were recruited from the University of Iowa community, provided informed consent, and were compensated $8 per hour. The final sample included 17 participants (12 female, 5 male; approximate age range: 18–30 years). Data from 3 of these were excluded from the P3 analyses because of problems with the labeling task that was run at the end of the experiment. All participants were included in the N1 analyses, as they did not rely on this task.

Stimuli

Stimuli were constructed using the KlattWorks front-end (McMurray, 2009) to the Klatt (1980) synthesizer. Stimuli began with a 5-ms burst of low-amplitude frication. To create the VOT continua, we cut back AV (amplitude of voicing) in 5-ms increments and replaced it with 60 dB of AH (aspiration). All other parameters were constant across VOT steps. For the beach-peach continuum, F1, F2, and F3 transitions had rising frequencies, and for the dart-tart continuum, F2 and F3 transitions had falling frequencies, and F1 had a rising frequency. Formant frequencies for vowels were based on spectrographic analysis of natural tokens.

Electroencephalogram (EEG) recordings

ERPs were recorded from standard electrode sites over both hemispheres (International 10-20 System sites F3, F4, Fz, C3, Cz, C4, P3, Pz, P4, T3, T4, T5, and T6), referenced to the left mastoid during recording, and rereferenced offline to the average of the left and right mastoids. Horizontal electrooculogram (EOG) recordings were obtained using electrodes located 1 cm lateral to the external canthus for each eye, and the vertical EOG was recorded using an electrode beneath the left eye. Impedance was 5 kΩ or less at all sites. The signal was amplified using a Grass (West Warwick, RI) Model 15 Neurodata Amplifier System with a notch filter at 60 Hz, a high-pass filter at 0.01 Hz, and a low-pass filter at 100 Hz. Data were digitized at 250 Hz.

Data processing

Data were processed using the EEGLAB toolbox for MATLAB (Delorme & Makeig, 2004). Trials containing ocular artifacts, movement artifacts, or amplifier saturation were rejected. Artifact rejection was performed in two stages. In the first stage, trials were automatically marked if they contained voltages that exceeded a threshold of 75 µV in any of the EOG channels or 150 µV in any of the EEG channels. In the second stage, the data were visually inspected, and trials with any
additional artifacts were rejected. The baseline for each epoch was computed as the mean voltage 200 ms before the onset of the stimulus.

**Results**

**Behavioral responses**

Participants’ behavioral responses reflected standard categorization functions for both continua, though boundaries were affected by whether the target began with a voiced or voiceless consonant (Fig. 1a). Participants’ responses in the labeling task performed after ERP recording also reflected standard categorization functions (Fig. 1b). (See Additional Results in the Supplemental Material available online.)

**N1 amplitude**

N1 amplitude was measured as the mean voltage from 75 to 125 ms poststimulus, averaged across the three frontal channels (Figs. 2a and 2b). N1 amplitude decreased with increasing VOT, and this effect was observed for both the relevant and the irrelevant continua.

The data were analyzed with a linear mixed-effects model (LMM) using the lme4 package in R (Bates, 2005); the within-subjects factors of VOT, stimulus continuum (beach-peach or dart-tart), target voicing (voiced or voiceless), and task relevance (relevant or irrelevant) were treated as fixed effects (see Additional Results in the Supplemental Material for results of analyses of variance). In all LMM analyses reported, participant was entered as a random effect, and the Markov chain Monte Carlo (MCMC) procedure was used to estimate $p$ values for the coefficients. In the analysis of N1 amplitude, the main effect of VOT was significant ($b = 0.215, p_{\text{MCMC}} < .001$), which confirmed our prediction that VOT would affect the magnitude of the N1 (see Figs. 2a and 2b). The main effect of stimulus continuum was also significant ($b = 0.744, p_{\text{MCMC}} < .001$), with N1 amplitudes being larger for beach-peach than for dart-tart (see Fig. 2b). Thus, the N1 encodes not only VOT, but also differences in acoustic information more broadly. All other main effects and interactions were nonsignificant.

We next asked whether the effect of VOT on N1 amplitude could be fit just as well by a categorical model in which the category boundary varied across individuals (which could produce results mimicking linearity across participants). We directly compared two mixed-effects models relating N1 amplitude to VOT (similar to the two models used by McMurray, Tanenhaus, & Aslin, 2009): a linear model defined by two parameters (slope and intercept) and a categorical model defined as a step function with three parameters (the lower bound, the upper bound, and the crossover point). In both models, parameters were fit to each participant’s data to ensure that linear results were not an artifact of averaging.

The linear model provided a better overall fit, as measured both by mean $R^2$ values (linear: 0.430; categorical: 0.343) and by the Bayesian information criterion (BIC; linear: 645.2; categorical: 745.6). BIC scores favored the linear model for 16 of the 17 participants (binomial test: $p < .001$). Thus, even though the categorical model had more free parameters, the linear model provided a better fit. These results suggest that responses to acoustic cues are predominantly linear.

The final analysis was intended to examine potential influences of phonological categories on the N1. That is, we asked whether, for a given VOT, the N1 differed on the basis of how the stimulus was classified. Thus, in addition to looking at differences between stimuli (i.e., effects of VOT), we looked at whether the way listeners categorized the stimuli (voiced or voiceless) had an effect on the N1. For this analysis, we restricted the data set to include only trials in which participants made a “target” response. When either beach or dart was the target, all the “target”-response trials were trials on which the participant categorized the stimulus as having a voiced onset; when either peach or tart was the target, the participant...
Fig. 2. N1 results: (a) grand-average event-related potential waveforms, averaged across frontal electrodes, for each voice-onset time (VOT); (b) mean N1 amplitude as a function of VOT and stimulus continuum; and (c) mean N1 amplitude as a function of VOT and target voicing for trials in which participants made “target” responses. In (b), error bars represent standard errors. In (c), the size of each data point is proportional to the number of trials for that condition.
categorized the stimulus as having a voiceless onset. This allowed us to confirm that listeners’ voicing categories for the stimuli could not account for the linear effect observed earlier, as all stimuli within a given target-voicing condition were identified as part of the same phonological category (voiced for the voiced blocks and voiceless for the voiceless blocks). The use of only trials with “target” responses meant that some conditions included more trials than others (e.g., participants indicated a voiced target on few trials with VOTs of 40 ms), so we weighted each data point by the number of trials in that condition. Figure 2c shows N1 amplitudes for the different conditions in this data set.

An LMM analysis with VOT and target voicing as fixed factors showed a significant main effect of VOT ($b = 0.317$, $p_{MCMC} < .001$); neither target voicing nor the interaction was significant. Thus, there was no evidence that phonological category information influenced N1 amplitude.

**P3 amplitude**

P3 amplitudes were measured from the average of the three parietal channels by computing the mean voltage between 300 and 800 ms after stimulus onset. Figure 3 shows the ERP waveforms as a function of VOT distance from the target (i.e., the number of VOT steps from the target) for the task-relevant continuum (e.g., beach-peach when beach was the target; Fig. 3a) and the task-irrelevant continuum (e.g., beach-peach when dart was the target; Fig. 3b). Regardless of which word served as the target, P3 amplitude was larger at the target end of the relevant continuum than at the nontarget end; P3 amplitude did not vary with distance from the target along the irrelevant continuum.

As with the N1, P3 amplitude was analyzed using an LMM with VOT, stimulus continuum, target voicing, and task relevance as fixed factors (see Additional Results in the Additional Results).
Discussion

Our results indicate that (a) both N1 and P3 amplitude reflect listeners’ sensitivity to fine-grained differences in VOT, and (b) whereas P3 amplitude is influenced by phonological categories, N1 amplitude is not. The N1 shows a one-to-one correspondence with VOT even when participants indicate that stimuli belong to different phonological categories and when differences in individual category boundaries are accounted for. Further, this effect is not specific to VOT, as N1 amplitude varies with place of articulation as well.

This experiment provides strong evidence that nonphonological representations of speech are maintained until late (> 100 ms) stages of perceptual processing and that listeners encode acoustic cues linearly prior to categorization. Our results fit with the hypothesis that speech perception is fundamentally continuous (Massaro & Cohen, 1983) and that effects of phonological information are a product of categorization and task demands, not perceptual encoding.

This conclusion contrasts with earlier claims that the morphology of the N1 reflects categorical perception (Sharma & Dorman, 1999; Sharma et al., 2000). However, as we noted earlier, differences in the construction of stimuli allowed us to observe effects that may have been masked in previous studies. Our study also contrasts with work suggesting an auditory discontinuity in VOT encoding near the phonological boundary (Kuhl & Miller, 1975; Sinex et al., 1991), which would lead to a nonlinearity in perceptual encoding. However, the evidence for such a fixed discontinuity is mixed, with estimates of its location ranging from 20 to 70 ms, depending on the specific characteristics of the stimuli and range of VOTs tested (Ohlemiller, Jones, Heidbreder, Clark, & Miller, 1999). Further, given that listeners must use VOT information flexibly in both developmental time (because the VOT categories for a particular language must be learned) and real-time speech processing (e.g., because of variation in speaking rate), such an auditory discontinuity could make speech perception a much more difficult task.

The P3 results demonstrate that graded acoustic detail is also preserved at postperceptual levels. The P3 occurs too late (~450 ms) to be an indicator of phonological processing per se (though we refer to phonological categories here, because they were the relevant distinction in this task). Thus, acoustic detail is maintained even at postphonological stages, which is consistent with behavioral and neuroimaging work showing graded phonetic categorization (Andruski et al., 1994; Blumstein et al., 2003; McMurray et al., 2002, 2009).

These results offer a basis for examining the nature of early processing of acoustic cues, and current work is extending this approach to other cues. The results may also have practical implications. Work on specific language impairment and dyslexia has suggested that impaired listeners show less categorical perception than nonimpaired listeners (Thibodeau & Sussman, 1979; Werker & Tees, 1987). However, if perceptual encoding is continuous and categorical effects emerge as an
effect of task differences, researchers may need to use other measures as a benchmark for understanding speech perception in this group (e.g., McMurray, Samelson, Lee, & Tomblin, 2010).

Together, the N1 and P3 results support a model of spoken-word recognition in which perceptual processing is continuous and categorization is graded. More important, this linear encoding of the acoustic input is exactly what is needed for processes that use such detail to facilitate language comprehension (Goldinger, 1998; McMurray et al., 2009). These results support an emerging view that language processing is based on continuous and probabilistic information at multiple levels.

Acknowledgments
We would like to thank Margot Schneider and Kelli Kean for assistance with data processing and Susan Wagner Cook for help with the statistical analyses.

Declaration of Conflicting Interests
The authors declared that they had no conflicts of interest with respect to their authorship or the publication of this article.

Funding
This research was supported by National Institutes of Health Grant DC008089 to B.M.

Supplemental Material
Additional supporting information may be found at http://pss.sagepub .com/content/by/supplemental-data

Notes
1. We use the term perceptual encoding to refer to the process by which continuous acoustic information (e.g., a particular voice-onset time) is converted to a representation usable by the system. Categorization refers to the process that uses the information provided by perceptual encoding to identify a phonological category.
2. Using the M100 magnetoencephalography component, Phillips et al. (2000) also obtained results suggestive of this conclusion.
3. A bug in the randomization software prevented this value from being perfectly equivalent across conditions. The average standard deviation in the number of repetitions was 4.3.
4. The model coefficients we report are unstandardized, so the numbers reflect values in microvolts per unit of the factor.
5. The linear model also showed a better fit for each stimulus continuum individually.
6. The unweighted model produced the same pattern of results.
7. Some participants had only three steps on one side of the category boundary for one continuum. Thus, each waveform contains data from every participant, but some participants contributed more data to the ±4 conditions than others did.
8. An analysis including the extreme rVOT values still produced a marginal main effect of rVOT ($b = 0.249$, $P_{MCMC} = .054$), with no other significant effects.

References
Experimental Psychology: Human Perception and Performance, 34, 1609–1631.