Running Head: TIME COURSE OF PHONETIC CUE INTEGRATION

Tracking the time course of phonetic cue integration during spoken word recognition

Bob McMurray
Department of Psychology
University of Iowa

Meghan A. Clayards

Michael K. Tanenhaus

and

Richard N. Aslin
Department of Brain and Cognitive Sciences
University of Rochester

Corresponding Author
Bob McMurray
Dept. of Psychology
E11 SSH
University of Iowa
Iowa City, IA 52240
Phone: 319-335-2408
Fax: 319-335-0191
bob-mcmurray@uiowa.edu

Word Count: 3997

Abstract

Speech perception requires listeners to integrate multiple cues that each contribute to judgments about a phonetic category. Classic studies of trading relations assessed the weights attached to each cue, but did not explore the time-course of cue-integration. Here we provide the first direct evidence that asynchronous cues to both voicing (b/p) and manner (b/w) contrasts become available to the listener at different times during spoken word recognition. Using the Visual World paradigm, we show that the probability of eye movements to pictures of target and competitor objects diverge at different points in time after the onset of the target word. These points of divergence correspond to the availability of early (voice-onset-time or formant transition slope) and late (vowel length) cues to voicing and manner contrasts. These results support a model of cue-integration in which phonetic cues are used for lexical access as soon as they are available.

Key Words: Spoken word recognition, Speech perception, trading relations, time course, cue integration.

Tracking the time course of phonetic cue integration during spoken word recognition

Spoken word recognition requires that the perceptual system cope with a noisy signal, sequential inputs that persist only a fleeting moment, and temporary ambiguity as words unfold. Two particularly challenging aspects of this are that phonemic and lexical contrasts are rarely instantiated along a single dimension (cue) and information from disparate cues often arrives asynchronously. Therefore, at any point in time, the word-recognition system rarely has all of the relevant information for a phonetic distinction.

Integration of asynchronous cues has not been a focus of prior theories of spoken word recognition. One approach would be for the system to store early cues in a temporary buffer until the remaining cues have arrived. This would minimize premature (incorrect) commitments, but force the system to delay making even a partial decision. Alternatively, each cue could provide partial evidence for higher-level units (e.g., features, phonemes, or words) as soon as it arrived. This would speed up recognition, if early commitments are correct, but risks delay if they must be revised. Of course, the system could buffer some cues but treat others continuously; buffers may exist at multiple levels of representation, and some buffers might release preliminary analyses for higher-level processing.

We examined two phonetic distinctions that are determined by asynchronous cues to ask whether lexical activation is immediately sensitive to the early information or is delayed until both cues are available. Existing models have not addressed this question directly. The Fuzzy Logical Model of Perception (Oden & Massaro, 1978) assumes cues are simultaneously available (which would be true in a buffered system), but could likely function as an immediate integrator by treating cues that have not arrived yet as ambiguous (Oden, personal communication). TRACE (McClelland & Elman, 1986)—which is clearly consistent with

continuous integration—makes the simplifying assumption that cues are available simultaneously, and it is unclear if this is necessary for the dynamics of the model. There has also been debate about asynchrony in models of feature parsing. Gow (2003) argues that cues must be buffered and integrated while Fowler (1984) argues for continuous integration. Thus, answering this question would have implications for a number of approaches to cue integration.

We examined two word-initial consonant contrasts (voicing and manner) that are first cued by information at word onset and later by the length of the vowel. Word-initial voicing (e.g., /p/ versus /b/) is cued by voice onset time (VOT), along with other cues (e.g., pitch and first formant frequency). Vowel length varies systematically with voicing (Allen & Miller, 1999; Kessinger & Blumstein, 1998) and it participates in a trading relation with VOT (Summerfield, 1981, Miller & Volaitis, 1989); an ambiguous VOT is perceived as a voiced consonant when followed by a long vowel and as a voiceless consonant with a short vowel. Word-initial manner of articulation (e.g., /b/ versus /w/) is cued by the slope of the formant transitions. For medium transitions the likelihood of perceiving a /w/ increases with short vowels (Miller & Liberman, 1979; Miller & Wayland, 1993; though see Shinn, Blumstein & Jongman, 1995). Effects of vowel length can be separated from effects of sentential speaking rate (Summerfield, 1981; Wayland, Miller & Volaitis, 1994) and can be observed across a range of rates (Allen & Miller, 1999). Thus, vowel length appears to operate as an independent cue (Repp, 1982).

Although numerous studies have examined the final product of cue integration (Repp, 1982 for a classic review), few have assessed the time course of integrating asynchronous cues. Miller and Dexter (1988) is one notable exception. They found that vowel length had weaker effects on perceived voicing when participants responded quickly in a phoneme judgment task,

suggesting that participants based their earliest responses primarily on the VOT. However, the boundary indicated by these early responses favored more /p/ responses, suggesting that listeners treated the incomplete vowel length as short. Thus, it is not clear which model these data support.

In addition, their use of a metalinguistic phoneme decision task raises several possibilities. First, the system may be fundamentally buffered, but the act of making an overt phoneme response may force this buffer to be flushed (when it would not normally be during online recognition). Second, it is possible that a representation that is continuous at the phonemic level is buffered before it is available to lexical activation processes—measuring only lower level processes may miss this. Finally, given uncertainty about whether phoneme representations are used in word recognition (e.g., Gaskell, Quinlan, Tamminen & Cleland, in press), it is important to ask when cues affect *lexical activation* to determine if cues are obligatorily integrated prior to lexical access, or can affect activation directly.

The current work used 9-step VOT (voicing) and formant-transition (manner) continua, with two vowel lengths, to determine if early cues are immediately available to lexical access, or if these cues do not play a role in lexical access until later cues become available. We used eye movements to potential referents to sample listeners' lexical hypotheses as the signal unfolds over time (e.g. Tanenhaus, Spivey-Knowlton, Eberhard & Sedivy , 1995). This is a highly sensitive measure of lexical activation (Allopenna, Magnuson & Tanenhaus, 1998), showing effects of frequency and neighborhood density (Dahan, Magnuson & Tanenhaus, 2001; Magnuson, Dixon, Tanenhaus & Aslin, 2007), and mapping well onto lexical activation from models like TRACE (McClelland & Elman, 1986). Importantly, eye movements are sensitive to subtle variations in VOT (McMurray, Tanenhaus & Aslin, 2002), formant transitions

(Crosswhite, unpublished) and vowel duration (Salverda, Dahan & McQueen, 2003).

Methods

Participants

Thirty-three undergraduates from the University of Rochester, who were monolingual speakers of English with no known hearing problems, were paid $10 on each of two days.

Materials

Auditory stimuli were synthetic one-syllable words comprising four lexical/phonological contrast sets: two test contrasts (b/p and b/w) and two filler contrasts (d/g and l/r). Each set contained three minimal pairs representing the endpoints of a continuum. Within each continuum, vowel quality was constant. The b/p pairs were: *beach/peach*, *bees/peas*, and *beak/peak*; the b/w pairs were: *bell/well, bench/wench*, and *belt/welt*. Filler pairs were: *deuce/goose, dune/goon, dew/goo, race/lace, ray/lay,* and *rake/lake*. Stimuli were synthesized using the KlattWorks (McMurray, in preparation) interface to the Klatt (1980) synthesizer. Formant frequencies were modeled after natural tokens produced by a native speaker of English. Other parameters were set with reference to the spectrogram of the recorded word. Parameters for the initial portion of the stimulus were identical across all continua (within a contrast) to reduce differences between words and allow analysis to be performed across all continua within a set.

Each stimulus file began with 100ms of silence. For the voicing continua, words started with a 5 ms release burst at 60db. For VOTs of 0 ms, voicing (the AV parameter set to 60 dB) started simultaneously with this burst. To construct each step of the VOT continuum, the onset

of AV was delayed in increments of 5 ms from the burst, and 60 dB of aspiration (AH) was added, starting from the onset of the burst and ending at the onset of voicing (Figure 1A).

The manner continua contained no release burst, but the amplitude envelope featured a sudden onset (rather than ramping up). Formant transitions were modeled with logistic functions, which are characterized by a smooth transition between steady states. Slope was varied in nine steps, from 40 Hz/ms for all three formants (a steep slope, /b/) to shallow slopes (/w/): 14 Hz/ms (for F2 and F3) and 8 Hz/ms (for F1). As the slope decreased, the midpoint of the function was also delayed (by up to 20 ms for the endpoint /w/) (Figure 1b).

For both continua, vowel durations were based on the duration of the naturally recorded utterance. We created two vowel length conditions by shortening or lengthening the original vowel by 50 ms (see Table 1). Details for how the l/r and d/g continua were constructed are available from the first author. The visual stimuli were 24 canonical pictures depicting each lexical item, edited to remove extraneous content.

Procedure

Participants were first familiarized with the (written) names of the pictures. During testing, four pictures were presented on each trial: two pictures for each experimental pair (e.g., *beach* and *peach*), and two pictures for a filler contrast. Voicing trials were paired with l/r fillers. Manner trials were paired with d/g fillers. The pairing of words was constant across trials (e.g., *beach/peach* was consistently paired with *lake/rake*) but randomized between participants. Pictures were positioned in the corners of the 19 in. monitor along with a central fixation circle. Participants clicked on the circle after it changed color (500 ms after display onset), triggering the auditory stimulus (presented through Sennheiser HD570 headphones). They then clicked on the referent with a computer mouse.

Stimuli were presented randomly and repeated five times. Because this resulted in a large number of trials (1080), the experiment was divided into two one-hour sessions on consecutive days.

Eye movements were recorded using a SR Eyelink II eye-tracker, sampling at 250 Hz. Eye movements were recorded from the onset of the auditory stimulus until the response. The eye movement record was parsed into saccades and fixations using the default parameter-set in the Eyelink software. Saccades that occurred too early to be driven by the speech signal were excluded (saccades generated during the first 300 ms of the trial: 100 ms of silence at the onset of the auditory stimulus file plus the 200 ms oculomotor planning delay).

Saccades were combined with the subsequent fixation into a "look" which began at the onset of the saccade and ended at the offset of the fixation. In classifying the object to which looks were directed, the object borders on the screen were extended by 150 pixels in each direction to account for noise in the eye-track. Adjacent pictures were separated by 480 pixels (horizontal) and 224 pixels (vertical) so there was little chance of a misclassification.

Results

First we analyzed the mouse-click responses to verify that the stimuli produced the desired shifts in category boundaries. Second, and, most importantly, the fixation data were used to measure the effect of each cue over time. Finally, we assessed whether participants' initial biases reflected an obligatory use of the short portion of the vowel heard at that point.

Voicing

Mouse-click results. Six of the 33 participants categorized end-point stimuli less than 75% correctly and were excluded from the analysis.

As shown in Figure 2, vowel length shifted the VOT boundary by 8 ms in the expected direction (more voiceless judgments for short vowels). A logistic function was fit to each participant's data for each of the three word-pairs and two vowel lengths. The boundaries of these functions were compared in a 2 (vowel length) x 3 (word-pair) ANOVA. We found a significant main effect of vowel length ($F(1,26)=46.7$, $p=0.0001$), and no effect of word-pair ($F(2,52)=2.3$, $p>0.1$). Vowel length was significant for each continuum (*Beach/Peach:* $T(26)=5.8$, $p=0.0001$; *Beak/Peak:* $T(26)=5.7$, $p=0.0001$; *Bees/Peas:* $T(26)=3.9$, $p=0.001$). There was a significant interaction ($F(2,52)=4.6$, $p=0.014$) because the vowel length effect for *bees/peas* was smaller than the other continua.

Timing of the Effects. Figure 3 shows the likelihood that a look (at any time) was directed to the /b/ or /p/ objects as a function of VOT and vowel length. There are clear effects of both VOT and vowel length throughout the time-course of processing.

To evaluate the timing of these effects, we first computed the proportion of fixating each object as a function of time (as in Figure 3), for each participant in each condition. This was smoothed with an 80 ms asymmetrical sawtooth window, in which points 84 ms prior to the current point received a weight of 0 and the weight rose linearly to 1 at the point in question. Thus, any given point was only smoothed by data that occurred *before* it (so later cues could not influence the estimate). We then computed a single variable—b/p-bias, the difference between the likelihood of fixating the voiced and voiceless competitor (at any given point in time). This variable will be near 0 if participants are equi-biased, or if the component values are small. To the extent that it is non-zero, it reflects a commitment to either the voiced or voiceless category.

From this variable, we computed, at each four ms time-step, a measure of the effect size of VOT and vowel length. For VOT this was the slope of the regression line relating b/p-bias to VOT. For vowel length, this was the difference in b/p-bias between long and short vowels. Figure 4 shows the mean effect sizes as a function of time. There is a clear difference in their time course: VOT departs from zero earlier than vowel length. Since each cue lies along a different scale, to compare their timing, the onset of a cue was determined as the point at which each effect reached some proportion of its maximum value (e.g., Miller, Patterson & Ulrich, 1998). We chose four points: 10%, 15%, 20% and 30%.

Because fixation data in the visual world paradigm are not typically reliable for individual participants (particularly with small effects), we adapted the jackknife method used by Miller et al. (1998) for event-related potentials. This technique extracts test statistics from the full dataset averaged across every participant *except one,* and then repeats this process, excluding each participant in turn. The jackknifed data are then subjected to a statistical analysis in which error terms are adjusted to reflect the fact that each participant contributes N-1 times towards the variance, resulting in a very conservative analysis, particularly for large sample sizes (see Shao & Tu, 1995, for a review).

From the jackknifed data, we first computed the effect size at each time point, and from this, the time that the effect-size in each condition crossed 10%, 15%, 20%, and 30% of its maximum value and remained there for at least 40 ms. This criterion prevented small bumps in the function from driving these estimates.

Fixations were affected by VOT before vowel length. The vowel length and VOT effects were marginally different at the 10% point ($T_{jackknifed}(26)=1.87$, $p=.07$), though the means differed in the expected direction ($M_{vot}=279$ ms, $M_{vowel}=349$ ms). The effects crossed 15%, 20%

and 30% at significantly different times. At 15% of maximum, VOT appeared at 315 ms and vowel length at 395 ms ($T_{jackknifed}(26)=2.08$, p=.047). At 20%, VOT appeared at 344 ms and vowel length at 424 ms ($T_{jackknifed}(26)=3.07$, p=.0049). And at 30%, VOT appeared at 398 ms and vowel length at 463 ms ($T_{jackknifed}(26)=2.28$, p=.03).

Initial Bias. Our final analysis asked whether participants treated the early portion of the vowel as "short", suggesting an obligatory use of whatever vowel length information was available at that time (Miller & Dexter, 1988). If this were the case, early fixations should be biased to the voiceless object, particularly for VOTs near the boundary. To control for the effect of VOT, we examined tokens one, two and three steps from each subject's boundary and computed the proportion of fixations to the /b/ and /p/ objects generated prior to the end of the short vowels (M=296 ms). At three steps away, participants were significantly biased to /b/ on the voiced side (T(26)=3.7, p=.001) and to /p/ on the voiceless side (T(26)=2.7, p=.01). However, at steps closer to the boundary there was no significant bias (all p>.2).

Manner Results

Mouse-click results. Only one participant was excluded using the 75% end-point criterion, leaving 32 participants for analysis.

Figure 5 displays the percentage of approximant (/w/) responses as a function of formant transition slope and vowel length. The boundary is not well centered along this continuum, appearing between steps 6 and 7. Nonetheless vowel length shifts the boundary by approximately one step. Logistic functions for each participant for each of the word-pairs were fit and the boundaries of these functions were compared in a 2 (vowel length) x 3 (word-pair) ANOVA. As predicted, we found a significant main effect of vowel length (F(1,31)=18.4,

p=0.0001). Word-pair was unexpectedly significant (F(2,62)=13.9, p=0.0001) with the *bench/wench* boundary shifted toward the /b/ end of the continuum, perhaps due to the nasal or affricate. The interaction was marginally significant (F(2,62)=3.0, p=0.055), but all continua had boundary shifts in the expected direction. Paired T-tests revealed significant effects of vowel for *bell/well* (D=.52 steps; T(31)=3.6, p=0.001), and *bench/wench* (D=.97; T(31)=6.1, p=0.001), but not for *belt/welt* (D=.32; T(31)=1.1, p=0.2).

Timing of the Effects. Figure 6 shows the overall pattern of fixations to the stop and approximant competitors as a function of time, formant transition slope, and vowel length. Both experimental factors clearly affected fixations.

Figure 7 shows the effects sizes of the formant transition and vowel length measures on the b/w-bias measure (looks to /b/ minus looks to /w/) as a function of time after word onset. The effect of formant transition was the regression slope relating step number to b/w-bias; the effect of vowel length was the difference in b/w-bias between long and short vowels. The effect of the onset cue (formant transition) departs from zero earlier than vowel length.

We again used the jackknife procedure to determine when each effect size reliably crossed 10%, 15%, 20% and 30% of its maximum value and stayed there for 40 ms. The vowel length and formant transition slope effects crossed 10% of their maximum at significantly different times, with formant transition slope appearing at 289 ms and vowel length at 441 ms ($T_{jackknifed}(31)=4.94$, p=.0001). The 15% point was also significantly different, with formant-transition slope appearing at 337 ms and vowel length at 454 ms ($T_{jackknifed}(31)=3.93$, p=.0004). Likewise, the 20% point showed a significant difference ($M_{formant}$=383 ms, M_{vowel}=467 ms, $T_{jackknifed}(31)=2.82$, p=.008). The 30% point, however, was not significantly different ($M_{formant}$=457 ms, M_{vowel}=489 ms, $T_{jackknifed}(31)=1.24$, p>.2), reflecting the fact that while it starts

later, the effect of vowel length appears to "catch-up" rapidly.

Initial Bias. To assess whether participants treated the initial portion of the vowel as "short" (and were hence biased toward /w/), we compared looks to the /b/ and /w/ objects over the first 345 ms (the mean length of the short vowel). We examined three steps on either side of the boundary. On the stop side, participants were biased toward /b/ at three steps from the boundary ($T(31)=2.3$, $p=.03$), marginally biased toward /b/ at two ($T(31)=1.7$, $p=.1$), and unbiased at one step ($T(31)=.9$, $p>.2$). On the approximant side they were biased toward /w/, three steps from the boundary ($T(29)=2.8$, $p=.01$), marginally so at two steps ($T(31)=1.7$, $p=.09$) and w-biased one-step away ($T(31)=2.4$, $p=.02$). Thus, the only evidence for a bias toward /w/ early in processing comes from steps in which the formant-transitions were consistent with /w/.

Discussion

Our results provide support for continuous integration of word-initial voicing and manner of articulation based on early consonantal and later vowel length cues. In both cases, the effect of the onset cue (VOT and formant transition slope) preceded the effect of vowel length and there was little evidence that the early portion of the vowel was treated as short (biasing toward /w/ or /p/). The system does not appear to wait until both cues are available to make preliminary commitments. This offers a partial explanation for why lexical activation is gradiently sensitive to continuous cues like VOT (e.g., Andruski, Blumstein & Burton, 1994; McMurray et al., 2002). Preserving the continuity of cues like VOT is a fundamental requirement for continuous integration and updating. In contrast a categorical decision prior to accessing the lexicon would treat a VOT of 25 ms, which is close to the b/p boundary and strongly conditioned by vowel length, as equivalent to a VOT of 60 ms, which is less affected by vowel length.

Our results rule out both an encapsulated buffer for the cues we studied, and the notion

that integration of cues is a prerequisite to word recognition. We did not, however, assess other possibilities. Buffering could vary by cue or contrast, and where buffers are found, they could exist at subcategorical or categorical levels of processing. Such buffers might vary in their ability to release (cascade) preliminary analyses. For example, Kingston (2005) argues that *integral* cues are integrated at an auditory level and cannot have independent effects on perception, whereas *separable* cues can be weighted according to language-specific phonetic experience.

While we studied voicing and manner contrasts as representative of the problem, it will be important to consider other cues as well as cue-integration problems that cross word boundaries such as assimilation and long distance phonetic dependencies. These results, however, suggest that for at least some cues, word recognition can make immediate, partial commitments without waiting for disambiguating information.

Acknowledgements

The authors would like to thank Dana Subik for assistance with data collection, Steve Luck for suggesting the Jackknife analysis, and Joanne Miller and Gregg Oden for helpful comments during the development of this project. This work was supported by NIH grants DC006537 and DC008089 to BM, and DC005071 to MKT and RNA

References

Allen, J.S. & Miller, J.L. (1999) Effects of syllable-initial voicing and speaking rate on the temporal characteristics of monosyllabic words. *Journal of the Acoustical Society of America, 106,* 2031-2039.

Allopenna, P.D., Magnuson, J.S. & Tanenhaus, M.K. (1998). Tracking the time course of spoken word recognition using eye-movements: evidence for continuous mapping models. *Journal of Memory and Language, 38(4)*, 419-439.

Andruski, J.E., Blumstein, S.E. & Burton, M.W. (1994) The effect of subphonetic differences on lexical access. *Cognition, 52,* 163-187.

Dahan, D., Magnuson, J.S & Tanenhaus, M.K. (2001). Time course of frequency effects in spoken-word recognition: Evidence from eye movements. *Cognitive Psychology, 42,* 317-367.

Fowler, C. (1984) Segmentation of coarticulated speech in perception. *Perception & Psychophysics, 36,* 359-368.

Gaskell, M.G., Quinlan, P., Tamminen, J., and Cleland, A.A. (in press) The nature of phoneme representation in spoken word recognition. *Journal of Experimental Psychology: General.*

Gow, D. (2003) Feature parsing: Feature cue mapping in spoken word recognition. *Perception & Psychophysics, 65(4),* 575-590.

Kessinger, R.H. & Blumstein, S.E. (1998) Effects of speaking rate on voice onset time and vowel production: some implications for perception studies. *Journal of Phonetics, 26,* 117-128.

Kingston, J. (2005). Ears to categories: New arguments for autonomy. *Prosodies: With Special*

Reference to Iberian Language. S. Frota, M. Vigario & M.J. Freitas (eds), Berlin: Mouton de Gruyter. 177-222

Klatt, D. (1980) Software for a Cascade/Parallel Synthesizer. *Journal of the Acoustical Society of America, 67,* 971-995.

Magnuson, J. S., Dixon, J. A., Tanenhaus, M. K., & Aslin, R. N. (2007). The dynamics of lexical competition during spoken word recognition. *Cognitive Science*, *31*, 133-156.

McClelland, J. & Elman, J. (1986) The TRACE model of speech perception. *Cognitive Psychology, 18(1),* 1-86.

McMurray (in preparation). KlattWorks: A [somewhat] new systematic approach to formant-based speech synthesis for empirical research.

McMurray, B., Tanenhaus, M., & Aslin, R. (2002). Gradient effects of within-category phonetic variation on lexical access. *Cognition, 86(2),* B33-B42.

Miller, J.L. & Dexter, E.R. (1988) Effects of speaking rate and lexical status on phonetic perception. *Journal of Experimental Psychology: Human Perception and Performance, 14,* 369-378.

Miller, J.L. & Liberman, A.M. (1979) Some effects of later-occurring information on the perception of stop consonant and semi-vowel. *Perception & Psychophysics, 25,* 457-465.

Miller, J.L. & Volaitis, L.E. (1989) Effect of speaking rate on the perceptual structure of a phonetic category. *Perception & Psychophysics, 46(6*), 505-512.

Miller, J.L. & Wayland, S.C. (1993) Limits on the limitations of context-conditioned effects in the perception of [b] and [w]. *Perception & Psychophysics, 54*, 205-210.

Miller, J.O., Patterson, T., and Ulrich, R. (1998) Jackknife-based method for measuring LRP onset latency differences. *Psychophysiology, 35,* 99-115.

Oden, G. & Massaro, D.W. (1978) Integration of featural information in speech perception. *Psychological Review, 85(3),* 172-191.

Repp, B. (1982) Phonetic trading relations and context effects: New experimental evidence for a speech mode of perception. *Psychological Bulletin, 92(1),* 81-110.

Salverda, A.P., Dahan, D. & McQueen, J. (2003) The role of prosodic boundaries in the resolution of lexical embedding in speech comprehension. *Cognition, 90(1),* 51-89.

Shinn, P., Blumstein, S.E. & Jongman, A. (1985) Limits of context conditioned effects in the perception of [b] and [w]. *Perception & Psychophysics, 38(5),* 397-407.

Shao, J. and Tu, D. (1995). *The Jackknife and Bootstrap.* Springer, New York.

Summerfield, Q. (1981) Articulatory rate and perceptual constancy in phonetic perception. *Journal of the Acoustical Society of America, 7(5),* 1074-1095.

Tanenhaus, M.K., Spivey-Knowlton, M.J., Eberhard, K.M. & Sedivy, J.C. (1995) Integration of visual and linguistic information in spoken language comprehension. *Science, 268*, 1632-1634.

Wayland, S.C., Miller, J.L. & Volaitis, L.E. (1994) The influence of sentential speaking rate on the internal structure of phonetic categories. *Journal of the Acoustical Society of America, 95,* 2694-2701.

Table

Phonetic Contrast	Word-Pair	Vowel length (ms)		Total Length (ms)	
		Short	**Long**	**Short**	**Long**
b/p (VOT)	Beach/peach	105	200	315	415
	Bees/peas	145	245	335	435
	Beak/peak	105	205	240	340
b/w (formant transition)	Bell/well	150	250	290	390
	Belt/welt	165	265	340	440
	Bench/wench	180	280	405	505

Table 1: Vowel and total length of the stimuli in each of the six continua. For all stimuli, vowel length was measured from the onset of voicing to either 1) the offset of voicing, for *beach/peach*, *beak/peak* and *bench/wench*; 2) the onset of frication, for *bees/peas*; or 3) the end of the second formant transition indicating the /l/, for *belt/welt* and *bell/well.*

Figures

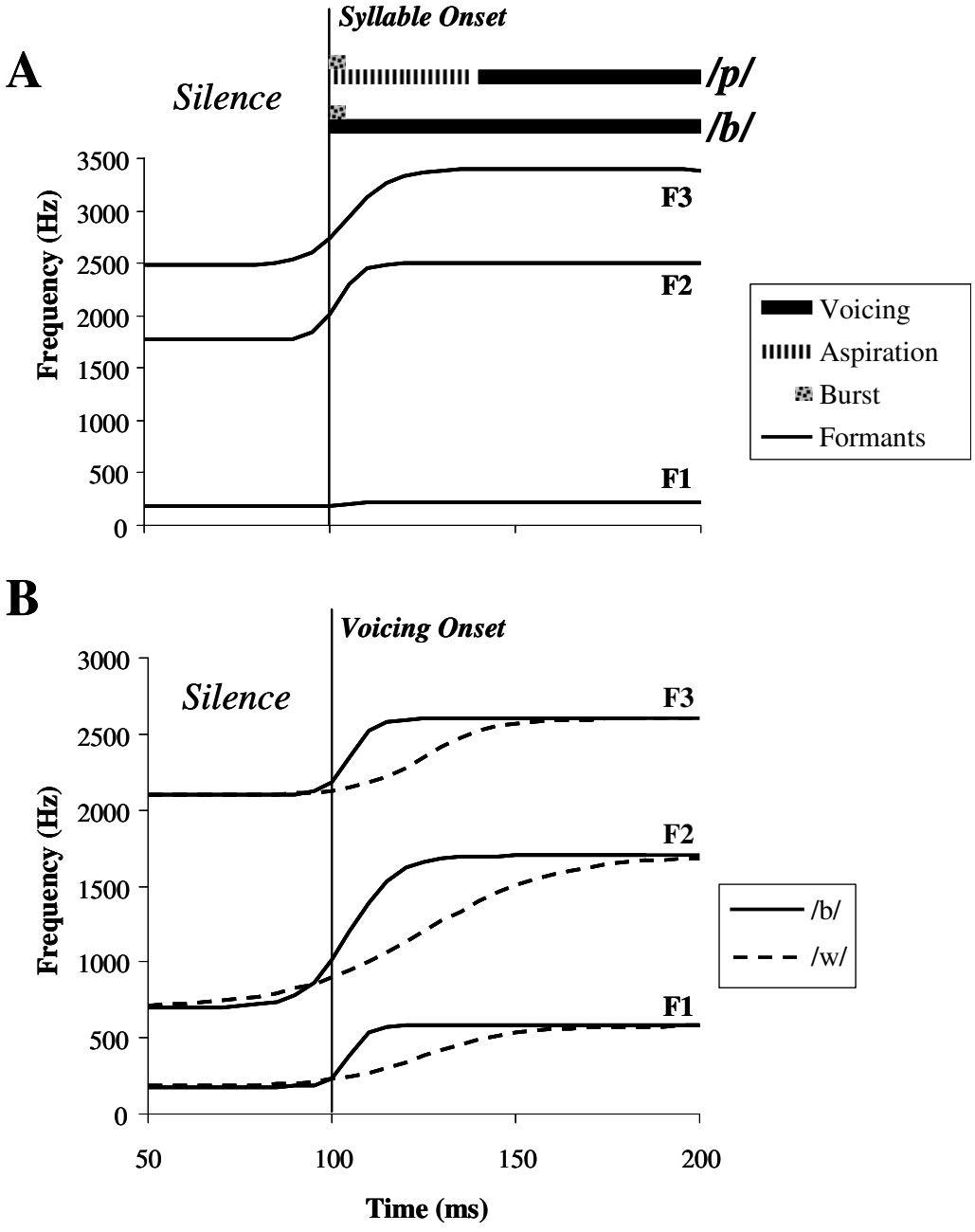

Figure 1: Schematics of the stimuli. A) For VOT continua, the frication burst and formant frequencies were kept constant across VOTs, but the relative onset of voicing (AV) and aspiration (AH) were manipulated. B) For manner continua, the slopes of the first three formants were manipulated simultaneously.

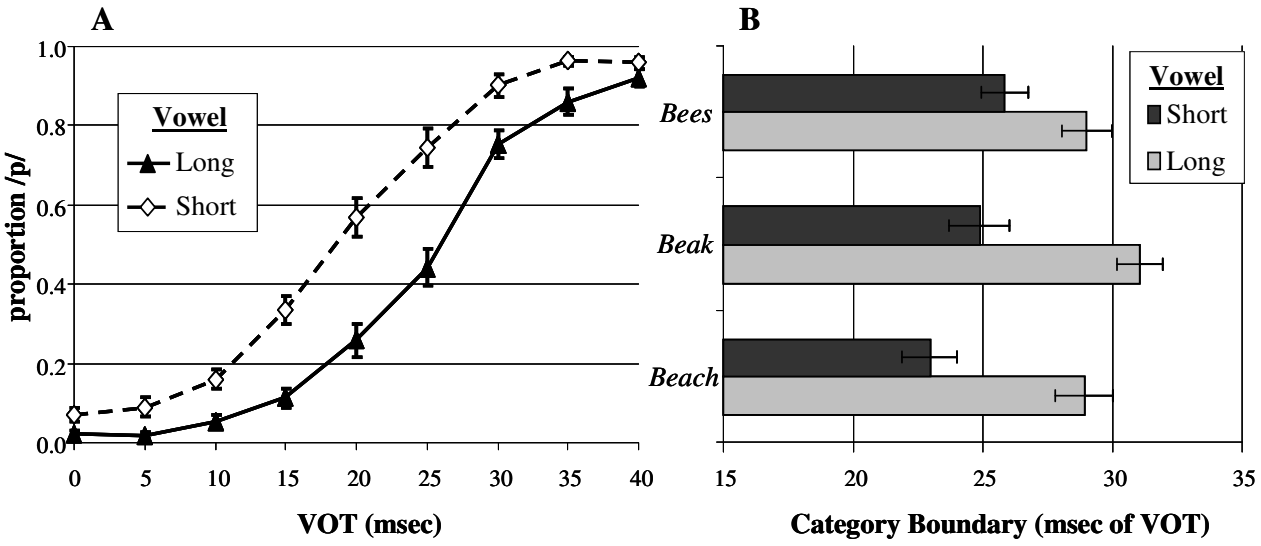

Figure 2: Identification data for VOT continua. A) Proportion /p/ responses as a function of VOT and vowel length averaged across the three voicing continua. B) Computed category boundaries for each continuum as a function of vowel length.

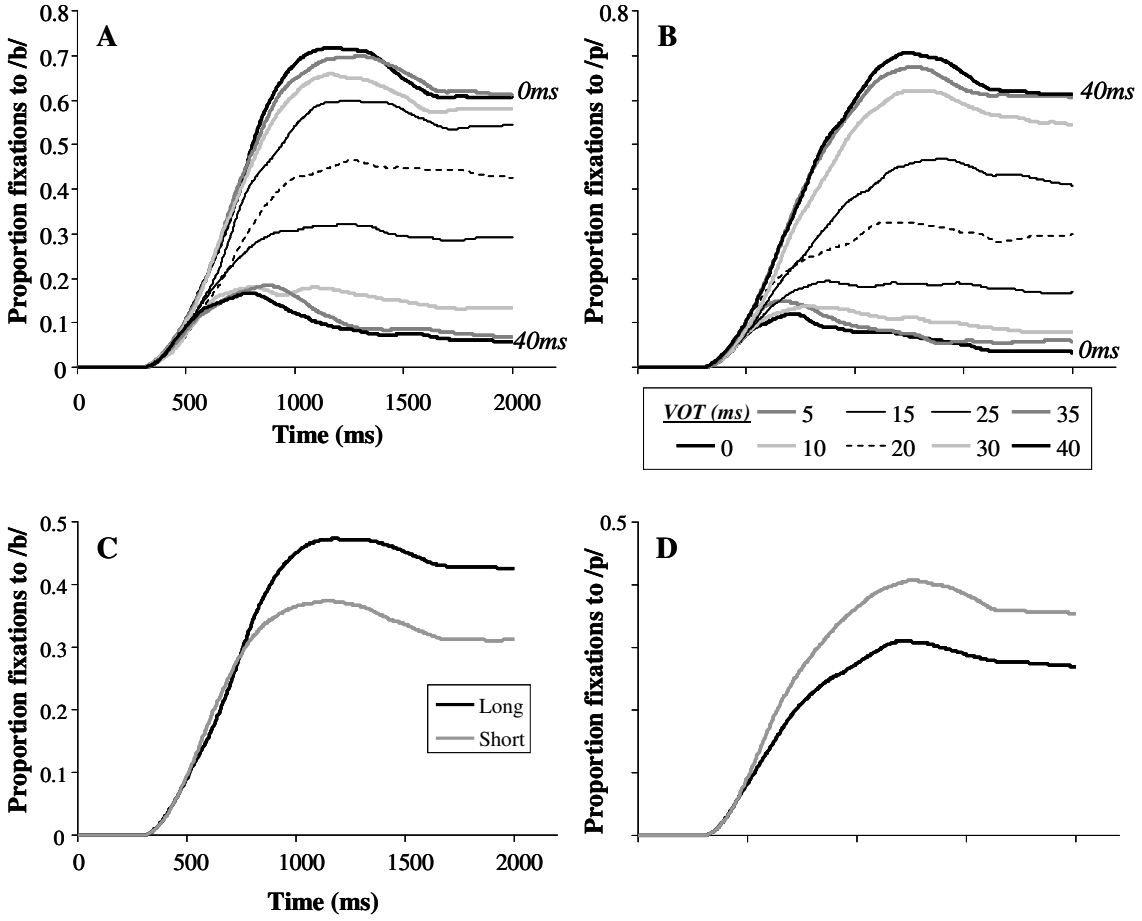

Figure 3: Effect of VOT and vowel length on fixations as a function of time. A) Fixations to voiced targets as a function of VOT and time. B) Fixations to voiceless targets as a function of VOT. C) Fixations to voiced targets as a function of vowel length and time. D) Fixations to voiceless tokens as a function of vowel length.

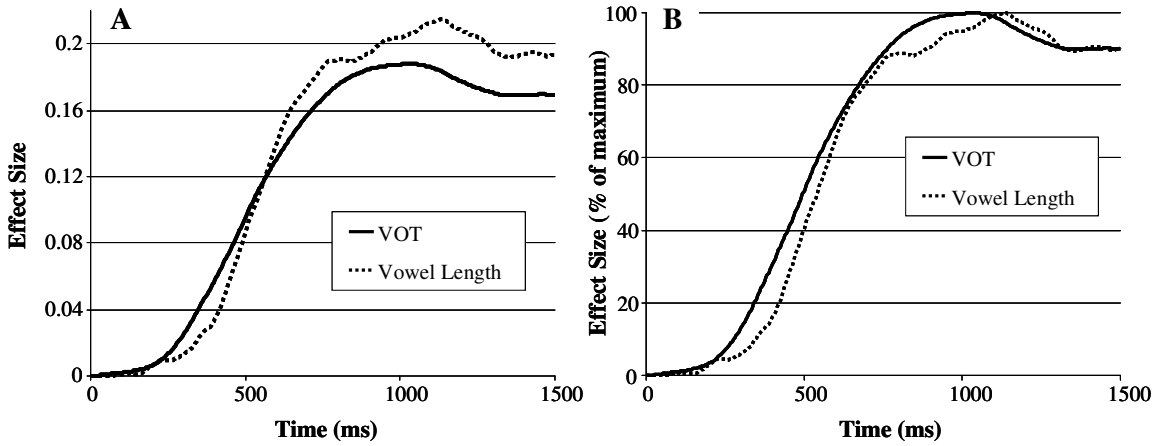

Figure 4: Effect of VOT and vowel length on fixations as a function of time for voicing continua. 0 ms represents the onset of the stimulus (taking into account oculomotor planning delays). A) Raw effect sizes. The effect of VOT is the regression slope relating b/b-bias to VOT at each timestep. Vowel is the difference in b/p-bias between long and short durations. B) Effect sizes are rescaled such that 1 is the maximum size (within an effect), and 0 represents no effect. This is the basis of the Jackknifing analysis in which effects were in terms of the percentage of their maximum values

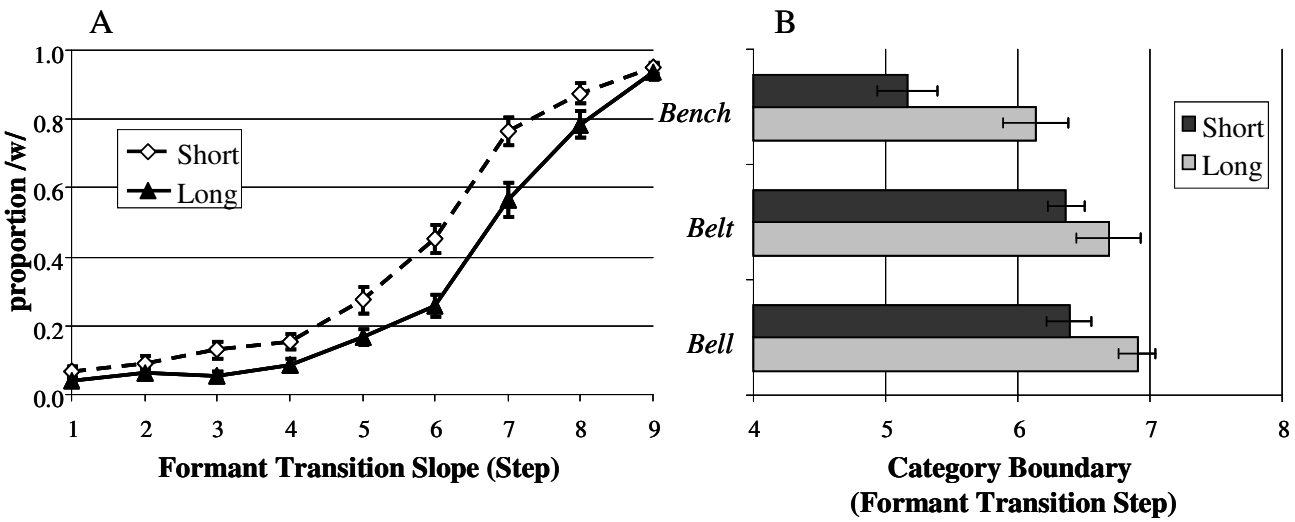

Figure 5: Identification data for Formant Transition continua. A) Proportion /w/ responses as a function of formant transition slope and vowel length averaged across the three voicing continua. B) Computed category boundaries for each continuum as a function of vowel length.

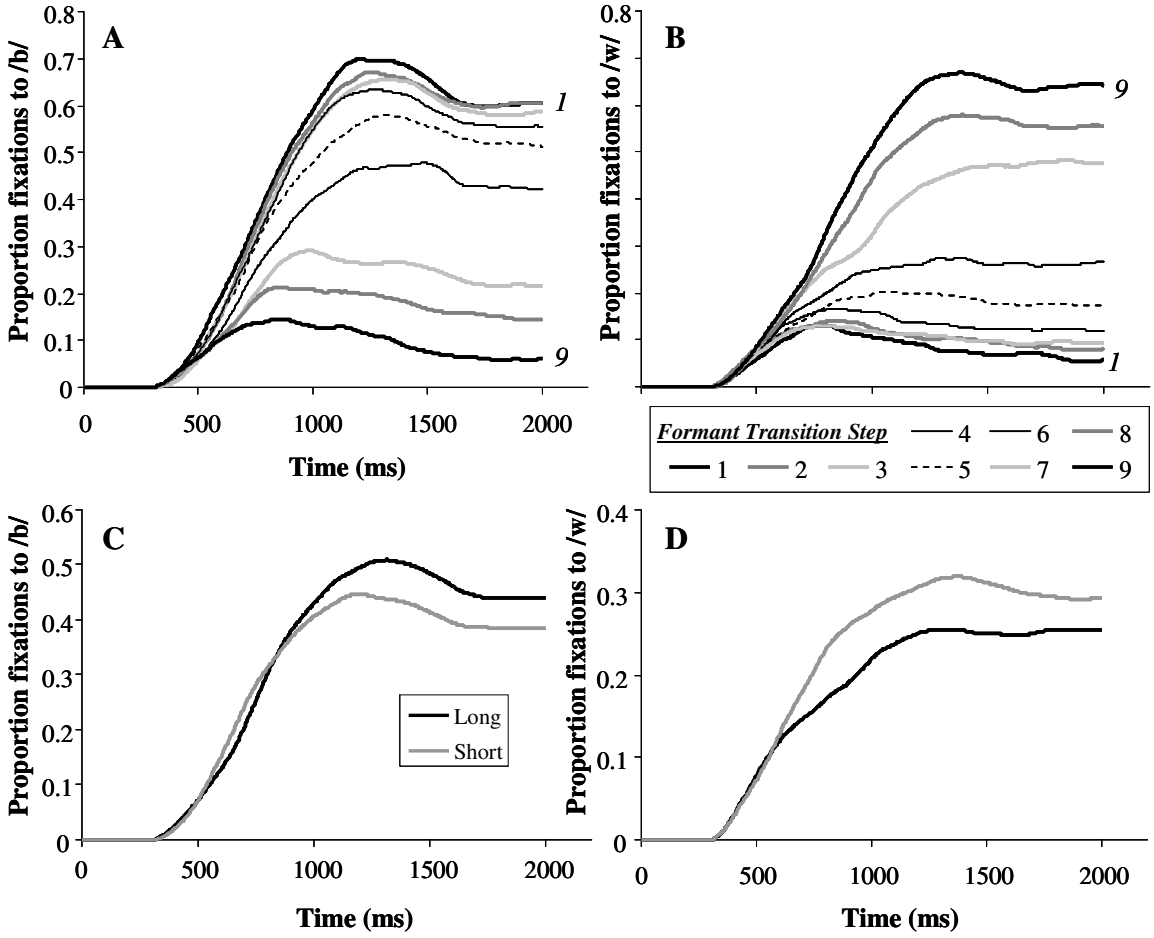

Figure 6: Effect of Formant transition and vowel length on fixations as a function of time. A) Fixations to stop targets as a function of formant transition and time. B) Fixations to approximant targets as a function of formant transition. C) Fixations to stop targets as a function of vowel length and time. D) Fixations to approximant targets as a function of vowel length.

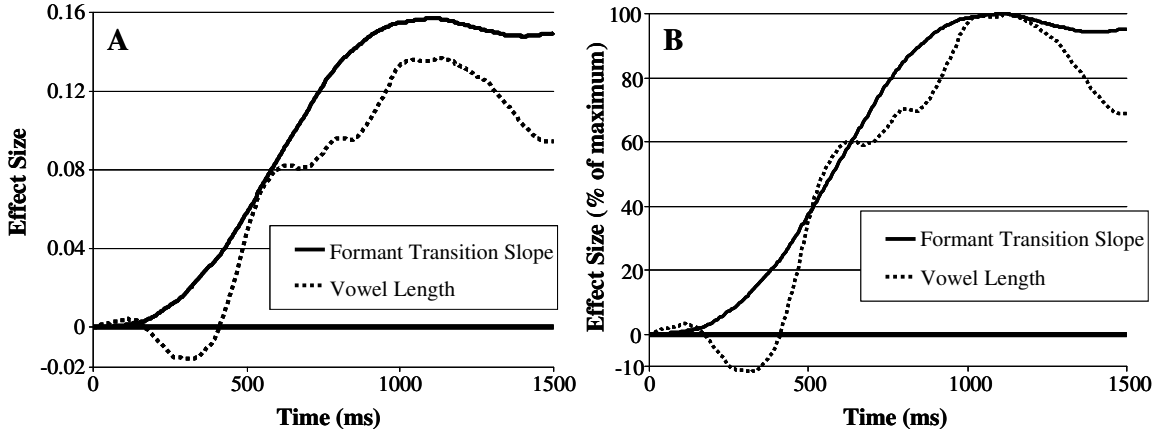

Figure 7: Effect of formant transition slope and vowel length on fixations as a function of time for manner continua. 0 ms represents the onset of the stimulus (taking into account oculomotor planning delays). A) Raw effect sizes. Formant transition effect is the regression slope relating b/w-bias to formant transition slope (in step number) at each timestep. Vowel is the difference in b/w-bias between long and short durations. B) Effect sizes are rescaled such that 1 is the maximum size (within an effect), and 0 represents no effect. This is the basis of the Jackknifing analysis in which effects were in terms of the percentage of their maximum values