Published in final edited form as:
Atten Percept Psychophys. 2012 August ; 74(6): 1284–1301. doi:10.3758/s13414-012-0306-z.

Cue integration and context effects in speech: Evidence against speaking rate normalization

Joseph C. Toscano[1,2] and **Bob McMurray**[2,3]

[1]Beckman Institute, University of Illinois at Urbana-Champaign

[2]Delta Center, University of Iowa

[3]Dept. of Psychology and Dept. of Communication Sciences & Disorders, University of Iowa

Abstract

Listeners are able to accurately recognize speech despite variation in acoustic cues across contexts, such as different speaking rates. Previous work has suggested that listeners use rate information (indicated by vowel length; VL) to modify their use of context-dependent acoustic cues, like voice-onset time (VOT), a primary cue to voicing. We present several experiments and simulations that offer an alternative explanation: listeners treat VL as an phonetic cue, rather than as an indicator of speaking rate, and rely on general cue-integration principles to combine information from VOT and VL. We demonstrate that listeners use the two cues independently, that VL is used in both naturally-produced and synthetic speech, and that effects of stimulus naturalness can be explained by a cue-integration model. Together, these results suggest that listeners do not interpret VOT relative to rate information provided by VL and that effects of speaking rate can be explained by more general cue-integration principles.

A fundamental issue in speech perception is understanding how the perceptual system maps acoustic cues onto phonological categories. This can be seen as an instance of a general categorization problem (Holt & Lotto, 2010), however, it is particularly complex because cues are influenced by many factors (Repp, 1982; McMurray & Jongman, 2011). For example, voice-onset time (VOT), a primary cue[1] for voicing (the difference between /b,d,g/ and /p,t,k/), is also influenced by place of articulation (Lisker & Abramson, 1964; Cho & Ladefoged, 1999), talker characteristics (Allen, Miller, & DeSteno, 2003), stress (Smiljanic & Bradlow, 2008), and whether the phoneme was mispronounced (Goldrick & Blumstein, 2006).

Perhaps most importantly, VOT is influenced by speaking rate. VOT corresponds to the time between the release of the articulators (e.g., the lips for a /b/ or /p/) and the onset of laryngeal voicing. Thus, as a temporal difference, it is heavily influenced by speaking rate (Kessinger & Blumstein, 1998; Allen & Miller, 1999; Miller, Green & Reeves, 1986; Pind, 1995; Beckman, Helgason, McMurray, & Ringen, 2011). Researchers have viewed rate as a source of contextual variability, suggesting that listeners use rate to normalize VOT (i.e., listeners treat VOT-values differently at different speaking rates). The present study reconsiders this view, suggesting that one of the most commonly used measures of rate, vowel length (VL), may actually be an independent phonetic cue for word-initial voicing. That is, speaking rate effects might not reflect normalization for context at all.

Corresponding Author: Joseph Toscano Beckman Institute for Advanced Science and Technology University of Illinois at Urbana-Champaign 405 N Mathews Ave Urbana, IL 61801 217-244-3977 (voice) 217-333-2922 (fax) jtoscano@illinois.edu.

[1]For the purposes of this paper, we will use the term *cue* to refer to any source of information in the sound signal that can be used to distinguish the relevant categories.

This has implications for many distinctions in speech, since temporal cues are used for various phonological contrasts (stop/approximate, Miller & Liberman, 1979; fricative/affricate, Repp, Liberman, Eccardt, & Pesetsky, 1978; fricative place and voicing, McMurray & Jongman, 2011), and VOT is commonly used as a model cue-category system (McMurray et al, 2002; Andruski, Blumstein & Burton, 1994; Pisoni & Tash, 1974). Thus, understanding the relationship between VOT and rate informs models of context compensation more generally.

As noted above, previous studies have generally treated the contextual variability as a normalization problem. For rate compensation, they have proposed that listeners adjust their use of VOT depending on the rate (indicated by VL). This may be done via *intrinsic normalization*, in which listeners compute the ratio between VOT and VL (Summerfield, 1981; Boucher, 2002) and use this continuous measure as the basis of phonological categorization. Alternatively, it may also be done via *extrinsic normalization*, where listeners use other indicators of rate, such as sentential rate (Dupoux & Green, 1997; Pind, 1995; Wayland, Miller, & Volaitis, 1994) to either adjust their VOT boundary or recode VOT as a rate-independent cue. Importantly, in both approaches, the relationship between VOT and estimates of speaking rate is based on how rate affects VOT (e.g., Miller et al., 1986) and should be relatively constant across situations.

We propose that more general-purpose computations may suffice to account for this apparent context effect. Specifically, we argue that VL effects could be handled in a *cue-integration* framework (Oden & Massaro; 1978; Nearey 1997; Smits, 2001a; Toscano & McMurray, 2010) in which multiple acoustic cues contribute directly to phonological categorization. Under this approach, cues that indicate properties of the context (e.g., estimates of rate) may also be weaker, first-order cues to a phonological distinction. VL is typically viewed as a cue to speaking rate, but short vowels (indicative of a fast rate) are also more likely to derive from the voiceless category, and long vowels from the voiced category (Allen & Miller, 1999; though see Kessinger & Blumstein, 1998). Thus, VL could simply serve as an independent cue to word-initial voicing. As a result, when the contributions of each cue are combined, VL biases the ultimate judgment, allowing cue-integration processes to mimic rate-sensitive ones (see Nearey, 1990, for an analogue in which diphone biases mimic the effect of compensation for coarticulation). Critically, this differs from normalization approaches in that it does not treat VL as a source of contextual variability that listeners normalize for, nor does it require VOT encoding to be explicitly tuned to rate. Rather, by treating VL as an additional phonetic cue, listeners responses may be explained with a much simpler model.

There is evidence that listeners use VL as a cue for other distinctions, such as word-final voicing (Warren & Marslen-Wilson, 1988), vowel quality (Hillenbrand, Clark & Houde, 2000), and syllable structure (Salverda, Dahan & McQueen, 2003). This work suggests that VL can directly bias listeners phonetic or lexical percepts. However, this has not been applied to word-initial voicing, and VL is typically thought of as a cue for speaking rate that participates only indirectly in word-initial voicing judgments via rate compensation. Thus, treating VL as a direct cue for word-initial voicing may provide a plausible and simpler explanation for rate effects that is more consistent with how VL is used for other distinctions.

Previous work provides some insights about whether a normalization or cue-integration approach offers a better explanation but does not rule out one or the other. The goal of our work is to provide converging evidence for cue-integration by testing several predictions made by the two approaches.

Speaking rate and VOT

Listeners consistently identify VOTs near 0 ms as voiced and VOTs near 40-60 ms as voiceless. However, at fast rates, the boundary shifts such that more tokens are identified as voiceless (Summerfield, 1981; Miller & Dexter, 1988; Pind, 1995; Miller & Volaitis, 1989; McMurray et al., 2008b; but see Diehl, Souther, & Convis, 1980). This can be observed for differences in preceding sentential rate and in isolated words with rate indicated by VL, though the effect of VL is larger (Summerfield, 1981; Wayland, Miller, & Volaitis, 1994).

Both normalization and cue-integration mechanisms can account for VL effects. Listeners might normalize VOTs by using the ratio between VOT and VL (rather than raw VOT) as the basis of phonetic categorization, or by computing VOT relative to the surrounding rate, and again using this new compound cue as the input to categorization. This would produce either a shift in listeners' perceived VOT as a function of VL, or a shift in the VOT boundary between voicing categories. The cue-integration approach, however, suggests that VL is simply a secondary cue to voicing: long VLs are associated with voiced sounds and short VLs with voiceless sounds. Work in several areas — phonetics, perceptual categorization, and online spoken word recognition — provides data about the relationship between VOT and VL that may be helpful for distinguishing these approaches. However, previous work does not provide a definitive answer.

Phonetic results

A number of studies have examined the acoustic properties of voicing as a function of speaking rate, though two different definitions of VL have been used. The first, *release-to-offset VL,* defines the beginning of the vowel as the release burst (inclusive of VOT) and the end as the offset of periodicity (Turk, Nakai, & Sugahara, 2006). This measure has been adopted in many studies of rate compensation (Miller & Liberman, 1979; Shinn et al., 1985; Miller & Wayland, 1993; Miller & Dexter, 1988). The second definition, *onset-to-offset VL,* instead defines the beginning of the vowel as the onset of voicing (e.g., Summerfield, 1981).[2]

Allen and Miller (1999) measured VOT and onset-to-offset VL for words spoken at different rates. As expected, VOT- and VL-values were shorter for fast rates than for slow rates. Crucially, at both rates, the difference between the voicing categories was signaled by changes in both VOT and VL. While VL distributions overlapped, voiced words had statistically longer VLs than voiceless words, and VL was not strongly correlated with VOT within each category (voiced: $r=-0.04$; voiceless: $r=-0.05$).[3] Thus, VL offers some information about word-initial voicing, in addition to word-final voicing (House, 1961; Peterson & Lehiste, 1960).

This phonetic work has been somewhat controversial with studies showing that onset-to-offset VL increases with VOT within a single category and rate (Kessinger & Blumstein, 1998) and others arguing that release-to-offset VL should be measured instead (Turk, Nakai, & Sugahara, 2006), which predicts that longer VLs should lead to more voiceless responses (though listeners may associate short vowels with voiceless sounds for the purpose of rate compensation).

In either case, the phonetic data suggest an approach for addressing whether rate effects reflect context normalization: small differences in VL that do not provide distinct rate

[2]In our experiments, all stimuli in the *short VL* condition for a given continuum are shorter than stimuli in the *long VL* condition, regardless of which definition is used.

[3]These values were estimated from the raw data from Allen and Miller (1999).

differences should still affect voicing categorization under a cue-integration model. For example, Allen and Miller (1999) found a mean VL difference of approximately 20 ms between voicing categories. This is much smaller than the variation produced by rate, and a perceptual effect of such a difference would support cue-integration. This will be examined in the present study.

Temporal asynchrony

Critical to the perceptual problem of integrating VOT and VL is the fact that VL can only be estimated *after* VOT is heard; the two cues are not temporally synchronous. Because of this, listeners could adopt one of several strategies for combining the two sources of information. For example, they could store VOT and wait to receive VL before activating higher level representations (like phonemes or words). This would be consistent with a normalization approach in which VOT is coded relative to context. Alternatively, they could use each cue as it becomes available, suggesting that they are treating the two as independent cues.

Recently, McMurray et al. (2008b) examined listeners' eye-movements in a visual world paradigm experiment to determine whether VOT and VL are used as soon as they become available or whether voicing decisions are delayed until listeners have both cues (see also Miller & Dexter, 1988). Participants heard synthesized minimal pair words (e.g., *beach/peach*) from one of several VOT continua that also varied in VL.[4] They clicked on the matching picture from a computer screen while their eye-movements to each object were monitored as a measure of real-time lexical activation. An eye-movement can only reflect processing that has occurred up to the point at which it was launched. Thus, by measuring the influence of VOT and VL on the fixation record at each time point, they could determine whether each cue was used as soon as it arrived or whether voicing decisions were delayed until both cues were available.

At the earliest points in time, listeners' fixations showed only the effect of VOT and no effect of VL. About 100 ms later, an effect of VL could be seen. Thus, both VOT and VL contributed to lexical activation, and they did so in a continuous cascade, rather than being buffered until both were received. These results seem to favor a cue-integration approach, since they suggest listeners treat VOT and VL as two independent cues rather than normalizing VOT on the basis of VL. Thus, an estimate of rate does not seem to be required to interpret VOT. However, this is only a preliminary conclusion since this study used synthetic speech, which, as discussed in the next section, may be processed differently with respect to rate than naturally-produced speech. Moreover, a normalization approach that assumes a default rate in the absence of VL could also account for these results. Thus, additional evidence is needed.

Stimulus naturalness

Work examining stimulus naturalness suggests that perceptual effects of VL might be limited to synthetic speech. The earliest hints of this come from work on the closely related manner of articulation distinction (/b/-/w/) which is primarily cued by formant transition duration and is also affected by speaking rate. Here, rapidly changing formants cue stops (/b/) and slower transitions cue approximants (/w/). This manner distinction shows a relationship similar to voicing: long VLs cue stops and short VLs cue approximants (Miller & Liberman, 1979; Shinn et al., 1985; Miller & Wayland, 1993; McMurray et al., 2008b).

Shinn et al. (1985) examined the effect of VL on manner identification using a continuum in which two additional cues (formant onset frequency and burst amplitude) co-varied with the

[4]/b/-/w/ words varying in formant transition duration and VL showed the same effects.

primary cue (transition duration). This led to stimuli that are more consistent with what listeners are likely to hear in natural speech. They found no effect of VL on the category boundary, but when transition duration was manipulated without the additional cues, the VL effect re-emerged, suggesting that naturalness can modulate the use of VL (though see Miller and Wayland, 1993).

Studies examining voicing in natural speech have found mixed results. Boucher (2002) presented listeners with a naturally-produced /d/ with a constant VOT (15 ms) and different VLs. VL affected listeners' voicing judgments, with more /t/ responses for short VLs. This contrasts with Shinn et al. (1985), suggesting an effect of VL in natural speech. In contrast, Utman (1998) found no effect of VL on voicing judgments with naturally-produced voiceless tokens that were manipulated such that the relationship between VOT and VL was either compatible with respect to the speaking rate (i.e., both short [fast] or both long [slow]) or incompatible. While a VL effect was observed in goodness ratings, this is suggestive of a weaker effect. Thus, listeners may treat natural and synthetic speech differently. However, neither study examined an entire continuum (only voiceless tokens were used by Utman, and a single VOT by Boucher), so it is difficult to compare the results to studies examining VOT continua with synthetic speech. Further, neither experiment explicitly compared naturally-produced and synthetic speech.

Given these results, it is unclear whether VL has an effect at all in natural speech. Understanding the effect of naturalness may also help to distinguish between normalization and cue-integration approaches. Normalization approaches predict that the effect of VL in natural speech should be similar to the one seen with synthetic speech, since the relationship between VOT and speaking rate is constant. If rate normalization is based on the fixed relationship between VOT and rate, it would not be affected by naturalness. Thus, to explain the differences between synthetic and natural speech, some researchers have argued that listeners simply do not normalize for rate in natural speech (Shinn et al., 1985), though others have argued that under more-realistic listening conditions (e.g., background noise), VL effects in natural-sounding speech can be observed (Miller & Wayland, 1993). However, it is unclear whether these effects are the same size as typical VL effects seen with synthetic speech; a smaller effect would be difficult to explain in terms of normalization given the fixed relationship between rate and VOT.

In contrast to either a constant VL effect across stimulus types or an absent VL effect in natural speech, a cue-integration account can allow for differences in the size of the effect between the two types of speech. Cue integration models (Oden & Massaro, 1978; Nearey, 1997; Smits, 2001a; Toscano & McMurray, 2010; McMurray & Jongman, 2011) posit that multiple cues are combined to determine the overall percept. In these models, cues are weighted and added together. Both the weight of the cue and the specific cue-values determine the likelihood of a particular response. Thus, if a VOT value is near a category boundary, it will contribute little to the voicing decision (which could be driven mostly by VL), even though it is weighted highly.

Consider Shinn et al. (1985)'s continua in which one or more secondary cues varied along with transition duration. Since secondary cues covary with VOT in natural speech, they will bias the response toward the voiced or voiceless category (in the same direction as VOT), and the effect of VOT will actually reflect the contribution of multiple cues. This would reduce the apparent effect of VL, even though the weight and actual VL values remain the same. In contrast, in synthetic speech, where these secondary cues are held constant at an ambiguous value, they will not bias the response much, and the apparent effect of VL will be larger. Crucially, these differences do not require cues to be weighted differently in synthetic and natural speech. This could explain the results of Shinn et al. (1985) and studies

using naturally-produced stimuli, which may also contain cues that covary with VOT. However, it is unclear from previous studies whether cue-integration approaches provide a better explanation.

Converging evidence

As the review above indicates, it is unclear whether a normalization or cue-integration model better describes the relationship between VOT and VL. Here, we present a series of experiments designed to look for converging evidence for one approach by testing four predictions made by normalization and cue-integration accounts.

First, for naturally-produced speech, normalization approaches predict the same effect as synthetic speech or no effect at all (if listeners simply ignore rate). In contrast, the cue-integration approach predicts an effect in naturally-produced speech that may be smaller due to additional cues that correlate with VOT. This is evaluated in Experiments 1 (synthetic speech) and 2 (naturally-produced speech).

Second, the cue-integration approach predicts that listeners' use of VOT and VL should be temporally asynchronous (since the two cues arrive at different times), while normalization accounts predict that the two should be used at the same time since listeners must adjust their use of VOT on the basis of rate. While this has been evaluated in synthetic speech, previous work suggests that rate compensation is quite different in natural speech. This is also examined in Experiments 1 and 2, which use the visual world paradigm to examine the timecourse of cue use.

Third, the cue-integration approach predicts that the effect of stimulus naturalness is due to additional voicing cues that covary with VOT. We motivate this by measuring additional cues in our naturally-produced stimuli and by running simulations with a cue integration model, the weighted Gaussian mixture model (WGMM; Toscano & McMurray, 2010). Experiment 3 tests this prediction by manipulating whether other cues covary with VOT (mimicking natural speech) or are held constant at an ambiguous value (as in synthetic speech).

Finally, the cue-integration approach predicts that small VL differences may have an effect on voicing, while normalization approaches predict that, since they do not indicate a large difference in speaking rate, they should not. This was examined in Experiment 4.

Together, these experiments and simulations allow us to distinguish between cue-integration and normalization explanations for VOT and VL effects in word-initial bilabial stops. Results consistent with cue-integration principles would offer a more parsimonious account than normalization and would suggest that listeners do not encode VOT relative to VL.

Experiment 1: Synthetic speech

Experiment 1 examined the timecourse of VOT and VL effects using synthetic speech. This allows us to replicate previous results with additional words (McMurray et al., 2008b, only used three VOT continua) and establish a better baseline for comparison to natural speech. Participants heard VOT continua (with either a long or short vowel) spanning two /b/-/p/ minimal pair words. They performed a 4AFC picture identification task while eye movements to the pictures were monitored. The proportion of looks to each object was used as a measure of listeners' ongoing lexical activation, and they were examined as a function of VOT and VL to determine when the effect of each cue occurred. If rate information is treated as a cue rather than as context, the effect of VOT should precede that of VL.

Methods

Participants—Monolingual native English speakers with normal or corrected-to-normal vision were recruited from the University of Iowa community in accordance with university human subject protocols and received course credit or $15 per hour. Twenty-seven participants completed the experiment, but three were excluded for having less than 80% correct on the endpoints of the VOT continua.

Design—Seven /b/-/p/ minimal pairs varied in nine steps of VOT and two VLs, yielding 126 total stimuli. Fourteen unrelated items (beginning with /l/ or /?/) were also included. All stimuli had either short or long VLs. Unrelated items served as the auditory stimulus on 50% of the trials. Stimuli were presented in random order, and each /b/-/p/ stimulus was repeated five times. This led to a total of 1260 trials (7 continua × 9 VOTs × 2 VL × 5 repetitions + fillers). Data collection was conducted over two sessions, lasting approximately an hour each.

Stimuli—Stimuli were synthesized using the KlattWorks front end (McMurray, 2008) to the Klatt (1980) synthesizer. For each word, parameter values for formant frequencies and amplitude components were set to closely match a recording from a natural utterance. VOT continua were constructed from the voiced endpoint by cutting back the AV (amplitude of voicing) parameter and replacing it with AH (amplitude of aspiration). Other than its duration, the characteristics of the aspiration were held constant across the VOT continua. F1, F2, and F3 transitions had rising frequencies at word onset, and were generally matched to the spectrogram of the natural recordings for the rest of their time course.

VL conditions differed by 100 ms for each minimal pair, though the overall duration of the vowel for a given pair depended on vowel quality (e.g., *buck/puck* had a longer VL than *bet/pet* because /?/ is typically longer than /?/). A 100 ms VL difference was chosen since, generally, asking listeners to speak quickly or slowly produces differences in VL of approximately 100 ms (Kessinger & Blumstein, 1997; Beckman et al., 2011; Magloire & Greene, 1999; Allen & Miller, 1999). We also wanted to replicate the relevant stimulus parameters from McMurray et al. (2008b) as closely as possible. Note that, because we systematically manipulated the VOT and VL values of the stimuli, they did not share the natural covariance between the two cues across voicing categories. Table 1 lists the VOT and VL values for the stimuli. Unrelated items beginning with /l/ (*lace, lap, leash, light, loaf, lock, loop*) and /?/ (*chef, shake, sheep, sheet, sheik, ship, shop*) were also synthesized with long and short vowels.

Visual stimuli were clipart images normed using a procedure designed to ensure that each was an acceptable referent for the target words (as in Apfelbaum, Blumstein & McMurray, 2011; McMurray, Samelson, Lee & Tomblin, 2010). For each word, several pictures were downloaded from a commercial clipart database. A team of graduate and undergraduate students examined each set of pictures and selected the most canonical exemplar for that item. The selected pictures were edited to obtain a consistent level of color and brightness, eliminate districting elements (e.g. objects in the background), and make minor modifications.

Procedure—The experiment was run using PsyScope (Cohen, MacWhinney, Flatt, & Provost, 1993) and the Eyelink PsyExtension (Bernstein, 2000). Participants were seated in front of an Apple Macintosh computer with a 19” CRT monitor. Eye movements were recorded using an SR Research Eyelink II head-mounted eye tracker. Before the experiment, the eye tracker was calibrated using a nine point calibration grid controlled by the Eyelink computer software. Auditory stimuli were presented binaurally over Sennheiser headphones.

Participants were able to adjust the volume on the headphones to a comfortable level using a Samson C-que 8 amplifier in front of them. Similarly to McMurray, Tanenhaus, and Aslin (2002), in the first session, participants were familiarized with the pictures of each stimulus in two blocks of training trials. First, each picture was shown, and its name appeared below. Next, participants were tested in a 4AFC picture identification task, in which they read the name of the item and clicked on the corresponding picture. Each word was presented twice in random order.

On each experimental trial, four pictures were shown: the two minimal pair referents and two unrelated items (an /l/ and /?/ item). A blue circle (100 ×100 pixels) at the center of the screen served as an initial fixation point. 500 ms after the beginning of the trial, the circle turned red, cueing the participant to click on it. 100 ms after they clicked on it, the auditory stimulus was played, and the participant clicked on the picture corresponding to the word they heard.

Since minimal pairs always appeared together (e.g., *beach* always appeared with *peach*), the same two unrelated items also always accompanied a particular minimal pair (similarly to McMurray et al., 2002). This was randomized between participants, and items that were semantically, visually, or acoustically similar were not paired.

The pictures appeared in each of the four corners of the display, and the relative positions of the four pictures were randomized on each trial. Each picture was 200 × 200 pixels in size (approximately 6.4° at a viewing distance of 50 cm) and they were separated by 780 pixels (24.5°) horizontally and 524 pixels (16.6°) vertically in the display.

Data analysis—Eye movement analysis proceeded similarly to McMurray et al. (2008a,b). Eye movements were automatically parsed into saccade and fixation events. Each saccade was paired with a subsequent fixation to create a single “look” that started at the onset of the saccade (the earliest moment the participant could be said to be attending to the object) and ended at the offset of fixation. At each 4 ms time step, the proportion of trials on which the participant directed a look to each object was computed. To account for drift during the experiment and noise in the calibration, the boundaries of the pictures were extended by 100 pixels for analysis.

Results

Mouse-click responses—Figure 1 shows participants’ responses as a function of VOT, VL, and continuum. Both cues affected participants’ decisions, with more /p/ responses for shorter VLs, particularly at the VOT boundary. A four-parameter logistic function (defined by its slope, midpoint, and lower and upper asymptotes) was fit to each participant's responses (proportion /p/) as a function of VOT. Separate functions were fit for each VL condition and continuum. Our analyses focused on the midpoint of the logistic (corresponding to the category boundary); slope showed no significant effects for this and all subsequent experiments.

We examined the midpoint using a VL (2) × continuum (7) within-subject ANOVA. A main effect of VL ($F(1,23)=80.4$, $\eta_p^2=0.78$, $p<0.001$) confirmed the observed difference between the VL conditions, with the boundary shifting 0.49 VOT steps (2.5 ms VOT) toward the voiceless end of the continuum for long VLs. There was also a main effect of continuum ($F(6,138)=32.3$, $\eta_p^2=0.58$, $p<0.001$). This is not unusual since the continua varied in vowel quality, formant frequencies, and overall duration, all of which influence voicing perception. The VL × continuum interaction was significant ($F(6,138)=4.71$, $\eta_p^2=0.17$, $p<0.001$), indicating that the size of the VL effect varied between continua. Follow-up tests within

each continuum found significant VL effects for bath-path, beach-peach, bet-pet, bike-pike, buck-puck, and a marginal effect for *beak-peak* (Table 2).

Eye movements—To estimate the time when each cue affected lexical activation, we used a technique similar to McMurray et al. (2008b). First, we computed *b-p bias*, the difference in proportion of looks to the /b/ and /p/ objects every 4 ms over the course of the trial. Next, we computed a measure of the effect of VOT and VL on *b-p bias* at each time-step. The VOT effect was the slope of a linear regression relating *b-p bias* to VOT step. The VL effect was the difference between long and short VL conditions, since there were only two levels. Figure 2 shows the effect of each cue over time and suggests that the VOT effect preceded VL effect, consistent with McMurray et al. (2008b) and the cue-integration hypothesis.

Finally, we applied the jackknife procedure, commonly used to measure the onset of ERP components (Miller, Patterson, & Ulrich, 1998; see Mordkoff & Gianaros, 2000, and Luck, 2005, for reviews; see McMurray et al., 2008b; Apfelbaum et al., 2011, for application to the visual world paradigm) to determine whether the timing of each effect was different. We first computed the average effect of VOT and VL over the entire dataset with one subject excluded. Next, we measured the onset of the effect as the time it crossed 50% of its maximum value. This procedure was repeated, excluding each subject in turn, yielding a dataset of the same sample size as the original. A paired t-test was then used to compare the two effect sizes, though the error term was adjusted to reflect the fact that each data point is derived from N-1 subjects.

VOT had a significantly earlier onset (705 ms) than VL (857 ms; $t_{jackknifed}(23)=3.03$, $p=0.007$). Thus, VL did not affect responses until after VOT, consistent with the cue-integration prediction. Moreover, the difference in the onset of the two effects (152 ms) is close to the earliest time when both cues could be used (136 ms; the difference between the maximum VOT and the end of the short vowel), suggesting that each cue was used as it became available.

Discussion

These results replicate those of McMurray et al. (2008b), demonstrating that the effect of VOT precedes that of VL, and they extend them to a larger set of stimuli. The temporal asynchrony of the effects argues against rate normalization accounts: listeners use each cue as it becomes available, indicating that they do not interpret VOT relative to VL. Next, we ask whether asynchronous effects are also observed in naturally-produced speech and whether the effect of VL is smaller, testing two additional predictions of the cue-integration approach.

Experiment 2: Natural speech

Experiment 2 used the same design as Experiment 1 but with stimuli constructed from natural speech. Here, we examine whether there is an effect of VL in naturally-produced speech and, if so, whether it shows the same timecourse as synthetic speech. Crucially, the presence of a smaller VL effect in naturally-produced speech would be difficult to rectify with normalization accounts. In addition, confirming that the earlier timecourse results apply to naturally-produced speech would provide an important extension of McMurray et al. (2008b) and Experiment 1 and offer additional evidence against speaking rate normalization.

Methods

Participants—Participants were recruited similarly to Experiment 1. Twenty-three completed the experiment, but four were excluded due to poor quality eye-tracks and two were excluded because they made no *buck* responses in the *buck/puck* continuum. This left 17 participants in the final dataset.

Design and task—The design and the task were the same as Experiment 1.

Stimuli—Stimuli were based on the same sets of minimal pair and unrelated words used in Experiment 1. The talker (author B.M.) was seated comfortably in a quiet room and wore a head-mounted microphone attached to a Kay CSL 4501. Recordings were made at 22.5 kHz. The talker recorded several tokens of each word in the context of the sentence "Click on the *x*" and the best token was selected as the basis of the stimuli.

Recordings were edited using Praat (Boersma & Weenink, 2009). The carrier phrase from each recording was removed and 100 ms of silence was added before the onset of the word. The stimuli were cross-spliced to create nine VOT steps (as in McMurray et al., 2008a). Nine points from approximately 0 ms to 40 ms (5 ms steps) after the onset of the word were marked at the nearest zero crossing for both the voiced and voiceless tokens. Words were cut at the marked points, and the beginning of each voiced token was replaced by a corresponding amount of aspiration from the voiceless token (see http://www.psychology.uiowa.edu/faculty/mcmurray/publications/matss_supplement/ for details).[5]

After creating the continua, words were modified using the pitch-synchronous overlap-add method (Moulines & Carpentier, 1990) to create long and short VLs. The period from the onset of the vowel steady state to the onset of energy from the following consonant was measured and lengthened or shortened by 40% to create the two VL conditions. Table 1 shows the VOT and VL values for the stimuli. The mean VL difference was 180 ms, longer than the 100 ms difference used in Experiment 1. If this has any effect on listeners' categorization responses, the VL effect should be larger than in Experiment 1.[6]

Results

Mouse-click responses—An effect of VL on the VOT boundary was observed (Figure 3). As in Experiment 1, logistic functions were fit to the data. A VL (2) × continuum (7)

[5]The cutback procedure requires that sounds be cut at zero-crossings. Thus, it is generally not possible to create continua manipulated in exactly 5-ms steps (as zero-crossings are not always available at exactly 5 ms increments). Across the set of continua, our stimuli were close to the VOT values used for the synthetic speech in Experiment 1 (see Table 1). This procedure can also produce stimuli with VOT values that are similar for the first two steps if the voiced token has a small positive VOT. This is what occurred for the *beach-peach* stimuli, which is why both have a 5 ms VOT (step 1 has the burst from the voiced token and step 2 has the burst from the voiceless token).
For the *best-pest* and *bet-pet* continua, the voiceless token did not cross a zero point at its onset. Thus, for the second step in each continuum, the burst form the voiced token was removed but not replaced with anything, and the stimuli from steps 2 through 9 varied from 0 to approximately 35 ms rather than from 5 to 40 ms. To make these two continua more consistent with the others, steps 2 through 9 were analyzed as the first eight steps of the continuum, and the original step 1 was excluded from analysis.
Also, a splicing error for step 4 of the *beach-peach* continuum caused it to have a shorter VOT than it should have. Trials with this stimulus were excluded from analysis.

[6]Besides stimulus naturalness, one other difference between the synthetic speech in Experiment 1 and the naturally-produced speech in Experiment 2 is the overall difference between the VL conditions, which was larger on average in Experiment 2. (Experiment 2 was designed before the comparisons with synthetic speech were conceived. Thus, a longer VL was used than in Experiment 1, which was designed to replicate the stimulus parameters from McMurray et al. [2008b] rather than match the parameters used in Experiment 2.) Although we might expect that this would lead to a larger VL effect, it could produce the smaller effect that was observed if listeners considered it an unnatural difference in rate and relied less on VL as a result. This would not provide an explanation for the results of Experiment 3, which shows different VL effects for the same-sized VL differences. However, it is possible that both factors affect the size of the VL effect.

within-subject ANOVA was run on the estimated boundaries, and a significant effect of VL ($F(1,16)=23.9$, $\eta_p^2=0.60$, $p<0.001$) was found. As in Experiment 1, long VLs shifted the boundary toward the voiceless end of the continuum, though this shift was only 0.30 VOT steps (≈1.5 ms VOT) compared to 0.49 VOT steps in Experiment 1. There was also a significant effect of continuum ($F(6,96)=7.6$, $\eta_p^2=0.32$, $p<0.001$) but no interaction ($F(6,96)=1.3$, $\eta_p^2=0.075$, $p=0.27$). Thus, using measures similar to those of previous studies, the results of this experiment indicate that there is a small effect of VL in naturally-produced speech.

To determine if this effect was smaller than in Experiment 1, we combined the results of both experiments in a mixed design ANOVA with the difference between the boundaries in the two VL conditions (i.e., the size of the VL effect) as the dependent variable, continuum as a within-subject factor, and experiment as a between-subject factor. There was a main effect of experiment ($F(1,39)=5.7$, $\eta_p^2=0.13$, $p=0.022$), confirming a smaller VL effect for the naturally-produced stimuli. There was also a main effect of continuum ($F(6,234)=3.9$, $\eta_p^2=0.091$, $p=0.001$), indicating that individual continua showed different-sized VL effects. Finally, there was a marginal experiment × continuum interaction ($F(6,234)=1.9$, $\eta_p^2=0.047$, $p=0.076$), indicating that the effects for specific continua may have differed between the two experiments.

Eye movements—The eye-movement data were analyzed in the same way as Experiment 1. Figure 4 shows the time course of VOT and VL effects. The effect of VOT was significantly earlier than VL ($t_{jackknifed}(16)=4.99$, $p<0.001$) with mean onsets similar to those in Experiment 1 (VOT: 666 ms; VL: 842 ms). As before, the difference in the onset of the two effects (176 ms) is close to the difference between the maximum VOT and the end of the short vowel (135 ms).

Discussion

The results of this experiment support two predictions of the cue-integration approach. First, the mouse-click responses show that VL is used for voicing judgments with naturally-produced speech. Second, the effect of VL is smaller in naturally-produced speech indicating that there is not a fixed relationship between VOT and VL. Finally, the timecourse of cue use is similar to Experiment 1, suggesting that each cue is used independently.

The smaller effect of VL for naturally-produced speech relative to synthetic speech also argues against most normalization approaches, since they predict either a fixed relationship between VOT and VL or an absence of VL effects in naturally-produced speech. A normalization approch that allows for an additional mechanism by which listeners weight VL less in naturally-produced speech (based on the assumption that listeners are less dependent on rate information) could explain these data. However, this is much more like a cue integration mechanism than a true context effect by which a lawful relationship between cue and context is used to compesate for contextual variability. It also constrasts with the prediction of the cue-integration hypothesis that the smaller VL effect observed with naturally-produced stimuli is the result of additional cues that covary with VOT. This prediction is examined in the next section.

Experiment 3: Effect of covarying cues

The previous experiment demonstrated that the effect of VL in naturally-produced speech is smaller than in synthetic speech. The cue-integration approach predicts that this will occur if additional voicing cues covary with VOT in the naturally-produced speech. This can happen as a consequence of the cutback procedure used to create the VOT continua. By removing

portions from the beginning of the voiced endpoint to create longer VOTs, some voicing cues, like formant onset frequencies, could be correlated with VOT.

Acoustic measurements confirmed that this was the case for our naturally-produced stimuli (see Online Supplement S1). For example, Figure 5 shows that in natural speech F1 was higher for stimuli with higher VOTs, while in synthetic speech, this only holds for part of the continuum. This relationship is not due to articulatory factors, nor was it intentionally manipulated in our stimuli. Rather, since F1s naturally rise at stimulus onset, if the entire onset is present (a short VOT), F1 will tend to have a low frequency. In contrast, if most of the transition takes place during the aspiration (due to our cross-splicing operation), the onset frequency will be higher. This effect was small for our synthetic speech where the transitions occurred quickly. Thus, in our naturally-produced continua (and, to a lesser extent, in the synthetic versions), the variation in VOT was also signaled by systematic variation in F1.

Is this a sufficient explanation for the differences in perceptual effects between synthetic and naturally-produced speech? We examined this using simulations with the WGMM, a general cue-integration model that can be used with different acoustic cues and phonological contrasts (Toscano & McMurray, 2010), and found that this was the case (Figure 6 and Online Supplement S2). This model consists of separate Gaussian mixture models (GMMs) that are used to estimate the statistical distribution of each individual cue (e.g., VOT and VL). These distributions are used to determine how much weight to assign each cue: cues that have highly distinct clusters get more weight, and those with overlapping cues get less weight. Finally, individual cue inputs are weighted and linearly combined (as in weighting-by-reliability approaches in vision; Ernst & Banks, 2002; Jacobs, 2002). The combined input is used to train a GMM that learns phonological categories from the combination of cues.

Using this approach, we can simulate conditions similar to those in the preceding experiments. By training the model on several cues to voicing, we can test it under conditions in which a cue either covaries with VOT during testing (mimicking naturally-produced speech) or is held constant at an ambiguous value (mimicking synthetic speech).

The simulations showed that the apparent effect of VL is smaller when the third cue covaries with VOT (as variation along the VOT dimension reflects the combined influence of VOT and the third cue) than when it is held constant. Critically, in both cases, the weight and contribution of VL is the same; its contribution simply appears to be smaller when additional cues bias the response in the same direction as VOT. This result also suggests that without stimuli that have the covariance of cues in natural speech, it may be difficult to draw conclusions about the effect of naturalness.

Do listeners also show different-sized rate effects without reweighting VL? To answer this question, we constructed a new set of synthetic stimuli modeled after our acoustic analyses. In one condition, the formants had a very short rise time at the onset of the word. This produced stimuli that had a relatively constant formant onset frequency at most VOT steps, since the formants had reached their maximum value by the onset of voicing. In the other condition, formants had a slower rise time yielding more substantial changes in formant frequencies over the first 40 ms. When the onset of voicing is cut-back to manipulate VOT, this yields concurrent changes in formant onset frequencies. This produces stimuli that correspond to the naturally-produced continua used in Experiment 2. If the difference in the size of the VL effect as a function of naturalness is due to differences in other cues, it should be smaller if formant onset frequency varies with VOT (as in naturally-produced speech).

Most importantly, the only way a normalization model could account for the reduced effect in naturally-produced speech is by reweighting VL, though assigning a weight to the VL

dimension is itself similar to cue integration approaches. Shinn et al. (1985) found a smaller VL effect when stimuli with different VLs were presented in a blocked design, suggesting that listeners could learn to downweight VL. Thus, if synthetic and naturally-produced stimuli are presented in different experiments (as in Experiments 1 and 2), listeners could learn to weight VL differently in the two conditions based on the stimuli they are hearing in each experiment. However, if the conditions are presented randomly with no other cues to indicate the difference between them, listeners would not be able to re-weight cues for each condition over the course of the experiment. Thus, by using synthetic stimuli in a randomized design with formants onsets that either covary with VOT or are held constant, we can test whether changes in VL effects are due to cue re-weighting or the availability of additional cues.

Methods

Participants—15 monolingual native English were recruited in accordance with university human subject protocols and received either course credit or $10 for their participation.

Design—As in Experiments 1 and 2, participants performed a picture identification task. Stimuli consisted of a set of three minimal pair words (*batch-patch, bet-pet, buck-puck*) that varied in VOT, VL, and formant onset frequency (F1, F2, and F3). Three sets of /l/ (*light, lock, lake*) and /ʃ/ (*sheep, ship, shoe*) words served as unrelated items and varied in VL. Each stimulus was presented 4 times for a total of 864 trials. Participants completed the experiment in a single 1-hour session. Stimulus presentation equipment was the same as in the first experiment.

Stimuli—Stimuli were synthesized using the same software used in Experiment 1. VOT continua were created in the same way and ranged from 0 to 40 ms in 5 ms steps. VL was 110 ms in the short VL condition, and 210 ms in the long VL condition (Table 1). Formant frequencies increased linearly over syllable onset, starting and ending at the same frequency for all conditions for a particular continuum with the peak frequency determined by the vowel. The time it took the formants to reach their peak was varied by changing their rise time (either 10 or 40 ms), creating two conditions with either long or short formant transition durations (FTDs). Long FTDs created stimuli analogous to the naturally-produced stimuli (where formant onset co-varied with VOT) and short FTDs were analogous to synthetic stimuli (where it co-varied less).[7]

Task—This experiment was not concerned with the time course of processing, only with the effect of VL as a function of FTD. Thus, we used the same 4AFC task that was used in Experiments 1 and 2, but without eye-tracking.

Results

Figure 7 shows the average proportion of /p/ responses in each of the FTD conditions. A larger VL effect was observed for short FTDs, consistent with the prediction that fewer co-varying cues lead to larger VL effects. Logistic functions were again fit to participants' responses to determine category boundaries. A VL (2) × FTD (2) × continuum (3) within-subjects ANOVA found a significant effect of VL ($F(1,14)=21.65$, $p<0.001$, $\eta_p^2=0.61$), with the mean boundary in the long VL condition 0.38 VOT steps (1.9 ms) greater than for short VLs. VL did not interact with continuum ($F<1$), suggesting that when formant onsets are controlled, some of the variability in VL effects across words disappears.

[7]It seems plausible that FTD could also serve as a rate cue, and similar predictions would apply to it as well. However, the current experiment does not address this question.

There was a main effect of FTD ($F(1,14)=43.6$, $\eta_p^2=0.76$, $p<0.001$). This was due to the fact that when FTD was short, formants reached their vowel targets within the first few milliseconds after onset. As a result, formant onset frequencies were higher overall for short FTDs, leading to more voiceless responses. There was a significant effect of continuum ($F(2,28)=13.8$, $\eta_p^2=0.50$, $p<0.001$), suggesting that the category boundaries were different for the different continua, as observed in the previous experiments. In addition, there was a significant FTD × continuum interaction ($F(2,28)=9.4$, $\eta_p^2=0.40$, $p=0.001$), indicating that the overall effect of FTD on voicing differed across the continua, most likely because of the different vowels (and different onset F1-values) in each of them. Follow-up tests found an effect of FTD for *bet-pet* ($t(14)=5.9$, $p<0.001$), but only marginal effects for *batch-patch* ($t(14)=1.97$, $p=0.069$) and *buck-puck* ($t(14)=1.94$, $p=0.073$), suggesting that overall FTD may not have had an effect on the boundaries for the *batch-patch* and *buck-puck* continua.

Most importantly, there was a significant VL × FTD interaction ($F(1,14)=7.3$, $\eta_p^2=0.34$, $p=0.017$) with a larger VL difference for the short FTD condition than the long FTD condition. Paired t-tests were run for each FTD condition separately to assess the effect of VL. As predicted, for long FTDs (which covaried with VOT), no effect of VL was found ($t(14)=1.05$, $p=0.31$), and for short FTDs (which did not covary with VOT), a significant effect of VL was found (0.62 steps, 3.1 ms; $t(14)=6.3$, $p<0.001$). This confirms the prediction that the short FTDs would show a larger VL effect. The three-way interaction was not significant ($F(2,28)=1.72$, $\eta_p^2=0.109$, $p=.197$), suggesting that our primary effect was not different across continua.

Discussion

The results show that when additional cues covary with VOT, there is a smaller VL effect. This supports our hypothesis that differences in the size of the effect (or the absence of an effect) for VL can be due to differences in other cues that covary with VOT, providing an explanation for the different results obtained for synthetic and natural speech (see Note 6).

Another possible explanation for the difference between natural and synthetic speech is that listeners weight cues differently between these conditions. There are certainly situations in which listeners reweight cues during the course of an experiment (Francis & Nusbaum, 2002). Given this, normalization accounts might be able to predict that listeners reduce the importance of rate compensation processes when stimuli are natural (and rate compensation may not be needed). However, it is unclear how this could be done on a trial-by-trial basis when all the stimuli are presented randomly. In contrast, in our cue-integration framework, no additional mechanism is necessary to account for these differences. They derive directly from the use of multiple cues, a point that is underscored by the WGMM simulations.

Experiment 4: Small VL differences

The previous three experiments used VL differences that were similar to previous studies and reflect differences due to speaking rate. However, these VL differences are much greater than the phonetic difference in VL for word-initial sounds. Allen and Miller (1999) show a mean VL difference of 20 ms between voiced and voiceless sounds. This difference is small given the range of speaking rates that can be produced by a single talker (Kessinger & Blumstein, 1997; Miller, Grojean, & Lomanto, 1984; Magloire & Greene, 1999; Beckman et al., 2011). If VL is only used to estimate rate (as predicted by normalization approaches), a 20 ms difference is minimal and should show no effect on categorization. In contrast, if VL is treated as a cue, we should still see an effect with these small differences. To test this, our final experiment examined listeners' responses to synthetic stimuli with a 20 ms VL difference between conditions.

Methods

Participants—Twenty monolingual native English were recruited in accordance with university human subject protocols and received either course credit or $15 for their participation. Two subjects were excluded from analysis for having less than 80% correct responses on the endpoints of the VOT continua.

Design and task—The design and task were the same as Experiment 1 except that the eye-tracker was not used.

Stimuli—The stimuli were the same as those in Experiment 1 except that the difference in VL was 20 ms rather than 100 ms. The particular VL values used are shown in Table 1.

Results

Figure 8 indicates that listeners showed a small effect of VL (mean difference: 0.23 steps; 1.1 ms VOT). Logistic functions were fit to the data to obtain category boundaries. A VL (2) × continuum (6)[8] within-subjects ANOVA showed a main effect of VL ($F(1,17)=16.1$, $\eta_p^2=0.49$, $p=0.001$), confirming that VL affected voicing judgments, and a main effect of continuum ($F(5,85)=31.3$, $\eta_p^2=0.65$, $p<0.001$), since boundaries differed between continua. The interaction was not significant ($F<1$), indicating a similar VL effect across continua.

Discussion

This experiment confirms that small VL differences affect voicing categorization. These stimuli may be more characteristic of actual VL differences between voicing categories, such as those found by Allen and Miller (1999), providing additional evidence that VL is used as a phonetic cue rather than as context information.

General discussion

These experiments provide converging evidence for cue-integration as the mechanism that underlies the influence of speaking rate on voicing judgments. First, the time course data of Experiments 1 and 2 show independent effects of VOT and VL. This is difficult to rectify with normalization explanations that require the two sources of information to be combined before accessing higher level categorical or lexical representations. Second, we found smaller VL effects with naturally-produced speech. This suggests that listeners are not simply failing to use VL in naturally-produced speech, but rather, that its apparent contribution is smaller. This contrasts with normalization accounts that predict a fixed relationship between VOT and VL. Crucially, as Experiment 3 and our simulations show, this can be accounted for quite simply in a cue-integration model: when two cues are providing the same information (e.g. VOT and F1 onset), the apparent contribution of the third (VL) is decreased, and this can be observed when listeners do not have the opportunity to reweight cues. Finally, Experiment 4 showed an effect of VL for differences that are smaller than those that are relevant for speaking rate but are quite reasonable if listeners are using VL to distinguish voicing categories.

Context effects in speech

Phonetic cues vs. context effects—The critical finding from this study is that listeners treat VL as a phonetic cue rather than as a true context effect. This contrasts with previous normalization approaches, which argued that the perceptual effect of VL is to modify

[8]The *bike/pike* continuum was not analyzed because it was discovered that its F2 and F3 onsets inadvertently covaried with VL (due to formant changes for the diphthong).

listeners' use of other cues (Summerfield, 1981), not to directly cue a voiced or voiceless sound. Treating VL as a phonetic cue, similar to VOT, F1, or F0, allows us to explain previous data and the results of the current study in terms of general cue-integration principles.

Although the present experiments rule out the most extreme version of VL as a context effect (in which context compensation is obligatory) we cannot rule out all types of rate normalization. VL could serve as a context effect if listeners initially use VOT information by itself, assume a neutral rate or prior distribution of rates, and update their VOT estimate as VL information is received. This would be consistent with our results. Without additional mechanisms though, it likely cannot account for the effect of stimulus naturalness or small VL differences, and a cue-integration mechanism offers a more complete and parsimonious solution.

Reconciling previous results—These results explain some conflicting prior work. The lack of a VL effect in natural or natural-sounding synthetic speech has a simple explanation: when other cues covary with VOT (as they do in naturally-produced speech), the apparent effect of VL is smaller. This mirrors the results of previous studies (e.g., Shinn et al., 1985) and builds on them by extending them to voicing and showing that this effect is seen even if listeners do not reweight cues. The effects of VL are most apparent at the voiced/voiceless category boundary. Therefore, we agree with Boucher (2002) that this may explain why Utman (1998) did not find a VL effect in natural speech when examining voiceless tokens, while he did find an effect when testing a single /d/ token with a nearly ambiguous VOT.

Other context effects—The cue-integration approach also has implications for other types of context effects. Speaking rate is classically considered an example of broader normalization processes, where one cue-value is interpreted with respect to surrounding context. Our results suggest that such compensation is not obligatory — listeners do not need to wait until they have the information required to normalize a cue before they can begin using it to make preliminary inferences. This is consistent with exemplar models (Goldinger, 1998; Johnson, 1997) as well as compensation approaches like C-CuRE (McMurray & Jongman, 2011) that can use cues in relatively unmodified forms when context information is not available.

While it remains to be seen whether this applies to other context effects, our results raise this as a possibility. Thus, many other factors that are typically thought of as context (e.g., talker identity), may actually be first-order cues to categorization. Indeed, well known influences of talker identity on phoneme or lexical identification (e.g., Palmeri, Goldinger, & Pisoni, 1993; Creel, Aslin & Tanenhaus, 2008) may simply be the result of indexical cues serving as direct cues to phonemes or words (see also Apfelbaum & McMurray, 2011).

Methodological considerations—The results also suggest that, in some cases, listeners' responses to synthetic stimuli may not reflect how they process natural speech. This is not to say that studies using synthetic stimuli are uninformative. Indeed, our acoustic measurements of natural stimuli suggest the contrary: synthetic stimuli may allow researchers to isolate and manipulate certain cues better. Rather, researchers should remember that varying synthetic stimuli along a single dimension, holding other cues constant, may not reflect the way speech sounds vary naturally. Other authors have made similar points about the construction of VOT continua in particular (Kessinger & Blumstein, 1998; Allen & Miller, 1999). Both naturally-produced and synthetic stimuli may also contain unnatural covariances between cues depending on the conditions in an experiment. Given these issues, we argue that approaches harnessing both types of stimuli are likely to be more informative than those relying on one or the other.

Approaches to cue integration

Cue weighting—The simulations with the WGMM and the results of Experiment 3 both suggest that listeners do not need to reweight cues in order to produce different-sized VL effects. Other models of cue integration (Nearey & Hogan, 1986; Nearey & Assmann, 1986; Nearey, 1997; Oden & Massaro, 1978; Massaro & Oden, 1980; Smits, 2001a) combine cues in similar ways as part of computing the probability of each category. As a result, they can be expected to perform similarly to the WGMM. Thus, the results of Experiment 3 coupled with these models offer a critical insight into cue integration: cues do not need to be reweighted in order to produce a change in the size of each cue's effect. This is consistent with the cue-integration approach and contrasts with explanations for the difference between natural and synthetic speech that suggest that listeners rely less on VL in natural speech.

More broadly, the results suggest that listeners assign some weight to subtle and relatively unreliable acoustic cues. Experiment 3 suggests they can take advantage of differences in the juncture between aspiration and voicing, and Experiment 4 suggests they can use fairly small differences in a highly variable cue (VL) to bias voicing judgments. This reflects listeners' sensitivity to fine-grained acoustic differences (McMurray et al., 2002; Toscano, McMurray, Dennhardt, & Luck, 2010) and demonstrates that they can use weak cues even when a much stronger cue (e.g., VOT) is normally available for a given phonological distinction.

Continuous processing—The cue-integration approach raises the question of how multiple cues are combined during online processing, and the visual world paradigm offers a powerful way to assess this by examining the timecourse of integration. While understanding cue-integration in real time was not a fundamental goal of this study, the results are relevant to these issues. McMurray et al. (2008b) suggest that cue integration could occur via two mechanisms. First, in an *integrate-and-store* approach multiple cues are combined *before* they are used to update phonemic or lexical representations. This can reduce the risk of committing to an incorrect lexical representation (McMurray, Tanenhaus & Aslin, 2009b). Second, cues could be integrated via a *continuous-cascade* approach in which partial phonemic or lexical decisions (i.e., those made on the basis of only some cues) occur as soon as information is available and are updated continuously.

The latter approach allows lexical representations to be accessed more quickly, but with the risk that preliminary states are likely to be less accurate. This reduced accuracy is mitigated by not fully committing at any point in time and updating as more information arrives. The data from Experiments 1 and 2 rule out a completely buffered system for VOT and VL, though we cannot rule out systems that combine the two strategies (e.g., with cues integrated by sub-lexical units that continuously cascade preliminary states to lexical activation). However, this would be functionally equivalent to systems without sub-lexical integration. Thus, a continuous model seems more parsimonious.

These distinctions have generally not been considered by previous or current models of cue integration (though see Smits, 2001b), and in the cue-integration approach described here, temporal order of the cues is not taken into account. Nonetheless, these issues will be important as we translate principles of cue-integration into models of real-time perceptual processing.

Conclusion

In contrast to previous work, the results of this study show that listeners do use VL for word-initial voicing judgments in naturally-produced speech, but that they do not interpret VOT relative to this rate information, as predicted by normalization approaches. Rather, the

present study provides evidence that cue-integration mechanisms can best explain effects of speaking rate variability on voicing judgments: VOT and VL simply participate as two independent, additive cues. Thus, there is no need to posit intermediate processes that re-compute VOT relative to rate or assume that VOT boundaries are adjusted as a function of speaking rate.

The results demonstrate that independent effects of each cue can be observed during speech processing in both synthetic and naturally-produced speech, they provide an explanation for why behavioral results differ depending on the naturalness of the stimuli, and they show that these differences can be obtained without reweighting cues. Together, these results indicate that listeners do not encode VOT relative to rate information provided by VL, and they support models of speech perception in which multiple sources of information contribute to spoken word recognition in-the-moment as a function of their reliability as phonetic cues.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We would like to thank Kristine Kovack-Lesh, Molly Robinson, and Meghan Delaney for help with stimulus preparation and acoustic measurements; and William McEchron and the members of the MACLab for assistance with data collection. This research was supported by NIH DC008089 to BM.

References

Allen JS, Miller JL. Effects of syllable-initial voicing and speaking rate on the temporal characteristics of monosyllabic words. Journal of the Acoustical Society of America. 1999; 106:2031–2039. [PubMed: 10530026]

Allen JS, Miller JL, DeSteno D. Individual talker differences in voice-onset-time. Journal of the Acoustical Society of America. 2003; 113:544–552. [PubMed: 12558290]

Andruski JE, Blumstein SE, Burton M. The effect of subphonetic differences on lexical access. Cognition. 1994; 52:163–187. [PubMed: 7956004]

Apfelbaum K, Blumstein S, McMurray B. Semantic priming is affected by real-time phonological competition: Evidence for continuous cascading systems. Psychonomic Bulletin & Review. 2011; 18:141–149. [PubMed: 21327343]

Apfelbaum KA, McMurray B. Using variability to guide dimensional weighting: Associative mechanisms in early word learning. Cognitive Science. 2011; 35:1105–1138. [PubMed: 21609356]

Beckman J, Helgason P, McMurray B, Ringen C. Rate effects on Swedish VOT: Evidence for phonological overspecification. Journal of Phonetics. 2011; 39:39–49.

Bernstein, R. The Eyelink PsyExtension. 2000. Available at: http://www.psych.uni-potsdam.de/cognitive/eyetracker/eyelink-psyextension.html

Boersma, P.; Weenink, D. Praat: doing phonetics by computer. 2009. Available at : http://www.praat.org

Boucher VJ. Timing relations in speech and the identification of voice-onset times: A stable perceptual boundary for voicing categories across speaking rates. Perception & Psychophysics. 2002; 64:121–130. [PubMed: 11916295]

Cho T, Ladefoged P. Variation and universals in VOT: Evidence from 18 languages. Journal of Phonetics. 1999; 27:207–229.

Cohen JD, MacWhinney B, Flatt M, Provost J. PsyScope: A new graphic interactive environment for designing psychology experiments. Behavioral Research Methods, Instruments, and Computers. 1993; 25:257–271.

Creel S, Aslin RN, Tanenhaus MK. Heeding the voice of experience: The role of talker variation in lexical access. Cognition. 2008; 106:633–664. [PubMed: 17507006]

Diehl RL, Souther AF, Convis CL. Conditions on rate normalization in speech perception. Perception & Psychophysics. 1980; 27:435–443. [PubMed: 7383831]

Dupoux E, Green K. Perceptual adjustment to highly compressed speech: Effects of talker and rate changes. Journal of Experimental Psychology: Human Perception & Performance. 1997; 23:914–927. [PubMed: 9180050]

Francis AL, Nusbaum HC. Selective attention and the acquisition of new phonetic categories. Journal of Experimental Psychology: Human Perception & Performance. 2002; 28:349–366. [PubMed: 11999859]

Goldinger SD. Echoes of echoes? An episodic theory of lexical access. Psychological Review. 1998; 105:251–279. [PubMed: 9577239]

Goldrick M, Blumstein SE. Cascading activation from phonological planning to articulatory processes: Evidence from tongue twisters. Language & Cognitive Processes. 2006; 21:649–683.

Hillenbrand JM, Clark MJ, Houde RA. Some effects of duration on vowel recognition. Journal of the Acoustical Society of America. 2000; 106:3013–3022. [PubMed: 11144593]

Holt LL, Lotto AJ. Speech perception as categorization. Attention, Perception, & Psychophysics. 2010; 72:1218–1227.

House AS. On vowel duration in English. Journal of the Acoustical Society of America. 1961; 33:1174–1178.

Johnson, K. Speech perception without speaker normalization.. In: Johnson, K.; Mullennix, J., editors. Talker Variability in Speech Processing. Academic Press; New York: 1997. p. 145-165.

Kessinger RH, Blumstein SE. Effects of speaking rate on voice-onset time in Thai, French , and English. Journal of Phonetics. 1997; 25:143–168.

Kessinger RH, Blumstein SE. Effects of speaking rate on voice-onset time and vowel production: Some implications for perception studies. Journal of Phonetics. 1998; 26:117–128.

Klatt DH. Software for a cascade/parallel formant synthesizer. Journal of the Acoustical Society of America. 1980; 67:971–995.

Lisker L, Abramson AS. A cross-linguistic of voicing in initial stops: Acoustical measurements. Word. 1964; 20:384–422.

Luck, SJ. An Introduction to the Event-Related Potential Technique. MIT Press; Cambridge, MA: 2005.

Magloire J, Greene KP. A cross-language comparison of speaking rate effects on the production of voice onset time in English and Spanish. Phonetica. 1999; 56:158–185.

Massaro DW, Oden GC. Evaluation and integration of acoustic features in speech perception. The Journal of the Acoustical Society of America. 1980; 67:996–1013. [PubMed: 7358922]

McMurray, B. KlattWorks: A [somewhat] new systematic approach to formant-based speech synthesis for empirical research. 2008. Manuscript in preparation

McMurray B, Aslin RN, Tanenhaus MK, Spivey MJ, Subik D. Gradient sensitivity to within-category variation in words and syllables. Journal of Experimental Psychology: Human Perception & Performance. 2008a; 34:1609–1631. [PubMed: 19045996]

McMurray B, Clayards M, Tanenhaus MK, Aslin RN. Tracking the time course of phonetic cue integration during spoken word recognition. Psychonomic Bulletin & Review. 2008b; 15:1064–1071. [PubMed: 19001568]

McMurray B, Jongman A. What information is necessary for speech categorization? Harnessing variability in the speech signal by integrating cues computed relative to expectations. Psychological Review. 2011; 118:219–246. [PubMed: 21417542]

McMurray B, Samelson V, Lee S, Tomblin JB. Eye-movements reveal the time-course of online spoken word recognition language impaired and normal adolescents. Cognitive Psychology. 2010; 60:1–39. [PubMed: 19836014]

McMurray B, Tanenhaus MK, Aslin RN. Gradient effects of within-category phonetic variation on lexical access. Cognition. 2002; 86:B33–B42. [PubMed: 12435537]

McMurray B, Tanenhaus MK, Aslin RN. Within-category VOT affects recovery from “lexical” garden-paths: Evidence against phoneme-level inhibition. Journal of Memory & Language. 2009b; 60:65–91. [PubMed: 20046217]

Miller J, Dexter E. Effects of speaking rate and lexical status on phonetic perception. Journal of Experimental Psychology: Human Perception & Performance. 1988; 14:369–378. [PubMed: 2971767]

Miller JL, Green KP, Reeves A. Speaking rate and segments: A look at the relation between speech production and speech perception for the voicing contrast. Phonetica. 1986; 43:106–115.

Miller JL, Grojean F, Lomanto C. Articulation rate and its variability in spontaneous speech: A reanalysis and some implications. Phonetica. 1984; 41:215–225. [PubMed: 6535162]

Miller JL, Liberman AM. Some effects of later-occurring information on the perception of stop consonant and semivowel. Perception & Psychophysics. 1979; 25:457–465. [PubMed: 492910]

Miller JL, Volaitis LE. Effect of speaking rate on the perceptual structure of a phonetic category. Perception & Psychophysics. 1989; 46:505–512. [PubMed: 2587179]

Miller JL, Wayland SC. Limits on the limitations of context-conditioned effects in the perception of [b] and [w]. Perception & Psychophysics. 1993; 54:205–210. [PubMed: 8361836]

Miller JO, Patterson T, Ulrich R. Jackknife-based method for measuring LRP onset latency differences. Psychophysiology. 1998; 35:99–115. [PubMed: 9499711]

Mordkoff JT, Gianaros P. Detecting the onset of the lateralized readiness potential: A comparison of available methods and procedures. Psychophysiology. 2000; 37:347–360. [PubMed: 10860412]

Moulines E, Carpentier F. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. Speech Communication. 1990; 9:453–467.

Nearey TM. Speech perception as pattern recognition. Journal of the Acoustical Society of America. 1997; 101:3241–3254. [PubMed: 9193041]

Nearey TM, Assmann PF. Modeling the role of inherent spectral change in vowel identification. Journal of the Acoustical Society of America. 1986; 80:1297–1308.

Nearey, TM.; Hogan, J. Phonological contrast in experimental phonetics: Relating distributions of measurements in production data to perceptual categorization curves.. In: Ohala, J.; Jaeger, J., editors. Experimental Phonology. Academic Press; New York: 1986.

Oden GC, Massaro DW. Integration of feature information in speech perception. Psychological Review. 1978; 85:172–191. [PubMed: 663005]

Palmeri TJ, Goldinger SD, Pisoni DB. Episodic encoding of voice attributes and recognition memory for spoken words. Journal of Experimental Psychology: Learning, Memory, and Cognition. 1993; 19:309–328.

Peterson GE, Lehiste I. Duration of syllable nuclei in English. Journal of the Acoustical Society of America. 1960; 32:693–703.

Pind J. Speaking rate, VOT and quantity: The search for higher-order invariants for two Icelandic speech cues. Perception & Psychophysics. 1995; 57:291–304. [PubMed: 7770321]

Pisoni DB, Tash J. Reaction times to comparisons within and across phonetic categories. Perception & Psychophysics. 1974; 15:285–290. [PubMed: 23226881]

Repp BH. Phonetic trading relations and context effects: New experimental evidence for a speech mode of perception. Psychological Bulletin. 1982; 82:81–110. [PubMed: 7134330]

Repp BH, Liberman AM, Eccardt T, Pesetsky D. Perceptual integration of acoustic cues for stop, fricative, and affricate manner. Journal of Experimental Psychology: Human Perception & Performance. 1978; 4:621–637. [PubMed: 722252]

Salverda AP, Dahan D, McQueen JM. The role of prosodic boundaries in the resolution of lexical embedding in speech comprehension. Cognition. 2003; 90:51–89. [PubMed: 14597270]

Shinn PC, Blumstein SE, Jongman A. Limitations of context conditioned effects in the perception of [b] and [w]. Perception & Psychophysics. 1985; 38:397–407. [PubMed: 3831918]

Smiljanic R, Bradlow A. Stability of temporal contrasts across speaking styles in English and Croatian. Journal of Phonetics. 2008; 30:91–113. [PubMed: 19122747]

Smits R. Evidence for hierarchical categorization of coarticulated phonemes. Journal of Experimental Psychology: Human Perception & Performance. 2001a; 27:1145–1162. [PubMed: 11642700]

Smits R. Hierarchical categorization of coarticulated phonemes: a theoretical analysis. Perception & Psychophysics. 2001b; 63:1103–1139.

Summerfield Q. Articulatory rate and perceptual constancy in phonetic perception. Journal of Experimental Psychology: Human Perception & Performance. 1981; 7:1074–1095. [PubMed: 6457109]

Toscano JC, McMurray B. Cue integration with categories: Weighting acoustic cues in speech using unsupervised learning and distributional statistics. Cognitive Science. 2010; 34:436–464.

Toscano JC, McMurray B, Dennhardt J, Luck SJ. Continuous perception and graded categorization: Electrophysiological evidence for a linear relationship between the acoustic signal and perceptual encoding of speech. Psychological Science. 2010; 21:1532–1540. [PubMed: 20935168]

Turk, A.; Nakai, S.; Sugahara, M. Acoustic segment durations in prosodic research: A practical guide.. In: Sudhoff, S., et al., editors. Methods in Empirical Prosody Research. Walter de Gruyter; Berlin: 2006.

Utman JA. Effects of local speaking rate context on the perception of voice-onset time in initial stop consonants. Journal of the Acoustical Society of America. 1998; 103:1640–1653. [PubMed: 9514028]

Warren P, Marslen-Wilson W. Cues to lexical choice: Discriminating place and voice. Perception & Psychophysics. 1988; 43:21–30. [PubMed: 3340495]

Wayland SC, Miller JL, Volaitis LE. The influence of sentential speaking rate on the internal structure of phonetic categories. Journal of the Acoustical Society of America. 1994; 95:2694–2701. [PubMed: 8207142]

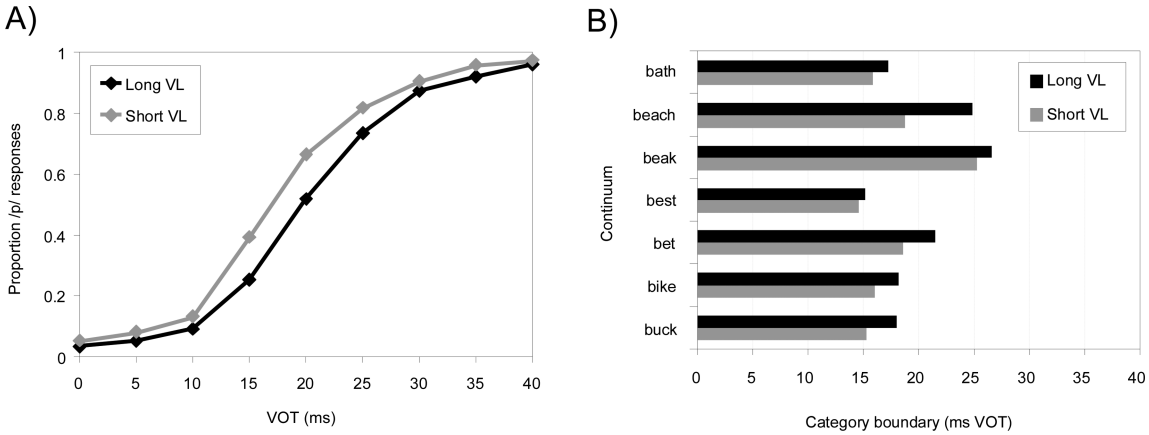

Figure 1.

(A) Proportion /p/-item responses as a function of VOT and VL in Experiment 1 showing a shift in the identification function for the two different VLs (i.e. more /p/ responses for short VLs; more /b/ responses for long VLs). (B) Category boundaries for each continuum in Experiment 1 as a function of VL.

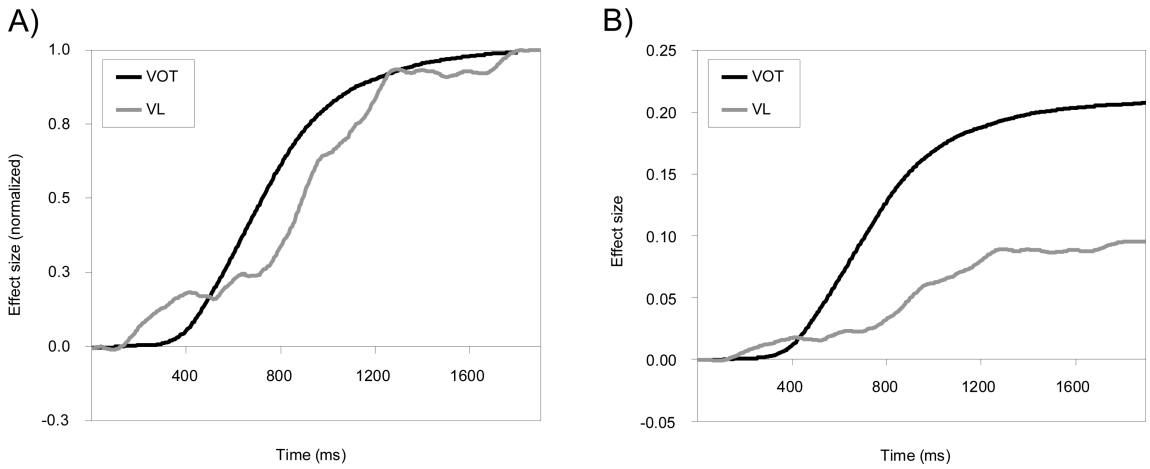

Figure 2.

(A) Time course of cue use for Experiment 1 with effect size normalized for each cue. The effect of VOT emerges earlier than the effect of VL. (B) Time course of cue use with raw effect sizes.

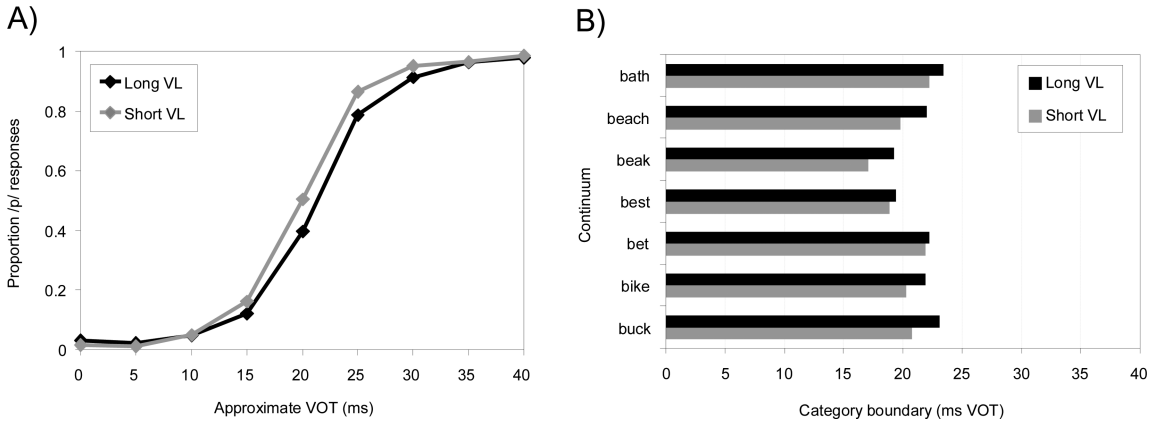

Figure 3.
(A) Proportion of /p/ responses in Experiment 2 as a function of VOT and VL. (B) Category boundaries for each continuum in Experiment 2 as a function of VL.

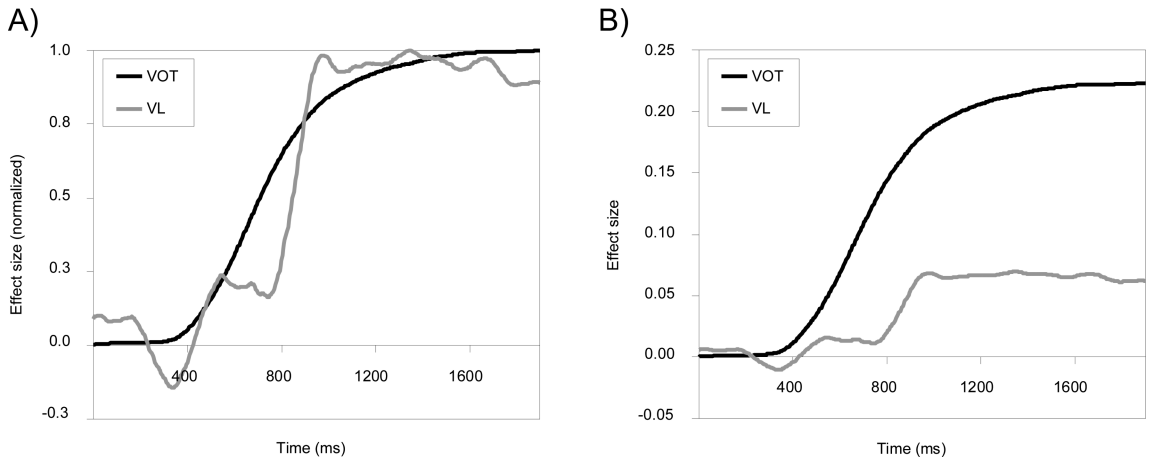

Figure 4.
(A) Time course of cue use in Experiment 2 with effect size normalized for each cue, showing that the effect of VOT precedes the effect of VL. (B) Unnormalized effect sizes.

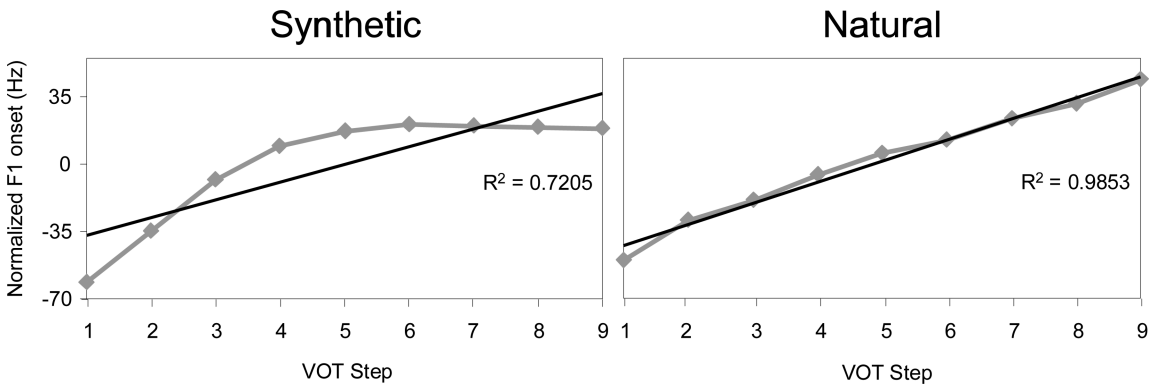

Figure 5.
Normalized F1 onset frequencies as a function of VOT and stimulus type. See Online Supplement S1 for additional details and measurements.

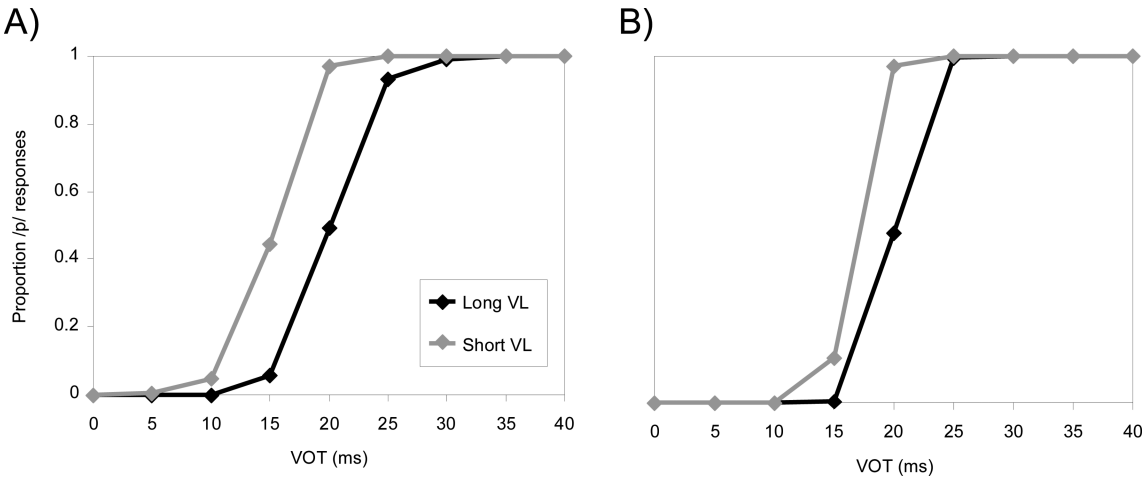

Figure 6.
(A) Model categorization responses when additional cues are held constant at an ambiguous value. (B) Model responses when those cues covary with VOT.

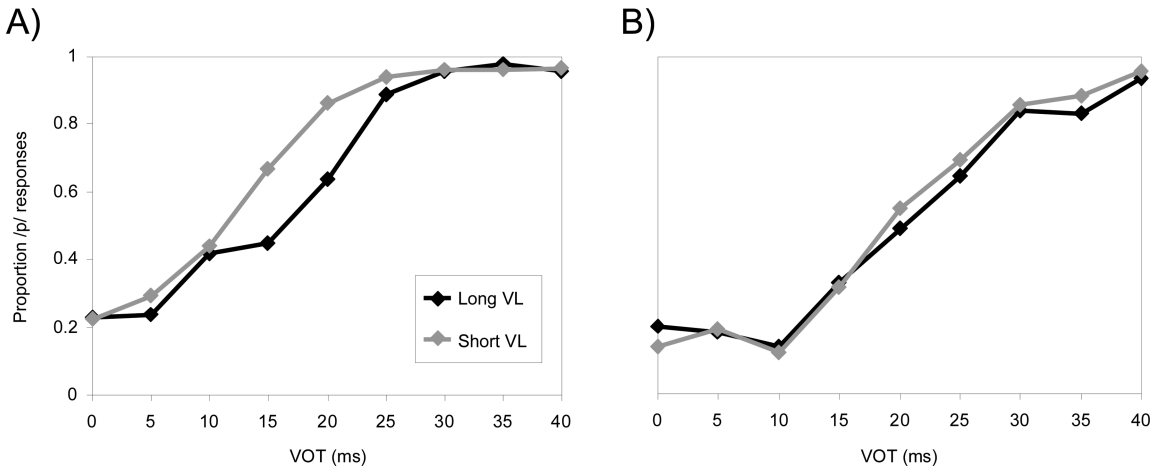

Figure 7.

(A) Proportion of /p/ responses in Experiment 3 as a function of VOT and VL for the short FTD (i.e., less variable formant onset) condition. (B) Responses for the long FTD (co-varying with VOT) condition.

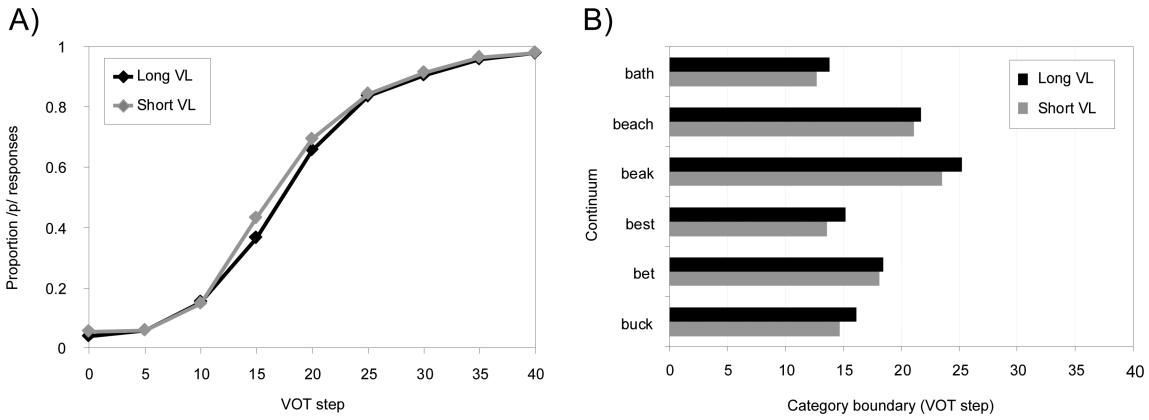

Figure 8.
(A) Proportion of /p/ responses in Experiment 4 as a function of VOT and VL. (B) Category boundaries for each word pair in Experiment 4 as a function of VL.

Table 1

Characteristics of the stimuli used in each experiment.

		VOT (ms)[a]	Release-to-offset VL (ms)	Word length (ms; short and long VL conditions)
Exp. 1	*bath-path*	0, 5, 10, 15, 20 25, 30, 35, 40	225, 325	321, 421
	beach-peach	0, 5, 10, 15, 20 25, 30, 35, 40	130, 230	329, 429
	beak-peak	0, 5, 10, 15, 20 25, 30, 35, 40	158, 258	327, 427
	best-pest	0, 5, 10, 15, 20 25, 30, 35, 40	169, 269	371, 471
	bet-pet	0, 5, 10, 15, 20 25, 30, 35, 40	141, 241	304, 404
	bike-pike	0, 5, 10, 15, 20 25, 30, 35, 40	203, 303	387, 487
	buck-puck	0, 5, 10, 15, 20 25, 30, 35, 40	236, 336	400, 500
Exp. 2	*bath-path*	4, 5, 9, 14, 18, 23, 31, 35, 40	209, 422	305, 518
	beach-peach	5, 5, 11, n/a, 23, 30, 34, 43, 45	155, 296	395, 536
	beak-peak	4, 7, 12, 17, 18, 25, 30, 36, 41	184, 368	390, 574
	best-pest	0, 5, 13, 16, 19, 22, 28, 37, n/a	152, 268	412, 548
	bet-pet	0, 7, 11, 14, 19, 27, 33, 38, n/a	185, 366	325, 506
	bike-pike	0, 4, 12, 17, 23, 27, 34, 39, 46	222, 449	426, 653
	buck-puck	0, 4, 8, 14, 19, 26, 32, 37, 40	179, 358	313, 492
Exp. 3	*batch-patch*	0, 5, 10, 15, 20 25, 30, 35, 40	155, 255	364, 464
	bet-pet	0, 5, 10, 15, 20 25, 30, 35, 40	155, 255	256, 356
	buck-puck	0, 5, 10, 15, 20 25, 30, 35, 40	155, 255	281, 381
Exp. 4	*bath-path*	0, 5, 10, 15, 20 25, 30, 35, 40	240, 260	336, 356
	beach-peach	0, 5, 10, 15, 20 25, 30, 35, 40	155, 175	354, 374
	beak-peak	0, 5, 10, 15, 20 25, 30, 35, 40	180, 200	349, 369
	best-pest	0, 5, 10, 15, 20 25, 30, 35, 40	190, 210	392, 412
	bet-pet	0, 5, 10, 15, 20 25, 30, 35, 40	165, 185	328, 348
	buck-puck	0, 5, 10, 15, 20 25, 30, 35, 40	250,270	414, 434

[a]"n/a" indicates stimuli that were not included in the analysis for Experiment 2 (see Note 5).

Table 2

Results of follow-up tests for mouse-click responses from Experiment 1.

	Difference (VOT steps)	t(23)	p
bath-path	0.292	2.65	0.014
beach-peach	1.215	6.37	<0.001
beak-peak	0.259	1.77	0.09
best-pest	0.118	0.93	0.363
bet-pet	0.582	2.90	0.008
bike-pike	0.430	3.65	0.001
buck-puck	0.554	2.61	0.016