

Probabilistic Constraint Satisfaction at the Lexical/Phonetic Interface: Evidence for Gradient Effects of Within-Category VOT on Lexical Access

Bob McMurray,^{1,3} Michael K. Tanenhaus,¹ Richard N. Aslin,¹ and Michael J. Spivey²

Research in speech perception has been dominated by a search for invariant properties of the signal that correlate with lexical and sublexical categories. We argue that this search for invariance has led researchers to ignore the perceptual consequences of systematic variation within such categories and that sensitivity to this variation may provide an important source of information for integrating information over time in speech perception. Data from a study manipulating VOT continua in words using an eye-movement paradigm indicate that lexical access shows graded sensitivity to within-category variation in VOT and that this sensitivity has a duration sufficient to be useful for information integration. These data support a model in which the perceptual system integrates information from multiple sources and from the surrounding temporal context using probabilistic cue-weighting mechanisms.

KEY WORDS: speech perception; spoken word recognition; invariance; lexical access; eye movements.

INTRODUCTION

The absence of an invariant mapping between the acoustic stream and plausible phonetic features that might be incorporated into lexical representations presents a major challenge for models of real-time speech perception.

This work was partially supported by NIH Grants NIDCD DC-005071 and NIDCD T30 DC00035. We would like to thank Dana Subik for help in testing participants.

¹ Department of Brain and Cognitive Sciences, University of Rochester.

² Department of Psychology, Cornell University.

³ To whom all correspondence should be addressed: Department of Brain and Cognitive Sciences, University of Rochester, Rochester, NY 14627. email: mcmurray@bcs.rochester.edu

Virtually every phonetic feature examined to date has been found to correlate with many acoustic properties. Voicing, for example, while primarily predicted by voice onset time (VOT) and aspiration, has been shown to have as many as 16 different acoustic correlates (Lisker, 1986). Thus, no single acoustic feature seems to correspond to linguistically relevant categories. Although one might be able to construct a notion of invariance around sets of such features, most acoustic/phonetic properties of speech also display a variety of context dependencies that further compound the problem. Therefore, in addition to standard voicing features such as VOT and aspiration, voicing is also determined in part by factors such as speaking rate (Summerfield, 1981), place of articulation (Lisker & Abramson, 1964), and position in a prosodic domain (Fougeron & Keating, 1997; Crosswhite *et al.*, in preparation). Thus, the speech perception system must cope with multiple graded constraints that often span large stretches of the surrounding acoustic context.

Two major lines of research have arisen in response to the problem of invariance. One approach argues that while there may be no invariant acoustic features corresponding to phonetic units, speech is characterized by invariance in the motor commands used for production (e.g., Liberman *et al.*, 1967; but see Fowler, 1991, for more recent thoughts). Thus, if the perceptual system mapped acoustic information directly onto motor representations (such as articulatory gestures), speech could then be decoded relatively easily.

The other approach to the invariance problem has been a search for alternative metrics (e.g., transformations of the acoustic signal) that might reveal *acoustic* invariance. For instance, Stevens and Blumstein, (1978) found some degree of acoustic invariance using gross spectral properties but were ultimately not successful in dealing with positional variance (e.g., identifying the same consonant in word-final vs. word-initial position). Later models that incorporated more subtle features (such as locus equations, Sussman *et al.*, 1998) proved more successful. The search for acoustic invariance has been motivated in part by advances in computational techniques that allow more fine-grained analyses of the signal and also by findings in neuroscience that suggest ways in which the brain may be sensitive to hypothesized invariant acoustic properties of the signal via transformations of the acoustic signal by the peripheral auditory apparatus (for example, Sussman *et al.*, 1998).

Research demonstrating categorical perception has been argued to support this fundamental claim that within-category variability is rapidly discarded (Liberman *et al.*, 1957). Categorical perception is marked by (a) steplike category boundaries and (b) a high correlation between discrimination and labeling performance. True categorical perception occurs when discrimination is completely predicted by categorization: Performance is at

chance for stimuli that are labeled equivalently but is well above chance only for stimuli labeled differently. Complete inability to discriminate stimuli within a phonetic category would demonstrate that the system discards all variability within a phonetic category and only retains category identity (in this case, the phoneme). Although true categorical perception has never been demonstrated, identification and discrimination studies have consistently demonstrated sharp category boundaries and high correlations between discrimination and labeling.

Ultimately, however, this search for invariance (either motor or acoustic) in the speech signal is motivated by an underlying (and often implicit) notion that the primary role of the speech perception system is to extract invariant structure, and thus to discard variation. Here, we argue for a different viewpoint: We suggest that many types of acoustic variation (if represented perceptually) would be quite helpful in decoding speech. Thus, it would be beneficial for the perceptual system to represent acoustic variation (and covariation) and use it probabilistically.

In this article we first review some of the work that supports a graded or probabilistic mapping between acoustics and phonetic features. After discussing the implications of categorical perception for this notion, we present recent findings from our own research demonstrating that (a) lexical access is sensitive to fine-grained acoustic variation within phonetic categories and (b) the system maintains sensitivity to these variations for periods of time that span several phonemes. We conclude with a discussion of the implications of these results for resolving acoustic/phonetic temporal ambiguity and long-distance acoustic/phonetic dependencies.

PROBABILISTIC CUE WEIGHTING AND CONTEXT DEPENDENCIES

Speech perception offers many examples of probabilistic cue weighting, in particular the large body of research examining the integration of multiple cues in speech perception (often termed *studies of trading relations*). These experiments primarily make use of categorization data from speech continua to show a shift in the category boundary (e.g., a voicing boundary) in response to a secondary factor ($F1$ frequency). For example, Summerfield (1981) asked subjects to categorize stimuli from two VOT continua. These continua were identical in all respects other than vowel duration, where vowel duration is an indicator of speaking rate. His results indicated that the boundary between voiced and voiceless tokens shifts as a function of vowel duration, such that longer vowels result in more voiced identifications than shorter vowels.

This method for assessing shifts in phonetic category boundaries as a result of acoustic/phonetic manipulations has become nearly universal in exploring how multiple features are combined during speech (but see Volaitis & Miller, 1992, for an interesting alternative). Results have revealed trade-offs between simultaneous cues to a single feature (e.g., *F1* and VOT for voicing; Summerfield & Haggard, 1977), effects of independent features on each other (e.g., the effect of place of articulation on voicing; Pisoni & Sawusch, 1974), coarticulation between consonants (Mann & Repp, 1981; Repp & Mann, 1982), lexical status on phonemic identity (e.g., Ganong, 1980), speaking rate on manner and voicing (Miller & Liberman, 1979; Summerfield, 1981), phonological knowledge (Massaro & Cohen, 1983a), visual-facial information (McGurk & MacDonald, 1976), and many other sorts of interactions (see McQueen, 1996, for a review).

These cue-weighting effects have typically been cast in terms of shifting category boundaries. That is, there is an underlying category boundary along some acoustic dimension that moves about in response to other acoustic cues (but cf. Oden & Massaro, 1978, and McClelland & Elman, 1986, for approaches compatible with probabilistic cue weighting). Although few experiments have looked at the integration of more than two features, it is implicit that all of these features (and many yet unstudied features) must be interacting during perception. Thus, a system for categorizing voicing would have to take into account traditional voicing features like VOT and aspiration, but also the place of articulation of the segment (velars have a different voicing category boundary than alveolars and bilabials), the length of the vowel (which would correlate with speaking rate and prosodic position), and potential coarticulatory or phonological processes (like voicing assimilation). Although one could think of each of these factors as moving the category boundary to various degrees, it seems more parsimonious to view each of these factors as affecting the probability that a particular category (e.g., voiced or voiceless) underlies a given acoustic representation. Under this probabilistic framework, these effects would be weighted to determine how they are integrated. Perception could then use these probabilities to recover the phonetic unit or word that is maximally likely given the input.⁴

Moreover, given the standard use of sigmoidal functions (e.g., the standard phoneme identification function) to map continuous variables onto probabilities, a shifting category boundary of this sort mathematically represents a simple reweighting of the probability mapping (toward one end-

⁴ Importantly, this might also motivate a model of perceptual development in which these weights or priors are the products of a statistically sensitive learning mechanism (e.g., Maye *et al.*, 2002).

point or the other). Thus, the entire body of evidence demonstrating shifting category boundaries is also compatible with probabilistic cue weighting.

Importantly, for a probabilistic cue-weighting system to succeed, information about the acoustic structure of the utterance must be stored in a graded or probabilistic manner. If the system discards fine-grained information about VOT and other features in favor of strict category information, there are no probabilities to integrate, only ones and zeros. Moreover, in a categorical system, if the information changes (e.g., when later-occurring lexical, coarticulatory, or vowel length information arrives), the system may find itself in a “phonetic garden-path” and be forced to revise earlier decisions. By preserving detail and making lexical or phonetic decisions on-line using probabilistic methods, the system would be able to more effectively integrate new information (for similar arguments, see Smits, submitted).

The idea that prelexical representations are updated probabilistically in real time has a counterpart in a large body of work on spoken word recognition (e.g., Marslen-Wilson, 1987; McQueen & Cutler, 2001). It is now well established that lexical candidates are activated probabilistically and in parallel as the speech stream unfolds. For example, at the onset of the word *beetle*, words like *beaker* and *beach* and *behind* will be simultaneously active (because at this point the acoustic/phonetic information is consistent with all of them). Lexical access is not a discrete selection process, but rather is characterized by on-line updating of probabilities given the available information. This approach is incorporated into most current models of spoken-word recognition (e.g., Elman & McClelland, 1986; Norris, 1994; Gaskell & Marslen-Wilson, 1996). Thus, these current models of spoken-word recognition are in a sense inconsistent with a “categorical” model of sublexical processing that makes discrete sublexical decisions.

In sum, a probabilistic approach to speech perception would allow temporal integration to occur efficiently and smoothly and to take advantage of acoustic information that is spread out over time. Thus, a system that is sensitive to fine-grained phonetic detail (even details that would normally be discarded during categorization) would be able to take advantage of the sort of on-line probabilistic processing we have described here. Rather than being discarded, subcategorical variation may instead be fundamental to speech processing.

CATEGORICAL PERCEPTION

To a large extent, the literature on categorical perception suggests that listeners are relatively *insensitive* to within-category subphonemic variation (Lieberman *et al.*, 1957; Sharma & Dorman, 1999). However, in at least some

tasks, subjects do have access to subphonemic acoustic information. Carney *et al.* (1973) showed that subjects could discriminate within-category stimuli in a *speeded* discrimination task. Likewise, in a 4AIX discrimination task (in contrast to the ABX task used by Liberman *et al.*, 1957⁵), Pisoni and Lazarus (1974) showed that subjects could also discriminate within-category VOT differences.

More recently, a number of studies have used goodness ratings to explore categorical perception and a range of effects on phonetic categorization. Massaro and Cohen (1983b) provided evidence favoring noncategorical perception when measuring discrimination on a same/different scale. A series of studies by Miller (Miller, 1997; Allen & Miller 1999) used goodness ratings for phoneme identity and found graded responses within a phonetic category. Importantly, throughout all of the studies, a consistent “prototype” category structure was found such that some ranges of stimuli were classified as “better” exemplars than others. This suggests a level of gradation within phonemic categories (although whether these are metalinguistic or perceptual categories is unclear). Although Miller’s studies did not test sensitivity to subphonemic differences directly, the fact that listeners assign different ratings to stimuli that are categorized the same way suggests that they are sensitive to this information.

These studies suggest that listeners are sensitive to subphonemic information but they do not conclusively demonstrate that subphonemic variation influences lexical access in spoken-word recognition for at least two reasons. First, these studies all use tasks that require off-line metalinguistic judgments about phonemes. Second, it is not clear how discrimination and goodness judgments relate to processes involved in spoken-word recognition.

The most direct evidence that sublexical variation within consonants affects lexical access comes from an important study by Andruski *et al.* (1994). Andruski *et al.* examined subphonemic effects in word recognition using a semantic priming paradigm. With stimuli that were either fully voiceless, one-third voiced (they had one-third of their voicing restored), or two-thirds voiced, they found significantly reduced priming for the two-thirds voiced condition with an interstimulus interval (ISI) of 50 ms, (e.g., *table* primed *chair* better than *table*_{2/3} did) but not with an ISI of 250 ms. Thus, Andruski *et al.* demonstrated that words with prototypical members of a phonetic category activate their lexical representations more strongly

⁵ In the more standard ABX tasks (e.g., Liberman *et al.*, 1957; Carney *et al.*, 1973), subjects hear two different sounds followed by a repetition of one them. They must decide whether the last sound is identical to the first or second one. In a 4AIX task, subjects hear two pairs of sounds (one pair is identical and one pair is different) and must decide which of the two pairs was the “different” pair.

than words with less prototypical representations that are closer to the category boundary. While this study is important in demonstrating lexical sensitivity to acoustic variation, it leaves open the question of whether lexical access exhibits a *gradient* effect of acoustic variation (because the effect could be carried by a single difference near the category boundary as opposed to a more linear relationship between VOT and lexical activation within the entire category). Moreover, it is unclear whether subphonemic sensitivity is found in the voiced end of the continuum, where the range of acceptable voiced stimuli is much smaller (e.g., Lisker & Abramson, 1964).

Thus, a growing body of evidence suggests that subjects are sensitive to within-category phonetic variation using metalinguistic tasks and that lexical activation is at least partially sensitive to within-category variability. However, it remains unclear whether lexical activation is *gradient* (i.e., whether lexical activation increases or decreases monotonically with changes in acoustic information throughout the category). Moreover, given that many sources of phonetic covariation stretch over relatively long periods of the signal (e.g., lexical information, prosodic strength, and coarticulation), it will be crucial to determine if sensitivity to gradiency is maintained long enough to accommodate the effects of long-distance phonetic dependencies.

We stress the importance of lexical activation here (as opposed to tasks requiring a phoneme decision) because most work on phonemic categorization rests on the assumption that the purpose of phonetic analysis is to identify words. Given this, it is crucial to examine lexical activation directly.

The current studies examined the effects of VOT on phoneme identification and on lexical access. Both experiments used eye movements as a dependent measure (Tanenhaus *et al.*, 1995). Eye movements were measured because a large body of evidence has demonstrated that the probability of fixating a picture of an object can be mapped onto activation from models of spoken-word recognition such as TRACE (McClelland & Elman, 1986) using a very simple linking hypothesis (Allopenna *et al.*, 1998; Dahan *et al.*, 2001; Tanenhaus *et al.*, 2000). Moreover, since eye movements driven by the auditory signal are *continuously* generated from 200 ms poststimulus, an analysis of fixations over time can reveal information about how representations change during the uptake of new acoustic information.

The first study was a partial replication of a classic categorical perception paradigm in which subjects performed a two-alternative phoneme identification task with nonword (CV) stimuli. Our addition to this paradigm is the simultaneous use of an eyetracker to provide both a detailed measure of the temporal dynamics of phoneme identification and also a measure that is sensitive to parallel activation of multiple alternatives. A primary purpose of the study was to determine whether we would replicate standard categorical identification functions using our VOT continua and to understand what

fixations might reveal in such a task. The second study examined lexical access using the same VOT continua to determine whether there is a gradient relationship between VOT and lexical access.⁶

EXPERIMENT 1: THE TEMPORAL DYNAMICS OF PHONEME CATEGORIZATION

A 9-step ba/pa VOT continuum ranging from 0 to 40 ms was created using the KlattWorks (McMurray, in preparation) interface to the Klatt (1980) synthesizer. Stimuli had identical 5 ms release bursts and rising formant trajectories. VOT was varied by cutting back the temporal onset of the amplitude of voicing (AV) parameter and replacing it with 60 dB of aspiration (AH).

Subjects heard 24 repetitions of each speech sound and categorized them by clicking on one of two “buttons” on a computer screen labeled “ba” and “pa.” In between trials, subjects fixated on a central point (to establish that they did not begin the trial looking at one or the other response region). The response buttons did not change locations between trials (so eye movements were not driven by search procedures).

During the experiment, eye movements were measured at 250 Hz. Although the eye-movement paradigm has typically been used with lexical responses (clicking on pictures), McMurray *et al.* (2000) present data suggesting that eye movements to labeled buttons correlate well with response-deadline (button-pressing) tasks.

Figure 1 shows mouse-click response data. The identification curve is quite sharp and there was a high degree of agreement between subjects on the category boundary (17.5 ms, 95% CI = $\pm .83$ ms).

For all of the analyses presented here, the data were split into two data sets straddling the 17 ms category boundary: VOTs in the /b/ range (0–15 ms) and those in the /p/ range (20–40 ms). All analyses of eye movements were performed on data that were filtered to remove trials in which subjects responded (with the mouse) with the low-frequency response (for a given VOT). For example, trials in which the subject heard a VOT of 0 ms and responded “pa” were removed from the data set. This filtering yields an essentially perfect identification function, with a step from 0% to 100% /pa/-responses between 15 and 20 ms of VOT. Any differences in eye movements are thus not due to small differences in off-line identification. Moreover, this provides the appropriately conservative test for gradient effects, because it

⁶ Complete methodological and statistical details for Experiment 1 can be found in McMurray *et al.* (in preparation). Experiment 2 is described more fully in McMurray *et al.* (2002).

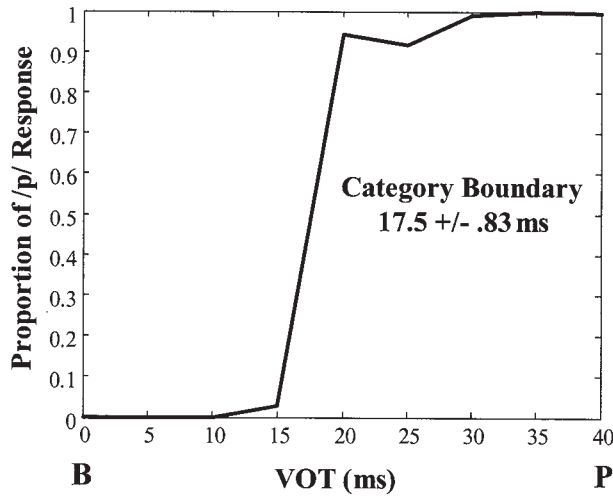


Fig. 1. Identification function (from mouse responses) as a function of VOT.

removes the data from trials where participants looked at the competitor because they labeled the input as an exemplar of the other category.

Figure 2 shows the proportion of fixations on “b” and “p” as a function of time for a VOT of 0 ms (left panel) and 40 ms (right panel). The thin curves show fixations to the target (“b” for a 0 ms VOT and “p” for a 40 ms

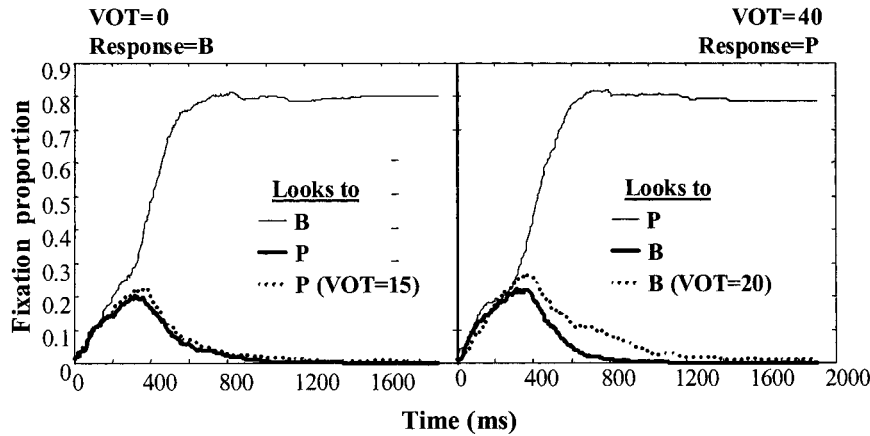


Fig. 2. Probability of a fixation as a function of time. Fixations to the target (B when the VOT was 0, or P when the VOT was 40) are shown with thin lines and to the competitor (P when the VOT was 0, or B when the VOT was 40) with thick lines. Fixations to the competitor when the VOT was near the category boundary are shown with dashed lines.

VOT). Thick curves are fixations to the competitor (“p” for a 0 ms VOT and “b” for a 40 ms VOT). Dashed lines show fixations to the competitor from VOTs close to the category boundary (“b”: 15 ms; “p”: 20 ms).

These figures show that early in the trial subjects look to both the target and competitor and that gradually subjects resolve any ambiguity by fixating consistently on the target. Moreover, there is a small, but significant, difference in the looks to the competitor near the category boundary in both cases (/b/: $F(3,48) = 2.91, p = .044$; /p/: $F(4,64) = 6.12, p < .001$), suggesting that the competitor is more active when the VOT is close to the category boundary, even for trials in which the subject responded with the “correct” label 100% of the time.

Thus, Experiment 1 shows that our stimuli replicate the classic phoneme identification findings from categorical perception (the extremely steep slope) and show excellent between-subject agreement for the category boundary. Moreover, by demonstrating a difference in the magnitude of the competitor effect (looks to the similar phoneme) between stimuli near the category boundaries and others, we have demonstrated that eye movements provide a measure of listener sensitivity to within-category differences near the category boundary (replicating Andruski *et al.*, 1994).

Experiment 2 tested whether this competitor effect holds for lexical activation, exhibits gradiency across VOTs, and lasts throughout the time-course of lexical processing.

EXPERIMENT 2: GRADIENT EFFECTS OF SUBPHONEMIC VOT VARIATION ON LEXICAL ACCESS

Experiment 2 used a more natural lexical identification task (clicking on pictures) that does not require an off-line metalinguistic judgment. Eye movements during these tasks have been shown to be sensitive to nondisplayed competitors (Dahan *et al.*, 2001; Magnuson, 2001) and lexical competitors from both lexicons in bilinguals (Spivey & Marian, 1999). Thus task-specific strategies (such as “prenaming” the stimuli) cannot account for the pattern of eye movements, and lexical processing appears to be taking place against the background of the full lexicon.

We created six 9-step b/p VOT continua ranging from 0 to 40 ms (beach/peach, bale/pail, butter/putter, bump/pump, bear/pear, and bomb/palm), along with six l- and six sh- initial filler words (lamp, ladder, lip, leg, leaf, lock, shark, shell, sheep, ship, shirt, and shoe) using the same synthesis procedure described in Experiment 1.

These stimuli were presented to subjects over headphones along with a visual display of four pictures: the “b-” and “p-” competitors (e.g. “bear” and

“pear”) and an “l-” and “sh-” filler item (e.g., “lamp” and “ship”). See Fig. 3 for an example display. Each pair of b/p pictures was paired with a single l/sh pair throughout the experiment (though this was randomized between subjects) to minimize the possibility that subjects would notice something about the b/p manipulation.

Subjects received pretraining (matching pictures with their printed names) on the items at the beginning of the experiment to ensure that subjects were comfortable using a few of the lower-frequency picture names (e.g., “putter” instead of “golf club”) and that the pictures were clear. During testing, at the beginning of each trial subjects were shown all four pictures (for that trial) for about 500 ms (to familiarize them with button locations). They then clicked on a central fixation point with the mouse and heard the name of one of the pictures. At this point their task was to click

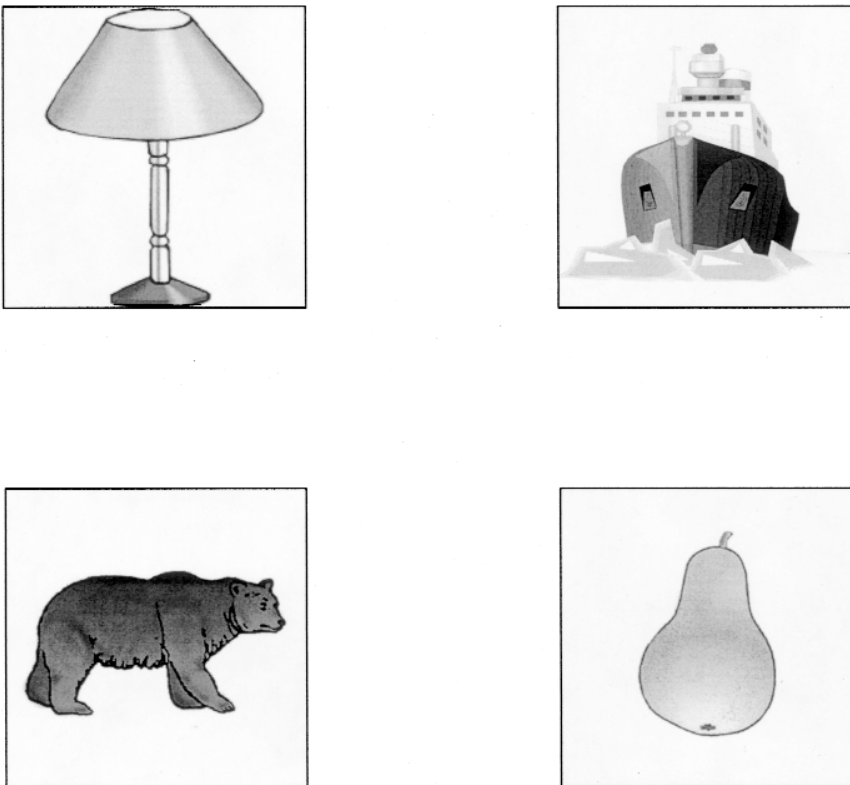


Fig. 3. A sample display. Each display contains a target (e.g., the “pear”), a competitor (e.g., the “bear”), and two filler items (“lamp” and “ship”).

on the picture corresponding to that auditory stimulus. Mouse clicks and reaction times were recorded along with eye movements (which were monitored at 250 Hz throughout the experiment).

Figure 4 displays the mouse-click data as a function of VOT for this experiment as well as for Experiment 1. There was considerable agreement between subjects and experiments over the category boundary (Exp. 2: 17.25, 95% CI = ± 1.33 ms; $t(32) = 0.55$, $p > .1$), suggesting that the synthesis methods used in the experiments were comparable. However, the slope of the identification function in this lexical task was considerably shallower than in the phoneme identification task ($t(32) = 9.36$, $p < .001$). This suggests that this simple change in task dramatically altered the degree of gradiency in the identification function: two alternative forced-choice metalinguistic tasks (e.g., the classic phoneme identification task) may underestimate such gradient effects.

Because we used six different items, there is a chance that the shallower slope of the mouse-identification function was due to variation in the category boundary between items. To assess this, we computed slope and category boundary of the identification function for each item within each subject. A one-way analysis of variance revealed significant differences in category boundary between items ($F(5,80) = 8.0$, $p < .001$). Post-hoc tests revealed that this difference was due to the height and/or frontness of the vowel: Beach, bale, and bear, items with high/front vowels, had more /p/ responses

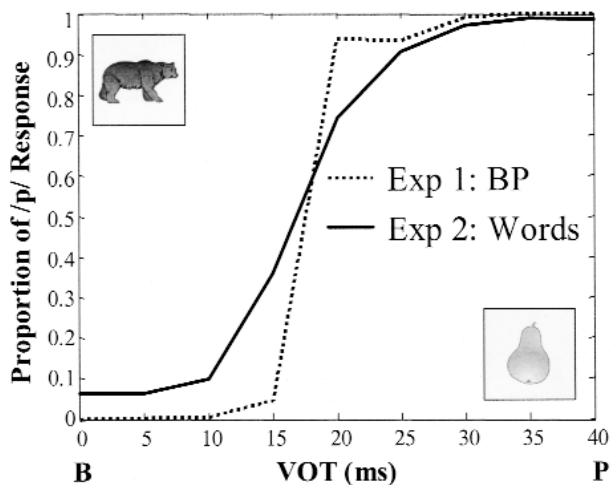


Fig. 4. Identification functions (mouse responses) for Experiments 1 and 2. Although the category boundaries were the same, a significant difference between slopes was found: The lexical identification task produced a much shallower slope than the phoneme identification task.

than bomb, butter, and bump, those with low/back vowels (High vs. Low: $F(1,16) = 27.1, p < .001$; Within High: $F(2,32) = 2.8, p > .05$; Within Low: $F(2,32) = 1.3, p > .1$).

To examine whether these between-item differences in category boundary could have resulted in the difference in slopes between Experiments 1 and 2, we used this same data set of slopes and category boundaries and averaged the slope *for each item* within each subject. This removes any effect of category boundary differences between items on the slope (because the slopes were computed with respect to the item). A comparison between these averaged slopes and the slopes from Experiment 1 revealed a significant difference ($t(32) = 4.4, p < .0001$); the identification curve from Experiment 2 was still shallower than Experiment 1. Thus, although there may have been some effect of category boundary differences on the slopes in Experiment 2, this difference was not sufficient to account for the differences between the experiments.

Analysis of the eye movements was performed on filtered data only (i.e., trials with the low frequency response were excluded). Figure 5 shows proportions of looks at the target (e.g., “bear” for a 0 ms VOT), competitor (“pear”), and the unrelated items (“lamp” or “ship”) as a function of time for VOTs of 0 ms (an endpoint /b/) and 40 ms (an endpoint /p/). In both figures, the proportion of looking at the target increases over time, peaking at

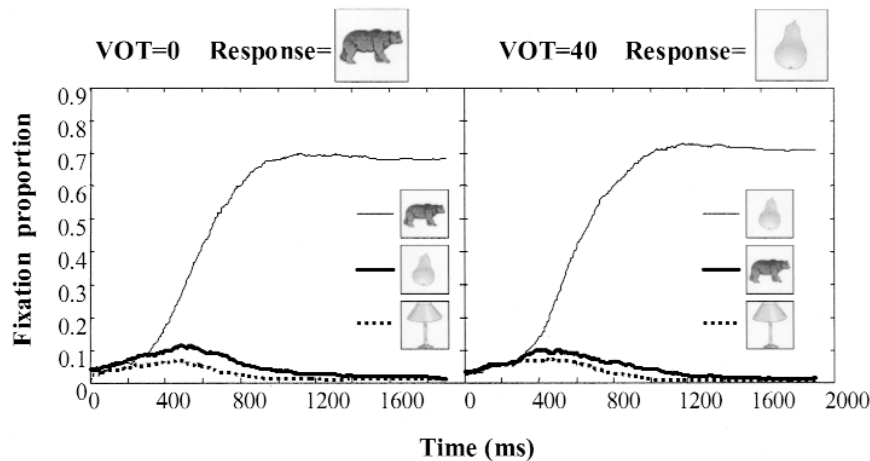


Fig. 5. Probability of a fixation as a function of time in the lexical identification task (Experiment 2). Again, fixations to the target (e.g., “bear” when the VOT was 0, or “pear” when the VOT was 40) are shown with thin lines and fixations to the competitor (“pear” when the VOT was 0, or “bear” when the VOT was 40) with thick lines. Fixations to the unrelated item (e.g., “lamp,” “ship”) are shown with dashed lines.

around 900 ms. The proportion of looks to the competitor also increase initially but falls off at around 450 ms. Overall, there are more looks to the competitor than the unrelated objects across all VOTs (/b/: $F(1,16) = 31.03$, $p < .001$; /p/: $F(1,16) = 22.57$, $p < .001$), demonstrating that the competitor effect is sensitive to the phonological similarity between the target and competitor (something we could not test in Experiment 1, due to the lack of unrelated items).

To evaluate gradiency, we examined looks to the competitor and target as a function of VOT to determine whether a linear trend describes the relationship between the proportion of fixations and VOT (i.e., do looks to the competitor increase as a linear function of VOT?). Additionally, eye-tracking data allow us to examine the effect of time and, crucially, the interaction of VOT and time (a lack of interaction would indicate that any effect of VOT does not change over time).

However, including time as a factor presents a difficult problem when using fixation data. Typically, these analyses would be done by dividing time into one or more bins and averaging fixation probabilities within each bin. However, because a single fixation could then contribute to more than one bin, assumptions of independence between levels of a factor would be violated. Therefore, to address this issue, we randomly assigned a *given trial* to either the “early” or “late” group and then only looked at one time bin within each group of trials. The early group included data for half of the trials (randomly selected) over the timespan of 300 to 1100 ms. The late group contained the other half of the trials from 1100 to 1900 ms.

Using this method, two-way ANOVAs were conducted with VOT and time as factors. Figure 6 shows the mean competitor activation for each VOT at each time. This is essentially the area under the competitor curve within a given time region. There was a significant effect of VOT for both “b” and “p” data sets (/b/: $F(3,48) = 3.84$, $p < .025$; /p/: $F(4,64) = 5.43$, $p < .005$): There were more fixations to the competitor as VOT neared the category boundary. More importantly, a significant linear trend of VOT (/b/: $F(1,16) = 6.42$, $p < .025$; /p/: $F(1,16) = 8.73$, $p < .005$) established a gradient relationship between VOT and competitor activation. Time was also significant (/b/: $F(1,16) = 17.20$, $p < .005$; /p/: $F(1,16) = 29.11$, $p < .001$) as there were more looks to the competitor in the early part of the trial. Importantly, the interaction was not significant (/b/, /p/: $p > .1$), suggesting that the effect of VOT did not change throughout the time-course (and extended remarkably late after word onset).

To verify that these effects were not driven solely by stimuli adjacent to the category boundary, these stimuli (15 ms for /b/ and 20 ms for /p/) were removed and the analyses repeated. Again, there was a significant main effect of VOT (/b/: $F(2,32) = 5.87$, $p < .01$; /p/: $F(3,48) = 4.01$, $p < .025$)

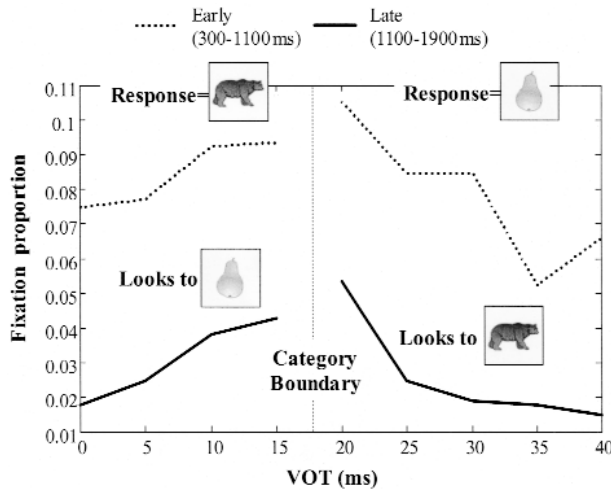


Fig. 6. Effect of VOT and time on fixations to the competitor.

and a significant linear trend (/b/: $F(1,16) = 8.06, p < .005$; /p/: $F(1,16) = 6.64, p < .05$). There was also a reliable effect of time (/b/: $F(1,16) = 16.80, p < .005$; /p/: $F(1,16) = 31.89, p < .001$), and neither interaction reached significance (/b/, /p/: $p > .1$).⁷

Figure 7 examines in more detail the temporal dynamics of the gradient effects of VOT. This figure shows the proportion of fixations to the competitor as a function of time across all VOTs. Importantly, although the overall amplitude (height) of these curves does vary with VOT, there is a much larger effect on the duration: The competitor effect lasts longer as VOT nears the category boundary.

Recall that there were small, but reliable, differences in category boundaries for words with high/front and low/back vowels. Because between-item differences in category boundaries might have affected the pattern of gradience in looks to the competitor, we collapsed items by vowel height to create two groups: high/front and low/back (recall that category boundaries did not differ significantly within these groups). These groups were included in

⁷ A separate analysis for the late time bin revealed a significant linear trend of VOT for both sides of the continuum using all of the stimuli (/b/: $F(1,16) = 4.80, p = .044$, /p/: $F(1,16) = 4.51, p = .05$). When stimuli near the category boundary were excluded, this effect remained for /b/ ($F(1,16) = 6.32, p = .023$) but was not significant for /p/ ($F(1,16) = 1.173, p > .1$) despite a visually present trend and a reasonable correlation ($R = -.31$).

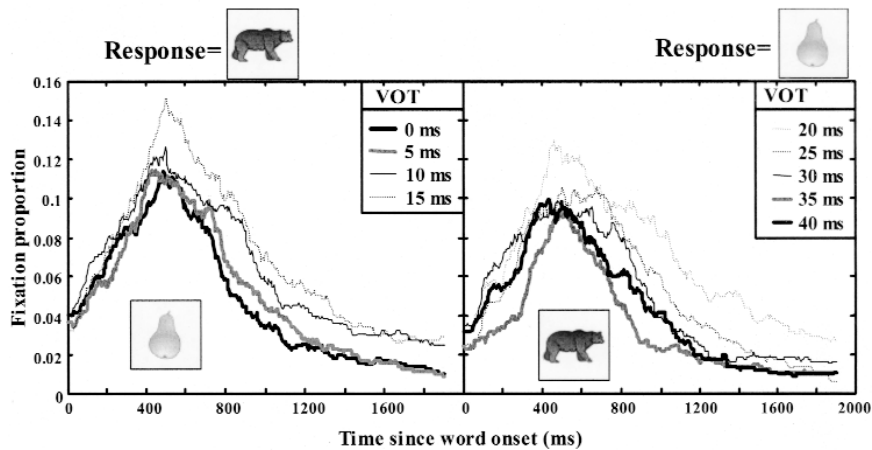


Fig. 7. Effect of VOT and time on fixations to the competitor. A clear gradient effect of VOT can be seen. Importantly, the effect of VOT is primarily on the duration of activation.

two-way VOT \times Vowel Height ANOVAs. If the gradient effect of VOT is independent of vowel-induced category boundary shifts, there should not be a reliable effect of vowel height or an interaction between vowel height and VOT. This is exactly what was found. For both /b/ and /p/ there was a significant effect of VOT (B: $F(3,48) = 2.9, p < .05$; P: $F(4,64) = 10.2, p < .001$), but crucially, there was no effect of vowel height (B: $F < 1$; P: $F < 1$) and no interaction (B: $F < 1$; P: $F(4,64) = 1.6, p > .1$). Removing stimuli adjacent to the category boundary yielded the same pattern of results: a significant effect of VOT (B: $F(3,32) = 5.1, p < .025$; P: $F(3,48) = 6.0, p = .001$), no effect of vowel height (B: $F < 1$; P: $F < 1$), and no interaction (B: $F < 1$; P: $F(3,48) = 1.9, p > .1$). Thus, vowel height/frontness, which had a dramatic effect on the category boundary, had no effect on looks to the competitor and, most crucially, did not modulate the gradient effect of VOT.

A similar set of analyses was performed on fixations to the target as well. Using all VOTs, we found a reliable main effect of VOT (/b/: $F(3,48) = 4.19, p = .01$; /p/: $F(4,64) = 5.36, p = .001$) and a linear trend (/b/: $F(1,16) = 4.79, p < .05$; /p/: $F(1,16) = 8.61, p = .01$). As VOTs approached the category boundary, there were fewer looks to the target. There were fewer looks to the target earlier than later (/b/: $F(1,16) = 54.83, p < .001$; /p/: $F(1,16) = 81.40, p < .001$), and there was no interaction of VOT and time ($p > .1$).

While these results parallel those found for the competitor effect, they are not statistically reliable when we remove the VOTs closest to the category boundary from the analysis. After removing trials with a VOT of 15 ms from the /b/ data set and 20 ms from the /p/ data set, the effect of VOT was

no longer significant ($p > .1$ for both data sets) nor was the linear trend (/b/: $F < 1$; /p/: $F(1,16) = 3.01$, $p = .108$). The interaction was not significant either ($p > .1$). Thus, the effect of VOT on the target is driven to a much larger extent by the difference between stimuli near the category boundary and those further from it.

To conclude, this experiment demonstrated clear *gradient* effects in lexical *competitor* activation as a function of VOT. Moreover, these gradient effects persist in time and seem to affect the duration as well as the overall amount of activation. The effect of VOT on the *target* was present but was much weaker.

RESOLVING PHONETIC CONTEXT DEPENDENCY AND TEMPORAL AMBIGUITY

Experiment 2 demonstrated that gradient effects of VOT are maintained over time during lexical activation. This effect suggests a potential mechanism for probabilistically integrating acoustic/phonetic context over time. Consider how gradiency could play a role in temporal phonetic ambiguity.

In sentence comprehension, a temporal *syntactic* ambiguity occurs when local syntactic information is consistent with more than one syntactic interpretation (and the disambiguating information has not yet arrived). For example, “The graduate student studied . . .” is locally consistent with a main clause interpretation (“The graduate student studied for his comprehensive exam”) or with a reduced relative clause (“The graduate student studied by Phil Zimbardo performed similarly to the hamster”). Although globally these sentences are unambiguous, given the serial arrival of information in spoken language, they are ambiguous at the point where the listener has only heard “The graduate student studied. . . .”

A similar process can occur at the acoustic/phonetic level. To take a simplified “phonemic” view of speech perception, consider a word like “*beetle*,” in which the initial consonant is *ambiguous* between /b/ and /p/. In this case, very early, when the subject has only heard /bi/, the information is compatible with words like *beetle* and *beaker* as well as with the competitors, *people* and *peeler*. However, when the remaining information arrives (the “tle”), the temporal ambiguity is resolved and the word (and its sublexical representation) can be resolved.

Speech perception provides a host of effects that could create such phonetic temporal ambiguities. Later-arriving lexical information (Ganong, 1980) could be used to disambiguate initial ambiguities (e.g., the fact that *kiss* is a word and *giss* is not could resolve any initial ambiguity between /k/ and /g/). Additionally, many of the effects of speaking rate on voicing or manner perception are conditioned on the length of the *subsequent* vowel

(Summerfield, 1981; Miller *et al.*, 1978). Finally, Gaskell (2001) reports regressive effects of place assimilation (in that the place of articulation of a consonant affects the perception of a *preceding* consonant).

In cases such as these, if the degree of ambiguity in the initial consonant is maintained and both potential referents are kept active to a certain extent, then the ambiguity could be resolved more efficiently than if a discrete decision is made and must later be revised.⁸ The competitor effect in Experiment 2 suggests exactly this behavior, as competitors with more ambiguous onsets are kept more active and for longer durations. Moreover, this activation persists for quite a while, at least into the 1100–1900 ms bin—more than enough time for the listener to have perceived later-occurring phonetic information such as vowel length, prosodic position, or lexically disambiguating information. Thus, it appears that the speech perception system has the information required to resolve ambiguity using a probabilistic constraint-satisfaction mechanism, rather than a discrete-choice, “garden-path” model.

Additionally, this mechanism would also be useful for integrating information forward in time (anticipating upcoming phonetic material). For example, Mann and Repp’s (1981) work on compensatory coarticulation shows that the identity of a fricative can influence the identity of a subsequent stop consonant. Local (2002) reports acoustic measurements of the second and third formants of vowels that indicate within-category differences that predict an /r/ or an /l/ up to 15 phonemes away. Finally, Gow (2001) reports a series of perceptual experiments demonstrating anticipatory place assimilation (in which the place of articulation of one consonant influences the perception of the following consonant’s place).

Our data suggest that the perceptual system is retaining information about the acoustic information for stretches of speech that span several phonemes and perhaps several words. This information could then be used to anticipate upcoming information. For example, in data reported by Gow (2001), when the subject has heard “run picks” (where the /n/ will be slightly labialized due to place assimilation with the /p/), the early acoustic information predicting /p/ could weakly activate this phoneme (or features, or a distribution of lexical items) to the degree that they match the input (the labialized /n/). This activation could be retained long enough so that when the segment (/p/) is actually heard it has been slightly primed.

Our experiments demonstrate that within-category subphonemic variation does have a lasting, gradient effect on lexical access. The temporal property of this effect (although not tested directly in our experiments) could provide a powerful mechanism for both anticipatory and regressive context sensitivities.

⁸ This is not unlike Frazier’s (1999) model of sentence parsing.

CONCLUSION

We have shown that VOT exerts a gradient effect on lexical access and that this effect persists over time. Importantly, these effects were reliable despite a conservative experimental approach in which

1. “Filtering” of the data removed trials in which the competitor was selected (these would be the trials with the most competitor activation).
2. Gradient effects remained even when we removed VOTs near the category boundary—thus the gradiency is not driven by differences between ambiguous stimuli near the category boundary and more prototypical stimuli.

Moreover, the temporal dynamics of the gradient effects are such that variation in VOT seems to affect the amount of time that competitors are active, suggesting that this gradient sensitivity could be used for integrating fine-grained detail over large temporal regions of context.

Important issues for future research will be to (1) determine whether the system makes use of systematic variation in the ways we have suggested, (2) determine the degree to which subphonemic variation is systematic in the signal, and (3) model lexical access using learning-based approaches in which priors, or weights, are sensitive to these systematic fine-grained contingencies in the input.

To conclude, the data presented here suggest that speech perception is sensitive to small within-category differences in VOT. Moreover, this sensitivity to fine-grained phonetic detail lasts long enough to be useful in integrating phonetic material over large temporal regions of context. Thus, the system is probabilistically sensitive to fine-grained variation, and this variation does not represent noise to be discarded, but rather a signal that is likely to be important in lexical access.

REFERENCES

- Allen, J. S., & Miller, J. L. (1999). Effects of syllable-initial voicing and speaking rate on the temporal characteristics of monosyllabic words. *Journal of the Acoustical Society of America*, *106*, 2031–2039.
- Allopenna, P., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory and Language*, *38*(4), 419–439.
- Andruski, J. E., Blumstein, S. E., & Burton, M. W. (1994). The effect of subphonetic differences on lexical access. *Cognition*, *52*, 163–187.
- Carney, A. E., Widin, G. P., & Viemeister, N. F. (1977). Non categorical perception of stop consonants differing in VOT. *Journal of the Acoustical Society of America*, *62*, 961–970.

- Crosswhite, K., Masharov, M., McDonough, J., & Tanenhaus, M. (in preparation). Phonetic cues to word length and online processing of onset-embedded words.
- Dahan, D., Magnuson, J. S., & Tanenhaus, M. K. (2001). Time course of frequency effects in spoken-word recognition: Evidence from eye movements. *Cognitive Psychology*, *42*, 317–367.
- Dahan, D., Magnuson, J. S., Tanenhaus, M. K., & Hogan, E. (2001). Subcategorical mismatches and the time course of lexical access: Evidence for lexical competition. *Language and Cognitive Processes*, *16*, 507–534.
- Fougeron, C., & Keating, P. (1997). Articulatory strengthening at edges of prosodic domains. *Journal of the Acoustical Society of America*, *101*, 3728–3740.
- Fowler, C. (1991). The perception of phonetic gestures. In I. G. Mattingly & M. Studdert-Kennedy (Eds.), *Modularity and the Motor Theory of Speech Perception* (pp. 33–59). Hillsdale, NJ: Lawrence Erlbaum Associates Inc.
- Frazier, L. (1999). *On Sentence Interpretation*. Boston: Kluwer Academic Publishers
- Ganong, W. F. (1980). Phonetic categorization in auditory word recognition. *Journal of Experimental Psychology: Human Perception and Performance*, *6*(1), 110–125.
- Gaskell, G. (2001). Phonological variation and its consequences for the word recognition system. *Language and Cognitive Processes*, *15*(5/6), 723–729.
- Gaskell, G., & Marslen-Wilson, W. (1996). Phonological variation and inference in lexical access. *Journal of Experimental Psychology: Human Perception and Performance*, *22*, 144–158.
- Gow, D. (2001). Assimilation and anticipation in continuous spoken word recognition. *Journal of Memory and Language*, *45*, 133–139.
- Klatt, D. (1980). Software for a cascade/parallel synthesizer. *Journal of the Acoustical Society of America*, *67*, 971–995.
- Ladefoged, P. (1993). *A Course in Phonetics*. New York: Harcourt Brace Publishers.
- Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, *74*, 431–461.
- Liberman, A. M., Harris, K. S., Hoffman, H. S., & Griffith, B. C. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology*, *54*(5), 358–368.
- Lisker, L. (1986). “Voicing” in English: A catalogue of acoustic features signaling /b/ versus /p/ in trochees. *Language and Speech*, *29*(1), 3–11.
- Lisker, L., & Abramson, A. (1964). A cross-language study of voicing in initial stops: Acoustical measurements. *Word*, *20*, 384–422.
- Local, J. (2002). Variable domains and variable relevance: Interpreting phonetic exponents. Paper presented at Temporal Integration in the Perception of Speech, Aix-en-Provence, April, 2002.
- Mann, V. A., & Repp, B. (1981). Influence of preceding fricative on stop consonant perception. *Journal of the Acoustical Society of America*, *69*(2), 548–558.
- Marslen-Wilson, W. (1987). Functional parallelism in spoken word recognition. *Cognition*, *25*(1–2), 71–102.
- Massaro, D. W., & Cohen, M. M. (1983a). Phonological context in speech perception. *Perception & Psychophysics*, *34*, 338–348.
- Massaro, D. W., & Cohen, M. M. (1983b). Categorical or continuous speech perception: A new test. *Speech Communication*, *2*, 15–35.
- Maye, J., Werker, J. F., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, *82*, 101–111.
- McClelland, J., & Elman, J. (1986). The TRACE model of speech perception. *Cognitive Psychology*, *18*(1), 1–86.
- McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*, 746–748.

- McMurray, B. (in preparation). KlattWorks: A [somewhat] new systematic approach to formant-based speech synthesis for empirical research.
- McMurray, B., Aslin, R. N., Tanenhaus, M. K., Spivey, M. J., & Subik, D. (in preparation). Two B or not Two B: Categorical perception in lexical and nonlexical tasks.
- McMurray, B., Spivey, M. J., & Aslin, R. N. (2000). The perception of consonants by adults and infants: Categorical or categorized? *Working Papers in the Language Sciences at the University of Rochester*, 1(2), 215–256.
- McMurray, B., Tanenhaus, M. K., & Aslin, R. N. (2002). Gradient effects of within-category phonetic variation on lexical access. *Cognition*, 86, B33–B42.
- McQueen, J. (1996). Phonetic categorization. *Language and Cognitive Processes*, 11(6), 655–664.
- Miller, J. L. (1997). Internal structure of phonetic categories. *Language and Cognitive Processes*, 12, 865–869.
- Miller, J. L., & Liberman, A. M. (1979). Some effects of later-occurring information on the perception of stop consonant and semi-vowel. *Perception & Psychophysics*, 25, 457–465.
- Norris, D. (1994). Shortlist: A connectionist model of continuous speech recognition. *Cognition*, 52(3), 189–234.
- Oden, G., & Massaro, D. W. (1978). Integration of featural information in speech perception. *Psychological Review*, 85(3), 172–191.
- Pisoni, D. B., & Lazarus, J. H. (1974). Categorical and noncategorical modes of speech perception along the voicing continuum. *Journal of the Acoustical Society of America*, 55(2), 328–333.
- Pisoni, D. B., & Sawusch, J. R. (1974). On the identification of place and voicing features in synthetic stop consonants. *Journal of Phonetics*, 2, 181–194.
- Repp, B., & Mann, V. A. (1982). Fricative-stop coarticulation: Acoustic and perceptual evidence. *Journal of the Acoustical Society of America*, 71(6), 1562–1567.
- Sharma, A., & Dorman, M. F. (1999). Cortical auditory evoked potential correlates of categorical perception of voice-onset time. *Journal of the Acoustical Society of America*, 106(2), 1078–1083.
- Smits, R. (submitted). Spoken word recognition benefits from deferring sublexical decisions.
- Stevens, K. N., & Blumstein, S. E. (1978). Invariant cues for place of articulation in stop consonants. *Journal of the Acoustical Society of America*, 64(5), 1358–1368.
- Summerfield, Q. (1981). Articulatory rate and perceptual constancy in phonetic perception. *Journal of Experimental Psychology: Human Perception and Performance*, 7(5), 1074–1095.
- Summerfield, Q., & Haggard, M. (1977). On the dissociation of spectral and temporal cues to the voicing distinction in initial stop consonants. *Journal of the Acoustical Society of America*, 62(2), 435–448.
- Sussman, H., Fruchter, D., Hilbert, J., & Sirosh, J. (1998). Linear correlates in the speech signal: The orderly output constraint. *Behavioral and Brain Science*, 21(2), 241–299.
- Tanenhaus, M. K., Magnuson, J. S., Dahan, D., & Chambers, C. G. (2000). Eye movements and lexical access in spoken language comprehension: Evaluating a linking hypothesis between fixations and linguistic processing. *Journal of Psycholinguistic Research*, 29, 557–580.
- Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. E. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, 268, 632–634.
- Volaitis, L. E., & Miller, J. L. (1992). Phonetic prototypes: Influence of place of articulation and speaking rate on the internal structure of voicing categories. *Journal of the Acoustical Society of America*, 92(2), 723–735.

