HHS Public Access

Author manuscript

Lang Cogn Neurosci. Author manuscript; available in PMC 2016 January 01.

Published in final edited form as:

Lang Cogn Neurosci. 2015 ; 30(5): 529–543. doi:10.1080/23273798.2014.946427.

The time-course of speaking rate compensation: Effects of sentential rate and vowel length on voicing judgments

Joseph C. Toscano and

Beckman Institute for Advanced Science and Technology, University of Illinois at Urbana-Champaign, 405 N Mathews Ave, Urbana, IL 61801

Bob McMurray

Dept. of Psychology and Dept. of Communication Sciences & Disorders, University of Iowa, E11 Seashore Hall, Iowa City, IA 52242

Abstract

Many sources of context information in speech (such as speaking rate) occur either before or after the phonetic cues they influence, yet there is little work examining the time-course of these effects. Here, we investigate how listeners compensate for preceding sentence rate and subsequent vowel length (a secondary cue that has been used as a proxy for speaking rate) when categorizing words varying in voice-onset time (VOT). Participants selected visual objects in a display while their eye-movements were recorded, allowing us to examine when each source of information had an effect on lexical processing. We found that the effect of VOT preceded that of vowel length, suggesting that each cue is used as it becomes available. In a second experiment, we found that, in contrast, the effect of preceding sentence rate occurred simultaneously with VOT, suggesting that listeners interpret VOT relative to preceding rate.

Keywords

speech perception; spoken word recognition; speaking rate; context effects; visual world paradigm

The time-course of speaking rate compensation: Effects of sentential rate and vowel length on voicing judgments

Despite enormous acoustic variability from factors like talker and speaking rate, listeners are remarkably adept at speech recognition. At a basic level, this process involves mapping continuous acoustic cues[1] onto categories (features, phonemes, or words). For example, stop voicing (/b,d,g/ vs. /p,t,k/) is distinguished by voice-onset time (VOT; the timing between closure release and the onset of voicing), with VOTs less than ≈20ms (in English) corresponding to voiced phonemes and VOTs longer than ≈20ms corresponding to voiceless phonemes (Lisker & Abramson, 1964). However, there is no one-to-one mapping between cues and categories (Repp, 1981; McMurray & Jongman, 2011). VOT varies as a

Corresponding Author (contact information as of August 2014): Joseph Toscano, Department of Psychology, Villanova University, 800 E Lancaster Ave, Villanova, PA 19085, joseph.toscano@villanova.edu.

[1]We define cues as measureable properties of the signal that serve as sources of information for perception.

function of speaking rate (Kessinger & Blumstein, 1997; Allen & Miller, 1999), talker (Allen, Miller, & DeSteno, 2003), place of articulation (Lisker & Abramson, 1964), and coarticulatory context (Nearey & Rochet, 1994). As a result, it must be interpreted with respect to context.

In describing how listeners cope with this variability, we can distinguish *phonetic information*, which directly signals a given phoneme, and *contextual information*, which modifies how phonetic cues are interpreted (Repp, 1982). Talker gender, for example, is not a cue for vowel identity (e.g., a male voice does not indicate what the vowel is) but may tell the listener to expect higher or lower formant frequencies. Similarly, speaking rate does not distinguish /b/ and /p/, but it indicates how to interpret VOT (since VOT varies with rate).

The process of integrating multiple *phonetic* cues is described by several models as an additive process, where each cue contributes some information to the percept (Oden & Massaro, 1978; Nearey, 1990; Toscano & McMurray, 2010; McMurray & Jongman, 2011). However, the mechanisms by which *context* cues are integrated with phonetic cues are less clear.

The goal of the present study is to understand one aspect of context, temporal asynchrony, the fact that context information does not always arrive simultaneously with phonetic cues. Speaking rate, for example, is inferred from sentential material surrounding the stop consonants it affects. As we describe, this creates opportunities to measure the perception of phonetic and context information independently and draw inferences about core processes in speech perception (McMurray, Clayards, Tanenhaus & Aslin, 2008b; Toscano & McMurray, 2012).

To do this, we used the visual world paradigm (Tanenhaus, Spivey-Knowlton, Eberhardt & Sedivy, 1995), which provides a real-time measure of online language processing. In these tasks, participants view pictures in a computer display (or physical objects on a table) and hear spoken instructions to select an item while their eye movements are recorded. Because they direct their gaze to an object before moving the cursor (or directing a reaching motion) to it, and because they make multiple eye movements per second, we can infer which referents the participant is considering *during* language processing (rather than simply at the *end* of processing, as with overt behavioral responses). This technique has been used to investigate sentence comprehension (Tanenhaus et al., 1995; Altmann & Kamide, 1998), lexical processing (Allopenna, Magnuson, & Tanenhaus, 1998), prosody (Salverda, Dahan, & McQueen, 2003), and, as we build on here, cue integration during speech perception (McMurray et al., 2008b; Toscano & McMurray, 2012). This allows us to investigate *when* specific phonetic or context cues affect higher level processing as the input unfolds, and it may distinguish approaches to cue integration much in the same way that the psychological refractory period has been used in other domains (Pashler, 1994).

For individual spoken words, the visual world paradigm is thought to measure lexical activation (Allopenna et al. 1998; Dahan, Magnuson & Tanenhaus, 2001), a process that occurs downstream from phonetic cue-integration. Its good temporal fidelity allows us to time-lock changes in the signal to effects on lexical activation (McMurray, Tanenhaus, &

Aslin, 2009; Salverda et al., 2003; Reinisch & Sjerps, 2013). If listeners wait to make lexical commitments until all information (both phonetic and contextual) is available, this would imply an independent pre-lexical processing stage. In contrast, if lexical activation is immediately sensitive to cues as they arrive, this would suggest a continuous cascade from perceptual to lexical processes. While there is evidence for continuous cascades in the integration of phonetic cues from eye-tracking (McMurray et al., 2008b; Toscano & McMurray, 2012), ERP (Van Petten et al., 1999), and gating studies (Warren & Marslen-Wilson, 1987; Marslen-Wilson & Zwisterlood, 1989; Gaskell & Marslen-Wilson, 1996), the story is less clear for contextual cues (Reinisch & Sjerps, 2013).

Here, we use the visual world paradigm to understand the role of context in speech perception using two cues for speaking rate that contribute to voicing perception. We first give an overview of phonetic and perceptual data on speaking rate and describe what is known about the time-course of processing. We then present two experiments examining the effects of rate and VOT on the time-course of spoken word recognition.

Speaking rate effects

The present study focuses on word-initial stop voicing, though rate compensation has also been examined in parallel situations involving durational cues, including manner of articulation (Miller & Lieberman, 1979), vowel quantity (Pind, 1995; Reinisch & Sjerps, 2013), fricative voicing (Denes, 1955), and word-medial voicing (Port & Dalby, 1982). Thus, while we focus on word-initial voicing, we also address broader issues of speaking rate compensation.

In syllable-initial stops, VOT is a reliable cue for voicing. Like many temporal cues, it is affected by speaking rate, such that at faster rates, VOT values tend to be closer to zero (Allen & Miller, 1999; Kessinger & Blumstein, 1997; Beckman, Helgason, McMurray, & Ringen, 2011). In English and aspirating languages, speaking rate primarily affects the voiceless category (/p,t,k/), with longer, more variable VOTs at slower rates, and does not exert much effect on the voiced category (/b,d,g/). In languages that use pre-voicing, the pre-voiced category changes with speaking rate (Kessinger & Blumstein, 1997; Beckman et al., 2011; Magloire & Green, 1999). These effects can cause VOT to be ambiguous by itself: a 25 ms VOT may indicate a voiceless sound in fast speech but a voiced sound in slow speech.

Listeners can estimate speaking rate from multiple sources of information in the signal, including preceding sentential rate (SR) and subsequent vowel length (VL)[2], and there is evidence that listeners use both cues (VL: Summerfield, 1981; Miller & Dexter, 1988; Pind, 1995; Boucher, 2002; Miller & Volaitis, 1989; McMurray et al., 2008b; Toscano & McMurray, 2012; for manner, see Miller & Lieberman, 1979; Miller & Wayland, 1993; SR: Summerfield, 1981; Miller & Grosjean, 1981; Wayland, Miller & Volaitis, 1994). Listeners identify more sounds as voiced at slow rates, and as voiceless at fast rates, although they do

[2]While the effect of VL on voicing *perception* has been amply demonstrated, there is ongoing debate about the relationship between VOT and vowel length (VL) in *production* (Kessinger & Blumstein, 1998; Allen & Miller, 1999), which may partially derive from differences in the definition of VL (Turk, Nakai, & Sugahara, 2006).

not fully adjust to the optimal boundary obtained from phonetic measurements (Miller, Green & Reeves, 1986).

Previous studies have also compared effects of VL and SR. Summerfield (1981) found that VL has larger effects than SR, leading him to conclude that information in the same syllable as the consonant may be more important (even though VL is a less direct measure of speaking rate). Miller, Volaitis and colleagues used goodness ratings to assess the category structure of voiceless sounds in response to changes in either VL (Miller & Volaitis, 1989) or SR (Wayland, Miller, & Volaitis, 1989). They found that, while SR effects were limited to the boundary and best exemplar, VL affected the entire range, suggesting somewhat different mechanisms by which each cue is used.

There is also debate about whether rate plays a role in naturalistic stimuli. Studies examining manner of articulation found no effect of VL when the stimuli reflected natural covariation between cues (Shinn, Blumstein, & Jongman, 1985; but see Miller and Wayland, 1993) and no effect of VL on voicing judgments in natural speech (Utman, 1998). Later work found an effect in stimuli constructed from natural speech (Boucher, 2002; Toscano & McMurray, 2012), even with VL differences that are as small as the mean difference measured in phonetic data (Toscano & McMurray, 2012, Experiment 4), but the effects for naturally-produced speech are smaller than those for synthetic speech (Toscano & McMurray, 2012, Experiment 1). Previous studies have not examined SR effects in naturalistic speech.

Limitations of existing models: The time-course of processing

Several models describe how listeners could compensate for these rate differences. Compound-cue models (Denes, 1955; Port & Dalby, 1982; Kessinger & Blumstein, 1998; Boucher, 2002) argue that listeners compute the ratio between VOT and VL, though it is unclear how these models would handle SR differences and they may be complicated by the finding that VL does not trade-off in a one-to-one relationship with VOT. Durational contrast models (Pisoni, Carrell, & Gans, 1983; Diehl & Kluender, 1989; Lotto & Kluender, 1998; Kluender, 2003; Holt, 2005) use contrasts between temporal events like VOT and VL or durations of syllables in other parts of the sentence. Cue-integration models (Nearey, 1997; Nearey, 1990; Nearey, 1986; Oden & Massaro, 1978; Massaro & Cohen, 1983; Toscano & McMurray, 2010) and exemplar models (Johnson, 1997; Goldinger, 1998) treat contextual information as additional phonetic cues. While they have not been explicitly applied to SR, these models could handle it similarly to talker effects, with SR serving as a direct cue to words or phonemes. Finally, explicit-compensation models (McMurray and Jongman, 2011; Cole, Linebaugh, Munson, & McMurray, 2010; Smits, 2001a; Smits 2001b) propose that listeners factor out predictable information (such as rate differences) when computing the values of phonetic cues.

However, none of these models address the time-course of processing. Recently, McMurray et al. (2008b) raised the issue of temporal asynchrony in rate compensation. They argue that since VL arrives *after* VOT, there are multiple mechanisms by which listeners could process information as it is received. Listeners could store VOT in a buffer and wait until VL arrives before making phonetic or lexical decisions. Alternatively, they could partially activate lexical candidates continuously when VOT is heard and update this activation when VL

arrives. McMurray et al. used the visual world paradigm to determine *when* (during real-time processing) the effects of each cue were observed. A buffered model predicts simultaneous late effects (i.e., listeners wait for both cues before accessing the lexicon) and a continuous cascade approach predicts an early effect of VOT followed by an effect of VL. For both voicing and manner, the evidence supported a continuous cascade—listeners showed an effect of VOT shortly after word onset and an effect of VL several hundred milliseconds later.

Toscano and McMurray (2012) followed up on this using stimuli constructed from natural speech, where the existence of VL effects has been debated. They found a smaller effect of VL, and confirmed that even with stimuli produced from natural speech, the eye-movement data support a continuous cascade, with VOT and VL used as each becomes available. Further, this study clarified why the effect of VL is smaller in natural speech: tertiary cues also play a role, diluting the apparent effect of VL. Given this, VL effects can be explained in a model with additive contributions from each cue that are used immediately. This observation, along with the fact that VOT can be used without VL, suggests that VL acts as a weak phonetic cue, rather than as a context effect. However, it is unclear whether SR uses the same continuous cascade process.

Other work has used the visual world paradigm to investigate the time-course of processing for contextual cues to vowel identity. Reinisch and Sjerps (2013) presented listeners with sentences ending in /ɑ/-/a:/ minimal pairs that varied in F2 and duration (both distinguish Dutch vowels). These cues were also varied in the preceding sentence to create different contexts. They examined the time-course of the effect of each cue relative to its corresponding context information to determine whether cues and context were used immediately. The effect of context occurred at the same time as that of the phonetic cues,[3] suggesting that sentential context may not be used immediately (or independently of the cues), but may be stored to modulate the use of phonetic cues. This would appear to conflict with Toscano and McMurray's (2012) claims that rate compensation may be handled by general cue-integration mechanisms, since those models predict that rate information should have been used as soon as it was available (during the sentence context), though Toscano and McMurray did not investigate SR effects.

Reinisch and Sjerps raise the possibility that context effects are handled by low-level auditory mechanisms such that context modulates the way similar acoustic cues are interpreted (Diehl & Kluender, 1989; Lotto & Kluender, 1998; Holt, 2005). However, because this study involved a phonological distinction that was directly related to duration (phonemic vowel length), it is unclear whether these findings extend to more abstract phonological contrasts (e.g., voicing differences, which relate to the presence of the spread-glottis articulatory feature). Such effects could pose a challenge for auditory accounts, since they involve comparisons between events that are less acoustically similar. Reinisch and Sjerps also used two-alternative forced-choice tasks with orthographic labels, which may be

[3]There appeared to be differences in the time-course of the spectral cues in the context vs. those in the target word, but this difference was not statistically significant. We focus on the effects of the durational differences, since those are most directly applicable to the current investigation.

less sensitive to fine-grained phonetic detail than the more typical four-alternative tasks using pictures (McMurray, Aslin, Tanenhaus, Spivey, & Subik, 2008a). This also raises questions about whether this result is specific to orthographic-phonological mappings or to lexical activation.

The purpose of the present study is to evaluate the effect of speaking rate on the time-course of listeners' use of voicing cues. Doing so will allow us to clarify issues in the literature for several reasons. First, while the effect of SR is unknown, these timing issues have been worked out for VOT and VL, providing a useful basis for studying effects of SR. Second, our task may be more sensitive to subtle effects than overt phoneme decisions because it provides a measure of the time-course of spoken language processing (as observed in other visual world experiments; e.g., McMurray et al., 2008a) and offers a clearer measure of lexical processes. Third, in addition to SR there are two durational cues (VOT and VL) that are relevant, allowing us to ask if SR is combined with phonetic information in a continuous cascade or whether it is buffered until certain types of cues are available (e.g., SR could be buffered until VOT arrives or even later to modulate the use of VL, as observed in Reinisch and Sjerps, 2013). Finally, investigating the issues regarding context information raised by Reinisch and Sjerps (2013) in a new domain (voicing) can illuminate the generality of these effects.

Predictions

The goal of the current study is primarily empirical: to measure the time-course of cue combination relative to the activation of lexical candidates. Few models of rate compensation make clear predictions about the time-course of these effects. Thus, these results may winnow the space of possible models and constrain models as they are developed to describe the time-course of processing.

A number empirical patterns are possible. Given prior results, the effect of VL should follow that of VOT (McMurray et al., 2008b; Toscano & McMurray, 2012). However, this has not been evaluated in sentential context, and it is possible that with a strong estimate of surrounding speaking rate, asynchronous VOT and VL effects may not be observed. Listeners may weight VL more since it can be estimated more accurately relative to the surrounding rate, which could change the time-course of the effects. Thus, our first experiment examined effects of VOT and VL adding a sentence context at a constant rate.

Assuming VL effects occur after VOT effects in sentential context, there are several possible relationships between these cues and SR. First, SR effects could precede VOT effects. If this is the case, we would predict a difference in fixations to /b/- and /p/-items as a function of SR (i.e., more /p/ fixations at fast SRs; more /b/ fixations at slow SRs) early in the time-course, and differences as a function of VOT at a later time. This would suggest that SR serves as a bias: fast speech biases the system toward voiceless items such that shorter VOTs can activate them. This is the prediction of cue-integration and exemplar models. In this case, SR serves (like VL) as another cue to voicing. This is similar to how normalized *a posteriori* probability models use bias terms to handle coarticulation (Nearey, 1990), how models of word frequency use pre-activation to simulate higher likelihoods (McClelland and Rumelhart, 1981; Marslen-Wilson, 1990; though see Dahan, Magnuson, &

Tanenhaus, 2001), how exemplar models handle talker specificity (Goldinger, 1998), and how Bayesian approaches might implement SR as a prior probability on voicing judgments.

Second, it is possible that SR effects will occur after VOT and coincidently with VL. That is, differences in fixations as a function of SR and VL could be observed at the same point in time (subsequent to VOT-related differences). This would suggest that the system buffers SR until VL is available (rather than using it to adjust how VOT is used). Given that vowel durations are likely to be salient differences in SR, this could favor an auditory account in which similar cues interact (Lotto & Kluender, 1998; Holt, 2005). The direction of the effect could reveal whether they interact contrastively (e.g., a short VL at a slow SR seems *even shorter)* or additively (a short VL and slow SR cancel out).

Finally, SR could have an effect at the same time as VOT and prior to VL. This would be consistent with approaches that treat SR as a true context effect that is used as soon as cues that must undergo compensation (here, VOT) are available, regardless of their acoustic similarity. It would also provide a counterpoint to evidence from VL that rate compensation can be handled using only cue-integration approaches (although, as we discuss, it doesn't rule out certain forms). Instead, an effect of SR that is coincident with VOT would support explicit-compensation approaches where SR modulates VOT encoding or is used to alter the VOT boundary.

Experiment 1: Effects of VOT and vowel length

Before examining the effect of sentential rate, we must confirm that previous findings hold for stimuli in sentential context. This experiment serves three purposes: (1) it provides a conceptual replication of Toscano and McMurray (2012); (2) it establishes that this paradigm can be used with stimuli in sentential context; and (3) most importantly, it allows us to evaluate whether listeners use VL when they have a robust estimate of rate from the preceding sentence.

Listeners performed a picture identification task while their eye-movements were recorded. An effect of VOT that occurs before VL would be consistent with previous work. Alternatively, listeners may ignore VL if they establish speaking rate based on information from the sentence. Such a result would fit with previous work arguing that listeners do not use VL in natural speech (Shinn, Blumstein, & Jongman, 1985; Utman, 1998).

Method

Design—Participants performed a 4AFC picture identification task. Auditory stimuli consisted of six sets of /b/-/p/ minimal-pair words *(bath-path, beach-peach, beak-peak, bet-pet, bike-pike, buck-puck*) preceded by one of five carrier sentences (Table 1). Referents varied along nine-step VOT continua and two VL conditions, with the carrier sentence spoken at a constant rate. Each minimal pair was grouped with two unrelated words starting with either /l/ or /ʃ/ (which were not minimal pairs; *lace, lap, leash, light, loaf, lock, chef, shake, sheep, sheet, ship, shop*). For each participant, /l/ and /ʃ/ items were pseudo-randomly assigned to specific /b/-/p/ words with the requirement that semantically related words (e.g., *pet* and *leash*) could not be paired together. Auditory stimuli for /l/ and /ʃ/ items also varied

in VL. This was done so that stimulus pairings were consistent across trials, and there was no reason to associate specific /b/ and /p/ stimuli.

Stimuli were presented in random order. Each combination of continuum (6), VOT (9), VL (2), experimental status (experimental vs. filler; 2), and carrier phrase (5) was repeated once, for a total of 1080 trials. The experiment was conducted over the course of two days and lasted ≈60 minutes each day.

Participants—Fifteen people (seven female) participated in the experiment. Participants were recruited from the University of Iowa (primarily undergraduate students) according to University human subjects protocols, provided informed consent, and received monetary compensation or course credit. Participants reported English as their only native language, normal hearing, and normal or corrected-to-normal vision.

Stimuli—Stimuli were recorded by a male talker (author B.M.) in a sound-attenuated room using a Marantz PMD670. Recordings were made at 22.05 kHz. Several tokens of each word were recorded, and those with the best audio quality were selected. Recordings were edited using Praat (Boersma & Weenink, 2010).

VOT continua were created from these recordings by cross-splicing the voiced and voiceless tokens (McMurray et al., 2008a) with the original voiced token serving as the /b/ endpoint. Each VOT step was generated by removing a period from the onset of the voiced token equal to the appropriate temporal duration for that VOT (e.g., a 20-ms portion was removed for the 20-ms VOT step) and splicing the corresponding amount of aspiration from the voiceless token onto the remaining vocoid from the voiced token. Each token was marked at zero-crossings in approximately 5-ms steps from the onset of the word to 40 ms after onset, and nine-step VOT continua were created (the endpoint stimuli at 0 and 40 ms were created in the same way as all other stimuli, except that the 0-ms step was identical to the voiced token, since the VOT is 0 ms).[4] Onsets of unrelated stimuli were unmodified.

The two VL conditions were created using the pitch-synchronous overlap-add method in Praat (minimum pitch: 75 Hz; maximum pitch: 600 Hz). The onset and offset of the vowel (measured from the release burst to the offset of voicing) was marked for each sound, and VLs were increased or decreased by 40% of their original duration (Toscano & McMurray, 2012). Mean vowel duration was 189 ms for the short VL condition and 377 ms for the long VL condition.[5] Mean duration of the entire target word was 359 ms for the short VL condition and 547 ms for the long VL condition (see Toscano & McMurray, 2012, for durations of each minimal pair). Vowel durations of unrelated stimuli were modified by the same proportions to create long and short VL conditions. Finally, carrier phrases were spliced onto the target and unrelated words, and sounds were normalized for intensity.

[4]After the experiments were run, we discovered a splicing error causing the VOT of one token to be inaccurate (specifically, step 4 along the *beach-peach* continuum). Removing this token from the analyses produced the same pattern of results.

[5]Note that, although these VL differences are larger than the mean voicing-related VL difference seen in phonetic data (Allen & Miller, 1999), the effect of VL is still observed with much smaller (20 ms) VL differences (see Toscano & McMurray, 2012, Experiment 4).

Visual stimuli were clipart images selected and processed using standard lab procedures used in several previous studies (Apfelbaum, Blumstein, & McMurray, 2011; Toscano & McMurray, 2012). For each word, several pictures were downloaded from a clipart database. A small focus group of students (close in age to our primarily undergraduate participant pool) selected the picture that depicted the most canonical representation of the word and recommended any changes to make it more canonical and remove distracting elements. After these edits, the final images were approved by a lab member with extensive experience using the visual world paradigm. The arrangement of pictures in the display was randomized such that, for a given VOT, VL, and carrier phrase, each relative arrangement of /b/-/p/ items (adjacent horizontally, vertically, or diagonal) occurred equally often.

Procedure—Participants were seated in a sound-attenuated room and wore an SR Research Eyelink II head-mounted eye-tracker and Sennheiser headphones. Auditory stimuli were presented at the participant's most comfortable level. The eye-tracker was calibrated using the standard 9-point calibration, and, participants then began the experiment. First, two sets of training trials were presented to familiarize participants with the pictures and words. In the first part of training, participants saw each picture in the center of the screen. The written word corresponding to the picture appeared below the image 500 ms later. After viewing the picture and reading the name, participants clicked the mouse button to continue to the next trial. Each picture was presented once in random order. In the second part of training, four pictures (one set of /b/, /p/, /l/, and /ʃ/ items) appeared in the four corners of the display (with each object randomly assigned to a corner), and the written word corresponding to one of the pictures appeared in the center of the display 500 ms later. Participants clicked on the picture corresponding to the written word to go onto the next trial (clicking the correct picture was required to continue). Each word was presented twice in random order.

After training, the experimental trials began. Each trial proceeded similarly to the second part of training. At the beginning of the trial, one picture appeared in each corner of the display, and a blue circle appeared in the center. After 500 ms, the circle turned red, participants clicked on it, and the auditory stimulus was played 100 ms later. Participants then made their response by clicking on the picture corresponding to the instruction they heard. Participants were given the opportunity to take a break every 45 trials, and a drift correction to account for movement of the eye-tracker was performed after each break.

Data processing—Eye-movements were recorded and automatically parsed into saccades and fixations by the Eyelink software. Each saccade was paired with the subsequent fixation to create a “look”, the onset of which reflects the earliest moment a participant could be directing an eye-movement to an object. The proportion of looks to each object was computed in 4-ms steps starting at the onset of the trial. During analysis, the boundaries surrounding each object were extended by 100 pixels to account for noise in the eye-track.

Results

Mouse-click responses—Listeners were highly accurate at identifying the endpoints of the VOT continua (99% correct). Figure 1 shows the proportion of /p/ responses as a

function of VOT and VL. Trials in which the participant selected a filler picture were excluded from analysis. There were more voiced responses for long VLs and more voiceless responses for short VLs, consistent with previous results showing a shift in listeners' category boundaries as a function of VL. Mean RT for the experimental trials (relative to target onset) was 1184 ms (SD: 507 ms).

These data were validated statistically using a logistic mixed-effects model with VOT, VL, and their interaction as fixed effects (centered), and by-subject and by-item random slopes for VOT, VL, and their interaction. The model showed main effects of VOT (b=2.26, SE=0.27, z=8.47, p<0.001) and VL (b=0.62, SE=0.24, z=2.59, p<0.01), confirming that both cues had an effect. The interaction was not significant.

Eye-movements—Looks to /b/-/p/ objects varied with both VOT and VL. Participants were more likely to fixate the /p/ object at longer VOTs and the shorter VL, mirroring the results observed in the mouse-click responses. The time-course of each effect was examined using an approach similar to McMurray et al. (2008b; Toscano & McMurray, 2012). For each subject, we computed the difference between the proportion of fixations to the /b/ and /p/ objects (*b/p-bias*) at each level of each cue. Differences in *b/p-bias* over time are shown in Fig. 2. Next, the effect of each cue was determined from the *b/p-bias*. For VOT, we computed linear regressions between the magnitude of *b/p-bias* and VOT step, and the slope was used as a measure of the size of the VOT effect. For VL, the effect size was simply the difference in *b/p-bias* between the two conditions (since VL had only two levels).

The results of this procedure are shown in Fig. 3. In this and subsequent figures, 0 ms corresponds to the onset of the target word. This was not adjusted for the known 200 ms occulomotor-planning delay, so 200 ms is the first point at which we would expect signal-driven eye-movements. The VOT effect begins shortly after 250 ms, and the VL effect begins shortly after 750 ms. The time-course functions for individual subjects are noisy, making it difficult to estimate accurate time parameters at an individual level. Thus, we used the jackknife procedure for statistical analyses (Miller, Patterson, & Ulrich, 1998; see McMurray et al., 2008b; Apfelbaum et al., 2011; Toscano & McMurray, 2012; Reinisch & Sjerps, 2013, for application to the visual world paradigm). To jackknife the data, we first computed the average effect size as a function of time for each cue with one subject excluded. We then fit a four-parameter logistic function to the jackknifed time-course data by minimizing the least-squares error between the function and the jackknifed data, as in Apfelbaum at al. (2011). The midpoint of this logistic was then used as a measure of the time of that effect. This procedure was repeated, excluding one subject at a time, yielding a dataset that had the same number of subjects as the original, non-jackknifed set. A paired t-test between the midpoints for the VOT and VL time-courses was performed, adjusting the error term to reflect the fact that each data point corresponds to N-1 subjects because of jackknifing (Miller et al., 1998).

We found that the average midpoint for VOT occurred at 547 ms, and the average midpoint for VL occurred at 910 ms. This difference was statistically significant ($t_{jackknifed}(14)=3.02$, p=0.009). These effects occurred well before the mean RT of the mouse-click responses

(1184 ms). Thus, the effect of VOT occurs earlier than the effect of VL, even when the stimuli are embedded in a sentential context.

Discussion

The mouse-click responses replicate Toscano and McMurray (2012), which showed that effects of VL can be observed in stimuli constructed from naturally-produced speech. Here, we see that this is also observed when stimuli are presented in a more natural sentential context.

The eye-tracking data extend Toscano and McMurray (2012) and McMurray et al. (2008b) to show that temporally-asynchronous effects of VOT and VL can be observed in a sentential context that offers significant information about speaking rate. These results are most consistent with models in which listeners treat VL as a phonetic cue rather than contextual information that is used to modify the perceived VOT. That is, listeners do not wait for VL before considering lexical candidates on the basis of VOT. Rather, the two cues exert independent effects on lexical activation. This rules out a model in which context (VL) is required to interpret VOT (which would have predicted that listeners wait for VL to use VOT) and suggests that a simpler cue-integration model can explain the results.[6] Given these results, we now ask how listeners use *preceding* SR.

Experiment 2: Effects of sentential rate

This experiment evaluated the effects of preceding SR on listeners' voicing judgments. The design was similar to Experiment 1, except that a third factor (SR) was added by manipulating the durations of the carrier phrases. Determining whether or not SR has an effect is important, since several studies have suggested that in natural speech listeners either do not use rate information (Shinn et al., 1985; Miller & Wayland, 1993; Utman, 1998) or show smaller effects (Toscano and McMurray, 2012; Boucher, 2002), but this has not been examined for SR. Most importantly, if SR has an effect, we can examine its time-course to determine whether it is processed independently of VOT (like VL) or whether it is used in conjunction with VOT or VL.

Method

Design—As in Experiment 1, the design included within-subject manipulations of VL and VOT. In addition, SR was manipulated within-subject by varying the rate of the carrier phrases in two levels (fast and slow). Because this doubled the number of conditions in the experiment, we did not use each variant of the five carrier phrases with each possible combination of experimental conditions. Rather, each combination of VOT, VL, SR, and continuum was repeated three times with a randomly chosen carrier phrase each time. Thus, the total number of trials in the experiment was 1296 ([VOT × SR × VL × continuum × repetition] + fillers). The experiment was conducted over two days (≈90 minutes per day).

[6]An alternative possibility is that VL modulates the use of VOT, but that VOT can be interpreted without it. Such a model cannot be ruled out. A simpler model, however, is that VL serves as an additional, secondary cue to voicing rather than as a true context effect (Toscano & McMurray, 2012).

Participants—Twenty participants completed the experiment. Participants met the same requirements as Experiment 1, provided informed consent, and received monetary compensation or course credit. One participant was excluded for having less than 80% correct on the continua endpoints and another was excluded because of a poor quality eye-track, leaving 18 participants (10 female) in the analysis.

Stimuli—The same recordings from Experiment 1 were used, except that the durations of the carrier phrases were modified to create fast and slow SR conditions. The durations of vowels and sentence contexts were modified via the same method used to create the VL conditions in Experiment 1. Carrier sentences were increased and decreased by 15% of their original duration to create the slow and fast SR conditions. These differences in sentence length produced speaking rates in a range similar to those reported by Miller, Grosjean, and Lomanto (1984). Carrier phrases were then spliced onto each referent. Visual stimuli were the same as Experiment 1. Randomization of picture locations followed a similar procedure, with each relative arrangement of minimal pairs occurring equally often for a given VOT, VL, and SR.

Procedure and data processing—The experiment and data processing procedures were the same as in Experiment 1, except that breaks and drift corrections occurred every 54 trials.

Results

Mouse-click responses—Listeners correctly identified the endpoints of the VOT continua (mean accuracy: 98%). Figure 4 shows the proportion of /p/ responses as a function of VOT, SR, and VL.[7] There were more voiced responses in the slow SR condition than in the fast SR condition, consistent with the prediction that preceding rate influences voicing judgments. The mean RT for the experimental trials (relative to target word onset) was 1278 ms (SD: 695 ms).

Responses were analyzed using a logistic mixed-effects model with VOT, VL, and SR as fixed effects; by-subject random slopes for main effects, the VOT×VL interaction, and the VOT×SR interaction; and by-item random slopes for main effects, and the VOT×VL interaction.[8] We found a main effect of VOT ($b=1.87$, $SE=0.24$, $z=7.85$, $p<0.001$) with more voiceless responses for long VOTs. There were also main effects of VL ($b=0.82$, $SE=0.17$, $z=4.90$, $p<0.001$) and SR ($b=0.41$, $SE=0.10$, $z=4.14$, $p<0.001$), such that listeners made more voiceless responses in the context of short vowels and fast sentences than in the context of long vowels and slow sentences. None of the interactions were significant. Thus, SR also has an effect on voicing in stimuli constructed from natural speech.

Eye-movements—As in Experiment 1, looks to each object as a function of VOT and VL reflected the mouse-click responses. Similar results were observed for SR: Listeners were more likely to fixate /p/ objects at longer VOTs, short VLs, and short (faster) SRs (Fig. 5).

[7]As in Experiment 1, trials in which participants clicked on a filler object were excluded from analysis.
[8]This was the most complex model that successfully converged.

The time-course of each effect was analyzed in the same way as Experiment 1. The effect of SR was found by calculating the difference in *b/p-bias* between the two SR conditions (Fig. 6). Planned paired t-tests (adjusting for jackknifing) between VOT and VL and between VOT and SR were performed. The average midpoint for VOT (586 ms) occurred significantly earlier than the average midpoint for VL (949 ms; $t_{jackknifed}(17)=3.83$, $p=0.001$), consistent with the results of Experiment 1. The average midpoint for SR (613 ms) was not significantly different from that of VOT ($t_{jackknifed}(17)=0.244$, $p=0.810$). Thus, although SR precedes VOT in the signal, listeners do not use it for voicing judgments until they hear the VOT.[9] All effects occurred well before listeners' mouse-click responses (mean RT: 1278 ms).

Discussion

These results show that SR has an effect on voicing in stimuli constructed from natural speech. They also suggest that listeners consider VOT relative to preceding SR and that this context compensation process occurs pre-lexically, since there is no evidence of direct lexical activation from SR that occurs prior to and independently of lexical activation from VOT. In conjunction with the results showing that VOT effects precede those of VL, this indicates that listeners use a hybrid strategy, combining aspects of cue-integration and explicit-compensation. Listeners modulate their use of VOT on the basis of preceding SR, but also use VOT without waiting to hear the length of the subsequent vowel. Thus, SR appears to adjust listeners' use of VOT, whereas VL directly influences voicing. In combination with the results of Reinisch and Sjerps (2013), this suggests that, while cue-integration need not be completed prior to lexical activation, it appears that it is for contextual factors like SR.

General discussion

These results provide evidence that listeners use context information in speech flexibly: they use preceding context (SR) when it is available (Experiment 2), but they recognize speech even in the absence of this information (Experiment 1).[10] That is, listeners compensate for contextual variability, but such compensation is not obligatory. In Experiment 1, we show that listeners use VOT and VL asynchronously, demonstrating that each cue is used as it becomes available; listeners do not have to wait for VL to use VOT. In Experiment 2, we found that preceding SR is processed *simultaneously* with VOT, while VL is still processed asynchronously. There was no evidence that SR is used independently of VOT (in contrast to predictions from exemplar or additive cue-integration models), and raw VOT does not appear to be used before SR modulates it.[11] Listeners use SR to adjust VOT as soon as that information is available, even while simultaneously cascading partial decisions to the lexicon.

---

[9] Although it appears there may be a small SR effect during the interval preceding the onset of the target word, this effect is not significantly different from zero ($t_{jackknifed}(17)=0.75$, $p=0.23$, one-tailed t-test).

[10] Because our stimuli were manipulated from natural speech (as opposed to unmanipulated natural productions), it is possible that a different pattern of results would be obtained with unmanipulated recordings. Nonetheless, our method of generating the stimuli preserves the fidelity and acoustic complexity of natural speech (unlike completely computer-generated synthetic speech). Thus, there is no obvious reason why these effects would not also be observed with completely unmanipulated speech.

[11] This would have appeared as an effect of VOT prior to the effect of SR.

This suggests that the speech system treats SR and VL differently (see also the contrast across Miller & Volaitis, 1989, and Wayland, Miller, & Volaitis, 1989)—VL is processed as a direct cue as soon as it arrives; SR, in contrast, modulates how other cues are used. One reason for this difference may simply be due to the time-course of processing and the order in which information arrives. Because SR is available when VOT arrives, it can modulate the way VOT is encoded or adjust the category boundary.[12] This suggests that SR effects are pre-lexical (since there are no effects of SR on lexical activation prior to the effect of VOT). In contrast, the system does not wait for VL before making initial lexical commitments.

Our results with VL are consistent with those of Toscano and McMurray (2012), and the overall findings corroborate those of Reinisch and Sjerps (2013). Here, we extend those results by showing how listeners use preceding contextual cues when making more abstract phonological distinctions (voicing decisions). This helps inform models of speech perception and context compensation more generally, as discussed below.

Models of context compensation

It is not clear that existing models can fully account for the complex pattern described above. In part, this is because models have not considered the time-course of processing. The present results may nonetheless constrain the selection of models.

Compound-cue models, for example, predict that listeners compute cues relative to information within the segment/syllable (Summerfield, 1981; Miller & Lieberman, 1979; Syrdal & Gopal, 1986; Christovich & Lublinskaya, 1979). For voicing, this suggests that listeners compute the ratio between VOT and VL (Boucher, 2002; Pind, 1995; see also Port and Dalby, 1982). However, the observation of asynchronous VOT and VL effects argues against this type of model. Perhaps a compound cue could be developed to evaluate VOT relative to SR, but there is not a clear motivation for such a combination of contextual and phonetic cues. Thus, without further development, these models do not appear sufficient to account for the data.

Durational contrast models have also been proposed for handling rate variability. This approach emphasizes general auditory principles with which listeners encode sounds relative to context in a contrastive way (Pisoni, Carrell, & Gans, 1983; Diehl & Kluender, 1989; Lotto & Kluender, 1998; Kluender, 2003; Holt, 2005). Durational contrast posits a pre-lexical stage that could be used for rate normalization (Reinisch & Sjerps, 2013). This is potentially consistent with our findings: VOT could be perceived in terms of its contrast with preceding durations. However, it is unclear whether the present results support a pure auditory approach, since there is a more abstract relationship between speaking rate and voicing (unlike the more direct relationship between rate and phonemic vowel length investigated by Reinisch and Sjerps). The specific predictions are unclear, particularly when

[12]The current data cannot distinguish these two explanations, as they are only measuring lexical activation and not lower-level perceptual encoding processes. At minimum, however, we can say that the effect of SR occurs pre-lexically. That is, listeners do not activate lexical candidates using VOT information independently of SR. In order to further identify the locus of the SR effect on VOT (i.e., to determine whether SR serves to modify the way that VOT is encoded or whether it serves to adjust a pre-lexical phonological category boundary), further data are needed. (See discussion below about future ERP studies that could address this point.)

multiple cues are involved. Why would preceding SR modulate VOT rather than the more acoustically similar VL? Do more reliable cues (VOT) take precedence over more acoustically similar ones (VL)? If so, is this a purely auditory account?

While the preceding approaches suggest that listeners use cue-values computed relative to context, cue-integration and exemplar models suggest that context is treated similarly to other sources of information (i.e., phonetic cues). For example, Toscano and McMurray (2010; 2012) suggest that long VLs could be mapped onto /b/ and short VLs onto /p/, just as short VOTs are mapped to /b/ and long VOTs are mapped to /p/. VL by itself carries some information about voicing, which is supported by phonetic measurements (Allen & Miller, 1999; McMurray, Kovack-Lesh, Goodwin, & McEchron, 2013). However, it does not appear that SR works the same way. Simple cue-integration models would predict an effect of SR followed by an effect of VOT (i.e., effects that occur in the same temporal order as the information in the signal), but this was not observed. Thus, while cue integration principles offer a solution to part of the problem, they may not be sufficient for explaining effects of SR.

Finally, there are models that combine cue-integration with an explicit form of compensation. This includes the hierarchical categorization (HICAT; Smits, 2001a,b) and computing cues relative to expectations (C-CuRE; McMurray and Jongman, 2011; Cole, Linebaugh, Munson, & McMurray, 2010) models. In these models, interpretation of phonetic cues is directly modified by context. While they have not been applied to speaking rate specifically, it is clear how they could be adapted: Listeners would compute an abstract representation of rate (perhaps by integrating many cues) and use this to alter how phonetic cues like VOT are treated with respect to voicing decisions. In HICAT, the rate estimate could be used to directly adjust the *VOT boundary*; in C-CURE, listeners could adjust expected *VOT values* as a form of predictive coding. Both models include cue-integration as a fundamental process and allow for an optional, additional context compensation process.

Overall, our results are most consistent with these types of models. The results also suggest that compensation for contextual variation can be explained by general processes. Compound-cue models rely on lawful relationships between phonetic cues, making the mechanisms highly specific to speech. Durational-contrast approaches, while more general, still rely on auditory principles to explain how listeners compensate for contextual differences. In contrast, cue-integration and explicit-compensation approaches use domain-general mechanisms that are not specific to speech or audition. For example, cue-integration models can be implemented using statistical learning mechanisms (Toscano & McMurray, 2010), and explicit-compensation models can be implemented using regression-based statistical models (McMurray & Jongman, 2011). Thus, general cognitive principles are sufficient to account for these results.

The results also suggest that SR and VL may not be used in the same way. SR affects lexical activation indirectly (by adjusting listeners' use of VOT), while VL appears to have a direct effect with its own time-course (though it is possible that VL effects occur by biasing VOT at a later time point when VL is available). This distinction between effects of SR and VL is similar to the distinction that Repp (1982) makes between cues and context effects: Factors

directly related to particular speech sounds are treated as cues, while those that are only indirectly related are treated as context. Listeners seem to process acoustic information related to context differently from information that may be directly related to phonological categories. This is consistent with Toscano and McMurray's (2012) suggestion that VL may be better thought of as a cue rather than a context effect. Given this, we must use caution when treating VL as a proxy for speaking rate. This also suggests that listeners are flexible, making immediate use of context when it is available, and otherwise, immediately using raw cues.

This result also seems reasonable for models of speech perception, since preceding SR does not, by itself, predict voicing categories (whereas VL does, albeit weakly; Allen & Miller, 1999). Unsupervised associative learning mechanisms, which have been proposed to describe speech sound acquisition in cue-integration models (McMurray et al., 2009; Toscano & McMurray, 2010), may therefore have a difficult time learning to map context information directly onto phonological categories, suggesting that some other mechanism may be needed (as in explicit-compensation models).

However, a crucial factor missing from all of these models is a tight integration with lexical activation. VOT and SR appear to be combined pre-lexically, but their output exerts only a partial (though immediate) constraint on lexical selection, which is sensitive to further cues (e.g., VL) as they arrive. This raises the possibility that some cue-integration occurs at a lexical level (that is, cues directly affect lexical activation), even though context compensation occurs pre-lexically (modulating the perception of the cues or how they are mapped to words). It is important to point out that even though compensation appears to be pre-lexical, it is not necessarily a variance-discarding process (Pisoni, 1997). Rather, models like C-CuRE stress that compensation can bias continuous perception of cue-values without discarding fine-grained detail and that it is likely to be sensitive to higher-level expectations (Apfelbaum, Bullock-Rest, Jongman, & McMurray, in press). At a broader level, it is clear that models of cue-integration and context compensation must consider a richer relationship with lexical processes.

Future directions

Although these results suggest a hybrid approach in which context compensation is possible but not necessary, they do not provide a specific mechanism by which compensation occurs. Compensation could be accomplished by re-encoding acoustic cues as values relative to context (McMurray & Jongman, 2011) or by adjusting the category boundary between two phonemes (Smits, 2001a).[13] Either process could precede the lexical activation we measured with the visual world paradigm.

Distinguishing these approaches requires methods that allow us to measure acoustic cue encoding more directly. Recently, Toscano, McMurray, Dennhardt, and Luck (2010) used the ERP technique to show that the auditory N1 varies linearly with VOT and the P3 component varies with VOT relative to listeners' category boundaries. Thus, the N1 gives us

[13]Nearey (1997) has also proposed that context can be accounted for by adjusting higher-level representations like diphones. While this seems like a plausible solution for compensating for coarticulation, it is not clear how this would be applied to speaking rate.

a measure of low-level encoding that can be used to assess listeners' representation of VOT. By examining whether the N1 changes as a function of preceding SR, we can determine whether SR affects encoding of VOT or whether it affects higher-level representations (e.g., diphones, phoneme boundaries). We are currently running experiments designed to address this question.

Finally, it is important to investigate these phenomena with other phonological contrasts and phonetic cues to determine whether these results generalize to the perceptual system more broadly. Indeed, work in progress on place of articulation in fricatives (/s/ vs. /ʃ/) suggests that these principles may not extend to this distinction, as information in the frication may be buffered until the onset of the vowel (Galle, 2014). Thus, some of these principles may break down in phonemes with a substantively different acoustic nature.

Conclusion

Understanding how listeners handle variability across contexts is critical for understanding speech perception. The results of these experiments provide us with a clearer picture of the processes that allow listeners to cope with variability in rate. Preceding SR is taken into account when processing temporal cues, like VOT, but does not exert an independent effect. However, compensation is not obligatory; listeners can use VOT independently of later-occurring information, like VL. This also suggests that VL may be better thought of as an independent phonetic cue to voicing. Moreover, while context integration likely occurs pre-lexically, the integration of multiple phonetic cues appears to cascade directly to the lexicon. Overall, these results suggest that listeners are flexible in how they handle speaking rate variability and suggest that models including principles of cue-integration, explicit compensation, and lexical activation dynamics are needed to provide a full account of context effects in speech.

Acknowledgments

We would like to thank Dan McEchron for assistance with data collection. This research was supported by a Beckman Postdoctoral Fellowship to J.C.T. and by NIH DC008089 to B.M.

References

Allen JS, Miller JL. Effects of syllable-initial voicing and speaking rate on the temporal characteristics of monosyllabic words. J Acoust Soc Am. 1999; 106:2031–2039.10.1121/1.427949 [PubMed: 10530026]

Allen JS, Miller JL, DeSteno D. Individual talker differences in voice-onset-time. J Acoust Soc Am. 2003; 113:544–552.10.1121/1.1528172 [PubMed: 12558290]

Allopenna PD, Magnuson JS, Tanenhaus MK. Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. J Memory Language. 1998; 38:419–439.10.1006/jmla.1997.2558

Altmann G, Kamide Y. Incremental interpretation at verbs: Restricting the domain of subsequent reference. Cognition. 1998; 73:247–264.10.1016/s0010-0277(99)00059-1 [PubMed: 10585516]

Apfelbaum KS, Blumstein SE, McMurray B. Semantic priming is affected by real-time phonological competition: Evidence for continuous cascading systems. Psychonomic Bull Rev. 2011; 18:141–149.10.3758/s13423-010-0039-8

Apfelbaum KS, Bullock-Rest N, Rhone A, Jongman A, McMurray B. Contingent categorization in speech perception. Language Cognitive Processes. in press. 10.1080/01690965.2013.824995

Beckman J, Helgason P, McMurray B, Ringen C. Rate effects on Swedish VOT: Evidence for phonological overspecification. J Phonetics. 2011; 39:39–49.10.1016/j.wocn.2010.11.001

Boersma, P.; Weenink, D. Praat: Doing phonetics by computer. 2010. Available from http://www.praat.org/ (Date last viewed: 11-February-2013)

Boucher VJ. Timing relations in speech and the identification of voice-onset times: A stable perceptual boundary for voicing categories across speaking rates. Percept Psychophys. 2002; 64:121–130.10.3758/bf03194561 [PubMed: 11916295]

Christovich LA, Lublinskaya VV. The 'center of gravity' effect in vowel spectra and critical distance between the formants: Psychoacoustical study of the perception of vowel-like stimuli. Hearing Res. 1979; 1:185–195.10.1016/0378-5955(79)90012-1

Cole J, Linebaugh G, Munson CM, McMurray B. Unmasking the acoustic effects of vowel-to-vowel coarticulation: A statistical modeling approach. J Phonetics. 2010; 38:167–184.10.1016/j.wocn.2009.08.004

Dahan D, Magnuson JS, Tanenhaus MK. Time course of frequency effects in spoken-word recognition: Evidence from eye movements. Cognitive Psychol. 2001; 42:317–367.10.1006/cogp.2001.0750

Denes P. Effect of duration on the perception of voicing. J Acoust Soc Am. 1955; 27:761–764.10.1121/1.1908020

Diehl RL, Kluender KR. On the objects of speech perception. Ecological Psychol. 1989; 1:121–144.10.1207/s15326969eco0102_2

Galle, ME. Doctoral dissertation. University of Iowa; 2014. Integration of asynchronous cues in fricative perception in real-time and developmental-time.

Goldinger SD. Echoes of echoes? An episodic theory of lexical access. Psychol Rev. 1998; 105:251–279.10.1037//0033-295x.105.2.251 [PubMed: 9577239]

Holt LL. Temporally nonadjacent nonlinguistic sounds affect speech categorization. Psychol Sci. 2005; 16:305–312.10.1111/j.0956-7976.2005.01532.x [PubMed: 15828978]

Johnson, K. Speech perception without speaker normalization: An exemplar model. In: Johnson, K.; Mullenix, JW., editors. Talker Variability in Speech Processing. London: Academic Press; 1997. p. 145-165.

Kessinger RH, Blumstein SE. Effects of speaking rate on voice-onset time in Thai, French, and English. J Phonetics. 1997; 25:143–168.10.1006/jpho.1996.0039

Kessinger RH, Blumstein SE. Effects of speaking rate on voice-onset time and vowel production: Some implications for perception studies. J Phonetics. 1998; 26:117–128.10.1006/jpho.1997.0069

Kluender K. Sensitivity to change in perception of speech. Speech Comm. 2003; 41:59–69.10.1016/s0167-6393(02)00093-6

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

Lotto AJ, Kluender KR. General contrast effects in speech perception: Effect of preceding liquid on stop consonant identification. Percep Psychophys. 1998; 60:602–619.10.3758/bf03206049

Magloire J, Green KP. A Cross-Language Comparison of Speaking Rate Effects on the Production of Voice Onset Time in English and Spanish. Phonetica. 1999; 56:158–185.10.1159/000028449

Marslen-Wilson, W. Activation, competition, and frequency in lexical access. In: Altmann, G., editor. Cognitive Models of Speech Processing: Psycholinguistic and Computational Perspectives. Cambridge, MA: MIT Press; 1990. p. 148-172.

Massaro D, Cohen M. Consonant/vowel ratio: An improbable cue in speech. Percep Psychophys. 1983; 33:501–505.10.3758/bf03202904

McClelland JL, Rumelhart DE. An interactive activation model of context effects in letter perception: Part 1. An account of basic findings. Psychol Rev. 1981; 88:375–407.10.1037//0033-295x.88.5.375

McMurray B, Aslin RN, Tanenhaus MK, Spivey M, Subik D. Gradient sensitivity to within-category variation in speech: Implications for categorical perception. J Exp Psychol: Human Percep Perform. 2008a; 34:1609–1631.10.1037/a0011747

McMurray B, Aslin RN, Toscano JC. Statistical learning of phonetic categories: insights from a computational approach. Developmental Science. 2009; 12:369–378.10.1111/j.1467-7687.2009.00822.x [PubMed: 19371359]

McMurray B, Clayards MA, Tanenhaus MK, Aslin RN. Tracking the time course of phonetic cue integration during spoken word recognition. Psychonomic Bull Rev. 2008b; 15:1064–1071.10.3758/pbr.15.6.1064

McMurray B, Jongman A. What information is necessary for speech categorization? Harnessing variability in the speech signal by integrating cues computed relative to expectations. Psychol Rev. 2011; 118:219–46.10.1037/a0022325 [PubMed: 21417542]

McMurray B, Kovack-Lesh K, Goodwin D, McEchron W. Infant directed speech and the development of speech perception: Enhancing development or an unintended consequence? Cognition. 2013; 129:362–378.10.1016/j.cognition.2013.07.015 [PubMed: 23973465]

McMurray B, Tanenhaus MK, Aslin RN. Within-category VOT affects recovery from ”lexical” garden-paths: Evidence against phoneme-level inhibition. J Memory Language. 2009; 60:65–91.10.1016/j.jml.2008.07.002

Miller JL, Dexter ER. Effects of speaking rate and lexical status on phonetic perception. J Exp Psychol: Human Percep Perform. 1988; 14:369–378.10.1037//0096-1523.14.3.369

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; 43:106–115.10.1159/000261764

Miller JL, Grosjean F. How the components of speaking rate influence perception of phonetic segments. J Exp Psychol: Human Percep Perform. 1981; 7:208–215.10.1037//0096-1523.7.1.208

Miller JL, Grosjean F, Lomanto C. Articulation rate and its variability in spontaneous speech: A reanalysis and some implications. Phonetica. 1984; 41:215–225.10.1159/000261728 [PubMed: 6535162]

Miller JL, Liberman AM. Some effects of later-occurring information on the perception of stop consonant and semivowel. Percep Psychophys. 1979; 25:457–65.10.3758/bf03213823

Miller J, Patterson T, Ulrich R. Jackknife-based method for measuring LRP onset latency differences. Psychophysiology. 1998; 35:99–115.10.1111/1469-8986.3510099 [PubMed: 9499711]

Miller JL, Volaitis LE. Effect of speaking rate on the perceptual structure of a phonetic category. Percep Psychophys. 1989; 46:505–512.10.3758/bf03208147

Miller JL, Wayland SC. Limits on the limitations of context-conditioned effects in the perception of [b] and [w]. Percep Psychophys. 1993; 54:205–10.10.3758/bf03211757

Nearey TM. Modeling the role of inherent spectral change in vowel identification. J Acoust Soc Am. 1986; 80:1297–1308.10.1121/1.394433

Nearey TM. The segment as a unit of speech perception. J Phonetics. 1990; 18:347–373.

Nearey TM. Speech perception as pattern recognition. J Acoust Soc Am. 1997; 101:3241–3254.10.1121/1.418290 [PubMed: 9193041]

Nearey TM, Rochet BL. Effects of place of articulation and vowel context on VOT production and perception for French and English stops. Journal of the International Phonetic Association. 1994; 24:1–18.10.1017/s0025100300004965

Oden GC, Massaro DW. Integration of featural information in speech perception. Psychol Rev. 1978; 85:172–191.10.1037//0033-295x.85.3.172 [PubMed: 663005]

Pashler H. Dual-task interference in simple tasks: Data and theory. Psychological Bulletin. 1994; 116:220–244.10.1037/0033-2909.116.2.220 [PubMed: 7972591]

Pind J. Speaking rate, voice-onset time, and quantity: The search for higher-order invariants for two Icelandic speech cues. Percep Psychophys. 1995; 57:291–304.10.3758/bf03213055

Pisoni, DB. Some thoughts on “normalization” in speech perception. In: Johnson, K.; Mullenix, JW., editors. Talker Variability in Speech Processing. San Diego: Academic Press; 1997. p. 9-32.

Pisoni DB, Carrell TD, Gans SJ. Perception of the duration of rapid spectrum changes in speech and nonspeech signals. Percep Psychophys. 1983; 34:314–322.10.3758/bf03203043

Port RF, Dalby J. Consonant/vowel ratio as a cue for voicing in English. Percep Psychophys. 1982; 32:141–152.10.3758/bf03204273

Reinisch E, Sjerps MJ. The uptake of spectral and temporal cues in vowel perception is rapidly influenced by context. J Phonetics. 2013; 41:101–116.10.1016/j.wocn.2013.01.002

Repp BH. On levels of description in speech research. J Acoust Soc Am. 1981; 69:1462–1464.10.1121/1.385779 [PubMed: 7240580]

Repp BH. Phonetic trading relations and context effects: New experimental evidence for a speech mode of perception. Psychol Bull. 1982; 92:81–110.10.1037//0033-2909.92.1.81 [PubMed: 7134330]

Salverda AP, Dahan D, McQueen JM. The role of prosodic boundaries in the resolution of lexical embedding in speech comprehension. Cognition. 2003; 90:51–89.10.1016/s0010-0277(03)00139-2 [PubMed: 14597270]

Shinn PC, Blumstein SE, Jongman A. Limitations of context conditioned effects in the perception of [b] and [w]. Percep Psychophys. 1985; 38:397–407.10.3758/bf03207170

Smits R. Evidence for Hierarchical Categorization of Coarticulated Phonemes. J Exp Psychol: Human Percep Perform. 2001a; 27:1145–1162.10.1037/0096-1523.27.5.1145

Smits R. Hierarchical categorization of coarticulated phonemes: A theoretical analysis. Percep Psychophys. 2001b; 63:1109–1139.10.3758/bf03194529

Summerfield Q. Articulatory rate and perceptual constancy in phonetic perception. J Exp Psychol: Human Percep Perform. 1981; 7:1074–95.10.1037//0096-1523.7.5.1074

Syrdal AK, Gopal HS. A perceptual model of vowel recognition based on the auditory representation of American English vowels. J Acoust Soc Am. 1986; 79:1086–1100.10.1121/1.393381 [PubMed: 3700864]

Tanenhaus MK, Spivey-Knowlton MJ, Eberhard KM, Sedivy JE. Integration of visual and linguistic information in spoken language comprehension. Science. 1995; 268:632–34.10.1126/science.7777863

Toscano JC, McMurray B. Cue integration with categories: Weighting acoustic cues in speech using unsupervised learning and distributional statistics. Cog Sci. 2010; 34:434–464.10.1111/j.1551-6709.2009.01077.x

Toscano JC, McMurray B. Cue integration and context effects in speech: Evidence against speaking rate normalization. . Attn Percep Psychphys. 2012; 74:1284–1301.10.3758/s13414-012-0306-z

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. Psychol Sci. 2010; 21:1532–1540.10.1177/0956797610384142 [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. Berlin: Walter de Gruyter; 2006. p. 1-28.

Utman JA. Effects of local speaking rate context on the perception of voice-onset time in initial stop consonants. J Acoust Soc Am. 1998; 103:1640–1653.10.1121/1.421297 [PubMed: 9514028]

Wayland SC, Miller JL, Volaitis LE. The influence of sentential speaking rate on the internal structure of phonetic categories. J Acoust Soc Am. 1994; 95:2694–2701.10.1121/1.409838 [PubMed: 8207142]

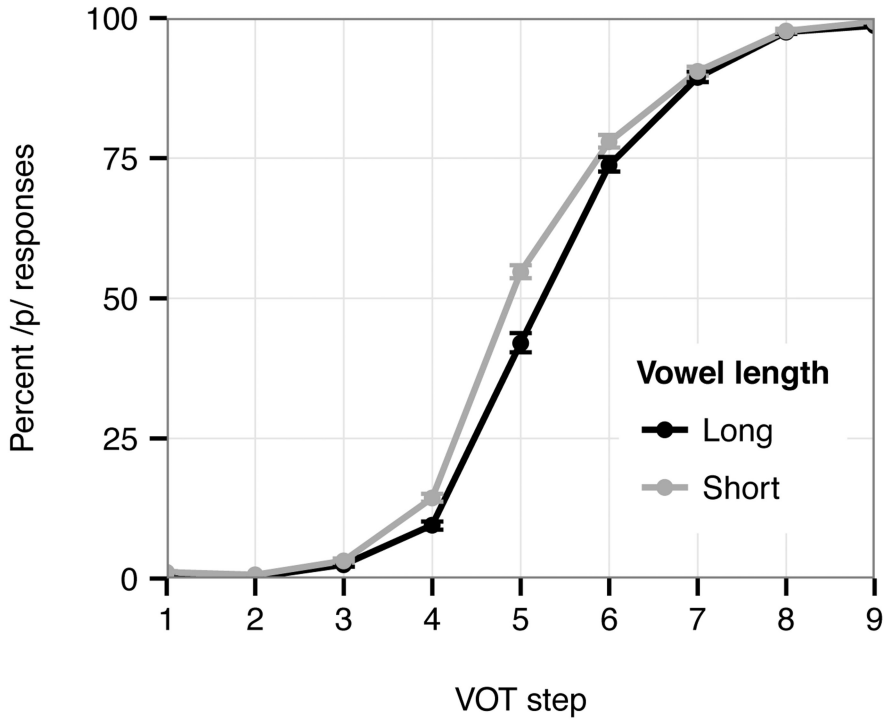

Fig. 1.

Percent of /p/ mouse-click responses in Experiment 1 as a function of VOT and VL. Error bars indicate standard error.

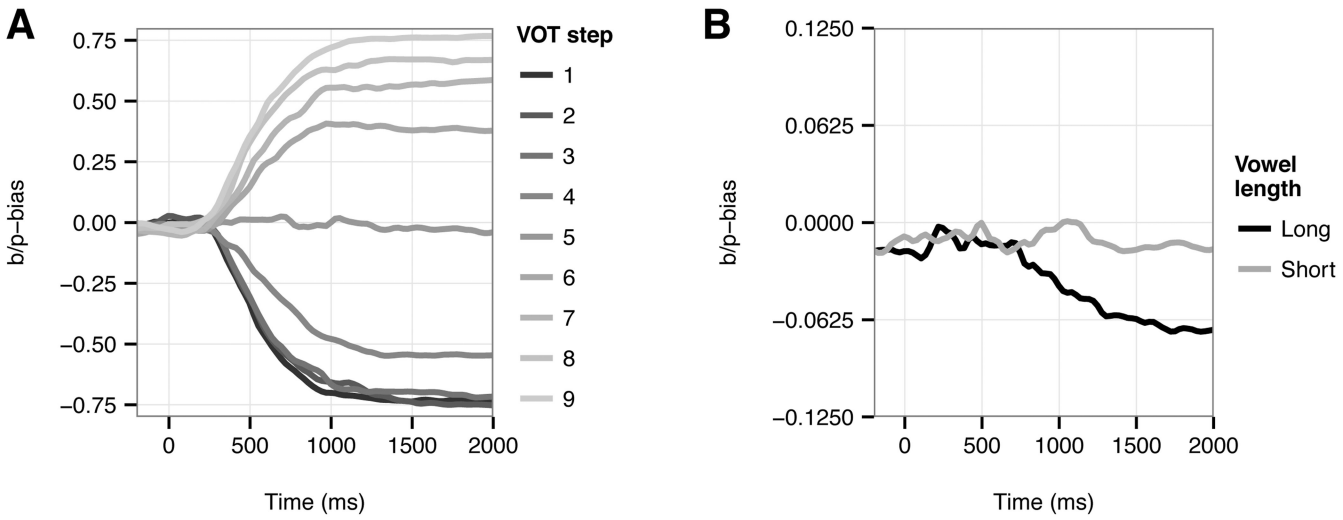

Fig. 2.

Time-course of *b/p-bias* as a function of **(A)** VOT and **(B)** VL. In each panel, the data are collapsed across the other acoustic dimension (i.e., in (A), the time-course reflects differences in VOT collapsed across the VL conditions). 0 ms is the onset of the target word.

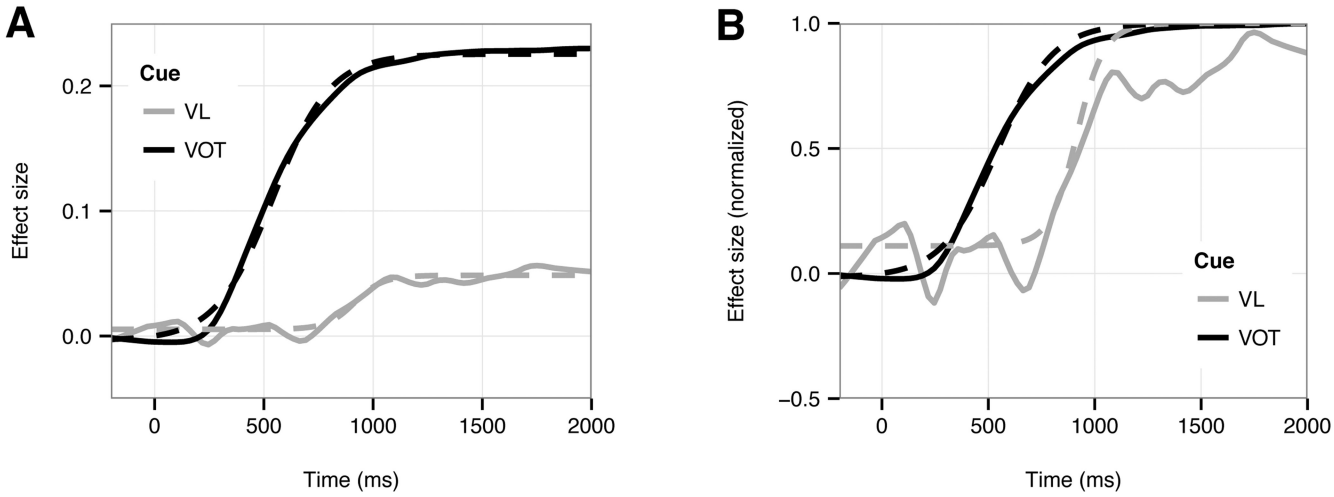

Fig. 3.

(A) Time-course of effect for each cue (VOT and VL). **(B)** Time-course with normalized effect sizes. Dashed lines indicate average model fits to data.

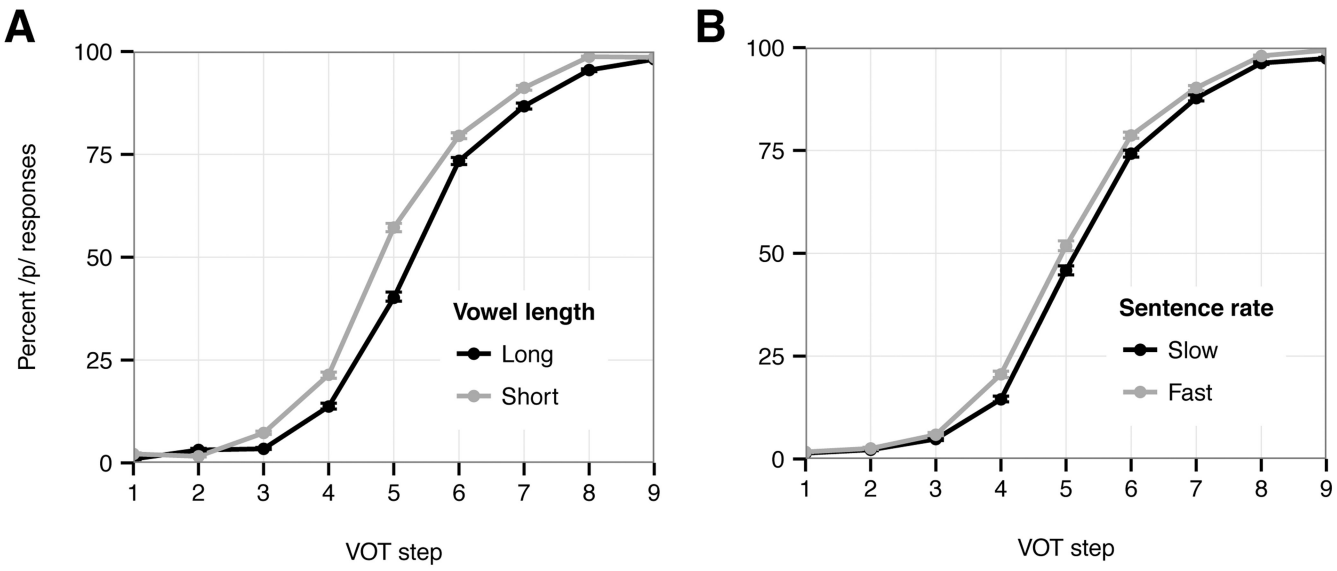

Fig. 4.

(A) Mouse-click responses in Experiment 2 as a function of VOT and VL. **(B)** Mouse-click responses as a function of VOT and SR. Error bars indicate standard error.

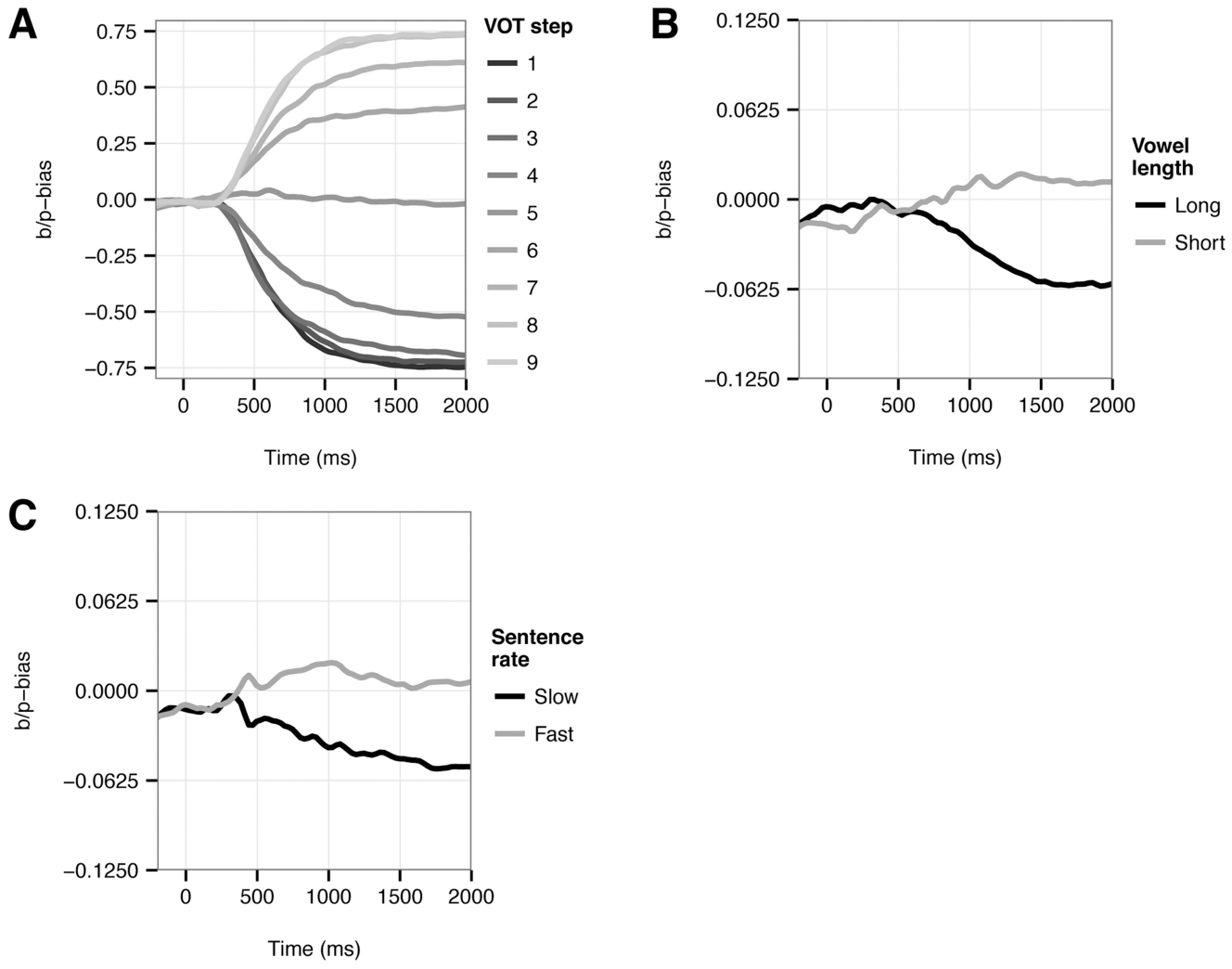

Fig. 5.

Time-course of *b/p-bias* as a function of **(A)** VOT, **(B)** VL, and **(C)** SR. The data in each panel are collapsed across the other two acoustic dimensions (e.g., in (A), the time-course shows *b/p-bias* as a function of VOT, averaged across the two VL and two SR conditions).

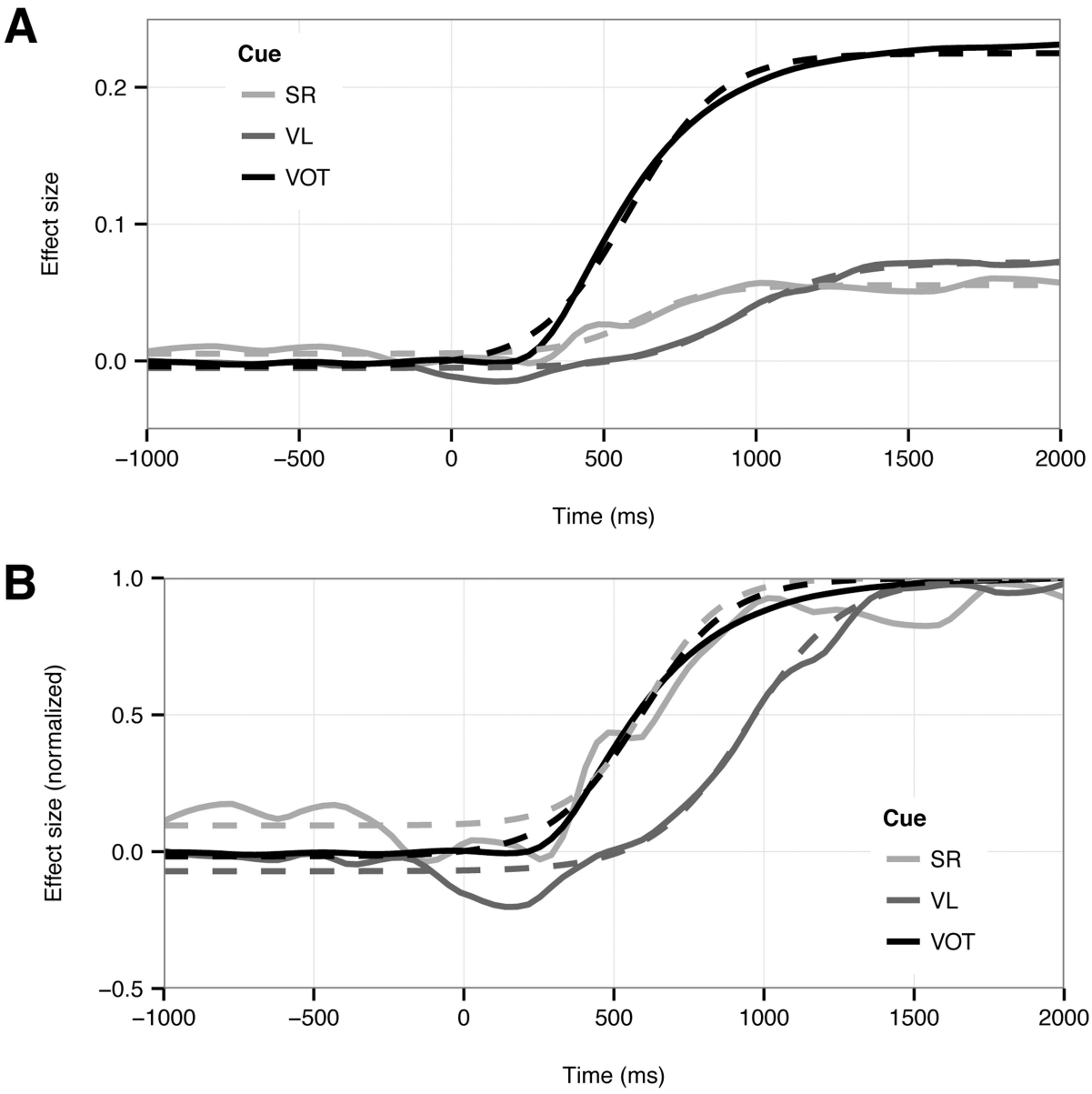

Fig. 6.

(A) Time-course of effect size for VOT, VL, and SR in Experiment 2. **(B)** Time-course showing normalized effect sizes. In both panels, dashed lines indicate average model fits.

Table 1

Carrier phrases used in the experiments.

On this screen, click on the…
In this display, choose the…
On this screen, select the…
In this display, pick the…
On this screen, please choose the…