

# Temporal integration in vowel perception

Andrew B. Wallace and Sheila E. Blumstein

*Department of Cognitive and Linguistic Sciences, Brown University, Box 1978, Providence, Rhode Island 02912*

(Received 7 January 2008; revised 16 December 2008; accepted 12 January 2009)

Psychoacoustic research suggests that multiple auditory channels process incoming sounds over temporal windows of different durations, resulting in multiple auditory representations being available to higher-level processes. The current experiments investigate the size of the temporal window used in vowel quality perception using an acoustic priming paradigm with nonspeech and speech primes of varying duration. In experiment 1, identification of vowel targets was facilitated by acoustically matched nonspeech primes. The magnitude of this effect was greatest for the shortest (25 and 50 ms) primes, remained level at medium (100 and 150 ms) duration primes, and declined significantly at longer prime durations, suggesting that the auditory stages of vowel quality perception integrate sensory input over a relatively short temporal window. In experiment 2, the same vowel targets were primed by speech stimuli, consisting of vowels using the same duration values as those in experiment 1. A different pattern of results emerged with the greatest priming effects found for primes of around 150 ms and less priming at the shorter and longer durations, indicating that longer-scale temporal processes operate at higher levels of analysis. © 2009 Acoustical Society of America. [DOI: 10.1121/1.3077219]

PACS number(s): 43.71.Es, 43.71.Rt, 43.66.Lj [AJ]

Pages: 1704–1711

## I. INTRODUCTION

Most models of speech perception assume a form of hierarchical processing, in which the acoustic input undergoes a number of transformations as it is mapped onto lexical form (cf. McClelland and Elman, 1986; Blumstein, 1995). The early stages of this process, which appear to involve the extraction of relatively simple auditory features (see Versnel and Shamma, 1998 for one possibility), are likely to play a critical role in speech perception, transforming the sensory input in a way that both shapes and facilitates higher-level phonetic perception. The current study examines one important aspect of this auditory processing: the size of the temporal window of auditory analysis in the early stages of vowel quality perception (i.e., the identification of a vowel's phonetic category, which can be contrasted with the perception of other information encoded in a vowel such as pitch, affect, or prosody).

Temporal integration—the summation of acoustic energy over time—is fundamental to virtually all models of auditory function (cf. Viemeister and Wakefield, 1991). Mathematically, this process is often modeled as a “leaky integrator” whose level of activation is increased by new acoustic stimulation (within a particular frequency band), then decays exponentially. The rate of decay can be characterized by a time constant in the exponential decay function or by the length of the window within which activity decays to a certain percentage of its original value. The size of this temporal window is of considerable interest in psychoacoustics, since it characterizes the temporal resolution of the auditory system.

In attempts to measure the size of this window with behavioral data, different perceptual tasks have yielded vastly different estimates. Studies of gap detection (Plomp, 1964; Penner, 1977; Fitzgibbons, 1983; Shailer and Moore,

1983, 1987), the perception of temporal order (Hirsh, 1959; Pisoni, 1977), the detection of temporal modulation (Viemeister, 1979), and asynchronous masking (Penner and Cudahy, 1973; Moore *et al.*, 1988) have all pointed to a relatively short temporal window of, at most, a few tens of milliseconds for auditory processing. On the other hand, studies of the effect of stimulus duration on hearing thresholds (Plomp and Bouman, 1959; Zwislocki, 1960) or overall loudness (Lifshitz, 1933; Munson, 1947) have pointed to integration over much longer windows of a few hundred milliseconds. Although other explanations have been proposed (cf. Penner, 1978; Viemeister and Wakefield, 1991), this discrepancy has generally been taken to indicate the existence of at least two distinct temporal windows, which comprise separate channels for auditory processing. Poeppel (2003) suggested a short window of 20–40 ms and a long window of 150–250 ms. Incoming sounds are presumably analyzed simultaneously in both channels—different kinds of sounds cannot be directed to different channels, since the sounds can only be identified *after* the initial auditory processing has taken place. The outputs of these channels then map onto higher-level perceptual mechanisms for speech perception, timbre analysis, and so on.

In the case of speech perception, many of the acoustic cues to consonants involve rapid spectral changes, and it is generally assumed that the processing of these spectral changes depends on input from a short-window channel (cf. Zatorre *et al.*, 2002; Poeppel, 2003). Vowels, on the other hand, are encoded by relatively stable spectral patterns. Spectral motion is still important, both in the form of consonantal formant transitions (cf. Strange, 1989) and vowel-intrinsic diphthongization (Nearey and Assmann, 1986; Assmann and Katz, 2000), but these cues are not essential, and synthetic stimuli with steady-state vowels can easily be

perceived as vowels (Delattre *et al.*, 1952). Because of this, some authors suggested that vowel quality perception depends on input from a long temporal window channel (Samson and Zatorre, 1994; Johnsrude *et al.*, 1997; Poeppel, 2003).

The suggestion that vowel quality perception takes input from a long-window auditory channel has not been empirically studied in great depth, but is tentatively supported by neuroimaging results. Since current neuroimaging methods lack the spatial resolution to isolate specific populations of neurons, this research has focused on potential differences between the two hemispheres in the computational mechanisms involved in auditory processing. Left-hemisphere auditory areas have been found to respond best to rapid acoustic changes, while right-hemisphere areas respond best to steady-state or slowly-changing stimuli (Belin *et al.*, 1998; Giraud *et al.*, 2000; Zatorre and Belin, 2001; Poeppel *et al.*, 2004; Zaehle *et al.*, 2004; Boemio *et al.*, 2005; Jamison *et al.*, 2006). Similarly, the response to the rapidly-changing segmental content of speech is left-lateralized, while the response to the slowly-changing prosodic content is right-lateralized (Hesling *et al.*, 2005; Wartenburger *et al.*, 2007). Further evidence for such hemispheric asymmetries in temporal processing comes from electro- and magnetoencephalographic studies showing stimulus-induced changes in rhythmic activity at frequencies related to the suggested short and long temporal integration windows (Luo *et al.*, 2005; Luo and Poeppel, 2007; Giraud *et al.*, 2007).

Poeppel (2003) suggested that these hemispheric differences can be explained by an asymmetric distribution of neurons with long and short temporal windows across the two hemispheres. Although both types of neurons are bilaterally distributed, long-window (150–250 ms) cells are proposed to be more abundant in the right hemisphere, while short-window (20–40 ms) cells are proposed to be more abundant in the left, creating an overall difference in temporal resolution between the two hemispheres. Differences in window length could also affect spectral resolution, since longer temporal windows allow for a more detailed spectral analysis. Thus, Poeppel (2003) proposed that the right-lateralized, long-window channel has a high spectral resolution, while the left-lateralized, short-window channel has a low spectral resolution.

Using Poeppel's (2003) framework, hemispheric asymmetries in auditory processing might offer evidence of the temporal window involved in the perception of vowel quality. A left-hemisphere advantage would suggest a short temporal window, while a right-hemisphere advantage would suggest a long window. Obleser *et al.* (2007) found bilateral activation during the perception of vowels, with what appeared to be a right-hemisphere preference, although this laterality was not tested statistically. In another functional neuroimaging study, Britton *et al.* (2009) studied the processing of vowels and pure tones at durations of 75, 150, and 300 ms. For both vowels and tones at the longer stimulus durations, more activation was found in the right hemisphere than the left. These findings are consistent with suggestions

that vowel quality perception draws its input from a right-lateralized auditory channel with a long temporal integration window.

Behavioral findings, however, seem to contradict the neuroimaging results, suggesting that the perception of vowel quality is based on auditory processing over a short temporal window. Holt *et al.* (2000), applying a method developed by Lotto and Kluender (1998), found a shift in the boundary between 60 ms tokens of the vowels [e] and [ʌ] in a continuum of synthetic stimuli varying only in F2 when these stimuli were flanked by 70 ms pure tones at the onset F2 frequencies of [d] and [b]. The direction of this shift was contrastive, meaning that vowels were less likely to be identified as belonging to the category whose F2 frequencies were closer to those of the flanking tones. The authors used these results to argue that perceptual compensation for coarticulation could be explained by a general auditory contrast mechanism rather than the speech-specific mechanism proposed by Mann (1980). In the context of the current study, however, their findings suggest that the auditory input for vowel quality perception has a high enough temporal resolution to resolve 60–70 ms stimuli, which in turn suggests a relatively short temporal integration window.

Additionally, a series of studies by Chistovich (1985) suggests that the auditory processing of vowels for phonetic perception involves low spectral resolution, a feature that has been linked to the short-window channel (Poeppel, 2003). Previous studies had observed that two formants, in close proximity, combine to form a single perceived spectral peak (Fant, 1956; Carlson *et al.*, 1970), a phenomenon with important implications for vowel quality perception (cf. Stevens, 1972, 1989). Chistovich and Lublinskaja (1979) and Chistovich and Sheikin (1979) showed that the range at which such mergers occurred was around 3–3.5 bark (expressed in the critical-band rate frequency scale of Zwicker (1961) and Zwicker and Terhardt (1980)). This 3–3.5 bark critical distance is large relative to the frequency resolution of the peripheral auditory system. Thus, Chistovich (1985) suggested that the auditory analysis of vowel spectra might involve a relatively wide-band spectral integration, which would result in a lower spectral resolution.

Low spectral resolution may not at first appear to be advantageous for vowel quality perception, since the primary determinant of vowel quality is a spectral cue, the pattern of formant frequencies (Delattre *et al.*, 1952). Formants, however, occur on a relatively large spectral scale. Finer levels of spectral detail provide information mainly about harmonics of the fundamental frequency of phonation ( $F_0$ ). This suggests that the perception of vowel quality can be accomplished at a relatively low spectral resolution. Indeed, increasing spectral resolution to the point where individual harmonics can be resolved might impede the extraction of formants, since the spectral peaks resulting from individual harmonics might be mistaken for formants. Consistent with this is evidence showing that vowel identification becomes less accurate as  $F_0$  increases and individual harmonics become more prominent (Ryalls and Lieberman, 1982; Diehl *et al.*, 1996). Thus, the optimal resolution for vowel quality perception may be found in the low-spectral-, high-temporal-

resolution channel rather than the high-spectral-, low-temporal-resolution channel (but cf. Cheveigné and Kawahara, 1999; Hillenbrand and Houde, 2003).

A number of behavioral methods could be used to experimentally test the temporal resolution of vowel quality perception. Past research has investigated vowel perception by measuring the temporal limits of forward and backward masking. Results have shown backward masking effects of vowel maskers on very short vowel targets, but the maximum temporal intervals for these effects ranged from 80 ms (Pisoni, 1972) to 250 ms (Massaro, 1972). Moreover, these studies found backward masking effects on vowel perception, but not forward masking, and there were large individual differences, all of which suggests that these effects were caused by higher-level interference effects on decision processes rather than low-level auditory integration effects (LaRivière *et al.*, 1975; Dorman *et al.*, 1977).

Another approach would be to test the accuracy and latency with which subjects can identify gated vowels of varying duration. The results of gating studies, however, are hard to interpret in the context of basic auditory processes. Because duration is itself a secondary cue to vowel quality (cf. Joos, 1948; Peterson and Lehiste, 1960), gating results would be influenced by higher-level processes, which map auditory information onto linguistic representations in order for the subject to ultimately make a phonetic category decision. As a consequence, higher-level linguistic factors could influence the perceptual results. Ultimately, studies that rely solely on the manipulation of speech stimuli will always be influenced by the effects of such manipulations on higher-level processing, making it difficult to isolate low-level auditory processing mechanisms.

An alternative approach uses the effects of nonspeech sounds on speech perception. Lotto and Kluender (1998) pioneered these methods, showing that preceding tones shifted the boundary between [da] and [ba] in a continuum of stimuli. Because the nonspeech tones do not activate higher-level representations strongly enough to create a percept of speech, the influence of such stimuli on perception is more likely to reflect early auditory stages of processing. On the other hand, because the nonspeech effect is measured in relation to subjects' responses to a speech target, it taps those auditory channels which contribute to phonetic perception. Thus, examining the effects of nonspeech sounds on speech perception provides a means of investigating the contribution of low-level auditory processing to the perception of speech.

The current study uses a paradigm similar to that employed by Lotto and Kluender (1998) and Holt *et al.* (2000), exploring the effects of nonspeech tones on vowel identification (Wallace and Blumstein, 2006). Stimuli consist of the vowel targets [i] and [ɑ], preceded by nonspeech primes composed of two sinusoidal components matched to the first two formant frequencies of one or the other of the targets. Initial experiments (Wallace and Blumstein, 2006) showed that subjects respond faster to the vowels in matched prime-target pairs than in mismatched pairs. This priming effect differs from Lotto and Kluender's (1998) method in one important way. In Lotto and Kluender's (1998) study, the effect of the nonspeech tone stimuli on the speech stimuli was

found primarily in the boundary region of a continuum of stimuli, where tokens were identified at or near chance for either phonetic category. The current study, in contrast, measures behavioral effects of nonspeech stimuli on good exemplars of phonetic categories. In this way, it is possible to investigate vowel identification independent of the processes involved in resolving phonetic category membership in ambiguous stimuli (though such processes are important in the perception of underspecified, hypoarticulated natural speech).

In the current study, the duration of nonspeech prime tones will be varied parametrically in order to test the size of the temporal window that contributes to vowel quality perception. Prime durations will be 25, 50, 100, 150, 200, 300, and 500 ms, a set that spans the proposed short and long window durations. Because priming effects depend on similarity between prime and target, response times (RTs) are expected to be faster for "match" than for "clash" trials at each prime duration. Prime duration is likely to affect the magnitude of this priming effect, since primes that are shorter than the temporal integration window produce a weaker auditory response than primes with durations equal to or greater than the temporal integration window. This difference in activation is likely to modulate the effect of similarity between primes and targets. Thus, whatever the duration of the window is, increases in prime duration up to this value should increase the magnitude of the priming effect, while increasing the prime duration beyond the length of the window should produce no further increases in priming.

If, as Poeppel (2003) suggested, vowels are analyzed over a long temporal window of 150–250 ms, the shortest primes of the current experiment will be too brief to fully excite that window. These short primes should, therefore, produce less activation and a weaker priming effect than longer primes. Maximum priming should therefore occur at prime durations of 150–250 ms, at which point the prime stimuli will be as long as the temporal integration window. If, on the other hand, vowel perception draws on a short window, then short duration primes should fully excite the auditory representation of the vowel. Maximum priming should then occur at prime durations of 25–50 ms. Long and short temporal integration windows thus make different predictions for the direction of the effect of prime duration on priming magnitude between 20 and 250 ms (i.e., in the 25, 50, 100, 150, and 200 ms conditions of the current experiment). A long window predicts increasing priming over this range, while a short window predicts level or decreasing priming.

## II. EXPERIMENT 1

### A. Methods

#### 1. Subjects

Subjects were 70 native speakers of North American English who reported having normal hearing and no history of speech or language disorders. They were paid for their participation.

## 2. Stimuli

Targets were two vowel sounds, [a] and [i], synthesized using the PRAAT software package (Boersma and Weenink, 2007). The linear predictive coding residual of a single token of [i] spoken in isolation by a male native speaker of North American English was passed through a five-formant cascade to create each target vowel.  $F_0$  ranged from 123 to 131 Hz and the stimulus duration was 280 ms. The first three formants were set to constant frequencies unique to each target vowel; F1 and F2 frequencies were determined by measuring the speaker's productions, while F3 was set to a value that produced the best vowel percept as judged by the experimenter (A. Wallace) ([i]: 362, 2191, and 3100 Hz; [a]: 772, 1308, and 2960 Hz). The fourth and fifth formants, as well as all formants' bandwidths, were not altered from the time-varying values measured in the original natural [i] (mean  $F_4=3487$  Hz,  $F_5=3997$  Hz; BW=82, 39, 261, 204, 238 Hz). The targets' amplitudes were scaled so that they would be equal in perceptual loudness, as predicted by the ISO532B standard [71 dB sound pressure level (SPL) for both targets].

Primes were two nonspeech tone complexes, one matched to each of the targets. Each prime contained two component tones at the frequencies of the first two formants of the matched target. Intensities of the individual components were controlled to ensure: first, that the two sinusoidal components of each prime were equally loud, based on ISO226 equal loudness contours; and second, that the two primes themselves were equally loud, as predicted by ISO532B, resulting in absolute levels of 79 dB SPL for primes matched to [i] and 83 dB SPL for primes matched to [a]. Primes were 25, 50, 100, 150, 200, 300, and 500 ms in duration, including 15 ms cosine-squared onset and offset intensity ramps, except in the 25-ms condition, where 12.5 ms ramps were used. After completing the experiment, subjects were asked what the primes sounded like, and none reported having heard them as speech. Subjects were then asked specifically whether the primes sounded like the target vowels; 3 of 70 subjects reported that they did.

## 3. Procedure

Subjects were assigned to one of the seven prime duration conditions with ten subjects participating in each condition. Each trial consisted of a prime and a target, separated by a 50 ms inter-stimulus interval. RTs were measured from

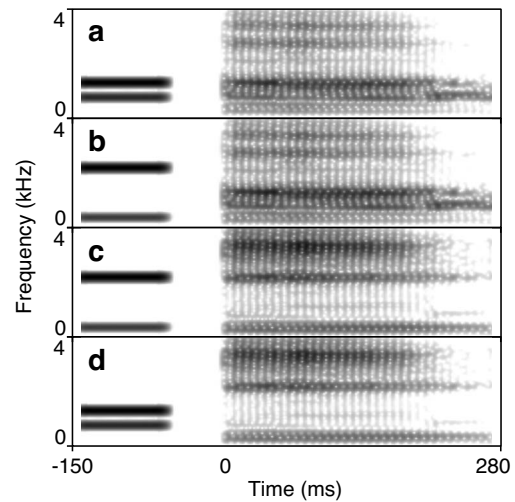


FIG. 1. Spectrograms illustrating four trial types in 100 ms prime condition: (a) [a] target, match prime; (b) [a] target, clash prime; (c) [i] target, match prime; and (d) [i] target, clash prime.

the onset of the target vowel. Each trial began 750 ms after the onset of the target of the previous trial, unless a response had not yet been recorded, in which case subjects were given up to 3 s to respond. In all, the two targets and two primes combined to form four unique trials (two match conditions and two clash conditions, see Fig. 1). Each of these four was repeated 250 times in pseudorandom order for each subject, while prime duration varied between subjects. Stimuli were presented binaurally over Sony MDR-V6 headphones and subjects responded by pressing one of two buttons labeled with the target vowels in both the international phonetic alphabet and a colloquial phonetic notation ([a]/"ah" and [i]/"ee"). Both performance and reaction-time measures were taken.

## B. Results

Mean RTs were calculated for each subject for all trials of a given type, excluding incorrect responses, responses that were made before the target onset, and trials in which the subject did not respond in the given 3000 ms (1.23% of trials). As shown in Table I, subjects' responses to match trials were both faster and more accurate than their responses to clash trials. Because of the high variability in both error rates (which were low overall) and absolute RTs (which varied greatly between subjects), statistical analyses relied on a

TABLE I. Cross-subject mean error rates and response times for experiment 1.

Prime duration (ms)	Error rate (per 1000 trials)			Response time (ms)		
	Match	Clash	Difference	Match	Clash	Difference
25	8.8	19	10	434	468	34
50	9.8	21	11	468	494	26
100	7.6	16	8.0	451	481	29
150	21	30	9.0	416	441	25
200	5.2	8.4	3.2	443	460	17
300	4.6	9.6	5.0	412	424	11
500	2.8	5.0	2.2	465	471	6

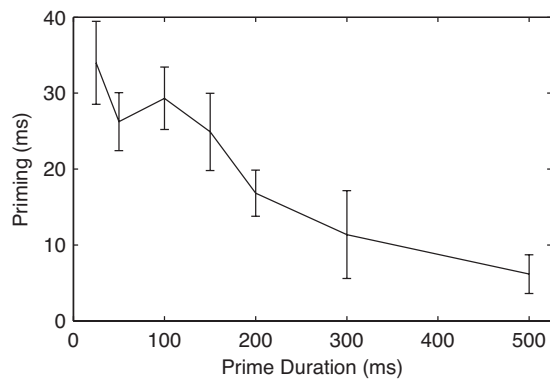


FIG. 2. Results of experiment 1. The magnitude of priming—RT for clash trials minus RT for match trials—is shown as a function of prime duration. Error bars indicate 1 standard error for the cross-subject mean.

single RT priming score, computed independently for each subject as the difference between mean RTs for match and clash trials. A one-sample *t*-test, with results pooled across all conditions, confirmed that priming scores were significantly greater than zero [ $t_{(69)}=11.3$ ,  $p<0.001$ ]. These results confirm the existence of the priming effect of nonspeech tones in a vowel-identification task.

As Fig. 2 shows, the magnitude of this priming effect varied across prime durations, with a greater magnitude of priming at shorter prime durations. A one-way analysis of variance (ANOVA) found a significant effect of prime duration on the magnitude of the priming effect [ $F_{(6,63)}=5.81$ ,  $p<0.001$ ]. Polynomial trend analysis revealed a significant linear trend [ $F_{(1,63)}=31.2$ ,  $p<0.001$ ] but no significant higher-order trends [quadratic  $F_{(6,63)}=1.31$ ,  $p=0.25$ ; cubic  $F_{(6,63)}=0.41$ ,  $p=0.52$ ; fourth-order  $F_{(6,63)}<0.001$ ,  $p=1.0$ ; fifth-order  $F_{(6,63)}=1.74$ ,  $p=0.19$ ]. To further explore the nature of this trend, contrasts were tested between short (25 and 50 ms), medium (100 and 150 ms), and long (200 and 300 ms) durations, excluding the 500 ms condition. Short and medium prime durations did not differ significantly from each other [ $F_{(1,63)}=0.52$ ,  $p=0.47$ ], but long prime durations produced significantly less priming than either short [ $F=15$ ,  $p<0.001$ ] or medium ( $F=9.7$ ,  $p=0.003$ ) durations. As discussed in the Introduction, a long temporal integration window would predict an increase in priming up to durations equal to that of the window, around 150–250 ms. The fact that such an increase did not occur and that maximal priming occurs at the shorter durations supports a short temporal integration window for vowel quality perception.

### C. Discussion

The results of experiment 1 indicate that priming magnitudes are highest at the shortest prime durations, remain level at durations of 100–150 ms, and decline at longer durations. The fact that the short duration primes did not produce a smaller priming effect than the medium duration primes is consistent with a short temporal integration window for the processing of vowels, but not a long one. As discussed in the Introduction, medium to long primes should excite a long window better than short primes. This should result in greater magnitudes of priming for either medium or

long primes than for short primes, a pattern which was not found in experiment 1. Thus, the current results indicate that the perception of vowel quality draws on an auditory processing channel with a short temporal integration window of 20–50 ms.

Finally, it is of interest that priming magnitudes actually declined at the longer prime durations. Standard auditory models (cf. Viemeister and Wakefield, 1991) predict that auditory activation (and thus, presumably, the magnitude of the priming effect) would plateau at durations longer than that of the relevant temporal integration window. A few factors that are not incorporated into the basic auditory models might explain the obtained decline in the magnitude of priming. For example, many parts of the auditory system have been found to respond strongly to stimulus onsets (Furukawa and Ishii, 1967; Smith, 1977; Rhode and Smith, 1985; Pfeiffer, 1966; Rhode *et al.*, 1983). Such an onset response might produce a decline in priming at longer prime durations, since increasing duration would increase the temporal separation between prime and target onsets, allowing more time for the auditory response to decay. In addition, it is of interest that no significant difference was found between short (25 and 50 ms) and medium (100 and 150 ms) prime durations, suggesting a delay before priming magnitudes begin to decline. Further research is needed to explore this phenomenon.

## III. EXPERIMENT 2

In experiment 1, it was shown that the priming effect produced by nonspeech tone complexes matched to a target vowel's formant frequencies is greatest at the shortest prime durations, suggesting that the spectral properties of vowels are extracted over a relatively short temporal window. To determine whether the effect of prime duration would be the same when primes engage higher levels of phonetic processing along with auditory levels, experiment 2 replicated the duration conditions of experiment 1 using matched or mismatched vowel primes in place of nonspeech tones.

### A. Methods

#### 1. Subjects

Subjects were 70 native speakers of North American English, who met the same criteria as the subjects who participated in experiment 1 but who had not taken part in experiment 1.

#### 2. Stimuli

Targets were the same tokens of [i] and [a] used in experiment 1. Primes were vowel stimuli, created by altering the duration of the vowel targets to 25, 50, 100, 150, 200, 300, and 500 ms using the PRAAT software package's (Boersma and Weenink, 2007) implementation of Moulines and Laroche's (1995) pitch-synchronous overlap-and-add algorithm.

#### 3. Procedure

All aspects of the experimental procedure, with the exception of the different primes, were the same as in experi-

TABLE II. Cross-subject mean error rates and response times for experiment 1.

Prime duration (ms)	Error rate (per 1000 trials)			Response time (ms)		
	Match	Clash	Difference	Match	Clash	Difference
25	2.2	11	9.2	446	503	57
50	5.4	21	16	456	536	80
100	9.0	22	13	504	598	94
150	3.4	18	14	480	582	102
200	4.2	11	6.6	476	550	74
300	6.2	13	6.6	417	461	44
500	7.8	11	3.2	507	534	27

ment 1. Subjects were told that vowels would come in pairs, and that they should ignore the first vowel in each pair and only respond to the second by pressing one of two buttons labeled with the target vowels. Both performance and reaction-time measures were taken. Prime duration varied between subjects, with ten subjects participating in each of the seven duration conditions. Subjects completed 250 match and 250 clash trials for each of the two target vowels.

## B. Results

Table II lists mean error rates and RT, for match and clash trials at each prime duration of experiment 2. As in experiment 1, mean RTs were calculated for each subject for all trials of a given type, excluding trials for the same reasons as in experiment 1 (1.05% of trials). The subject's mean reaction time for match trials was then subtracted from the mean for clash trials to produce a single difference value, the magnitude of priming for that subject. Figure 3 shows the magnitude of priming as a function of prime duration for the vowel prime stimuli of experiment 2 as well as the nonspeech primes of experiment 1. A single-factor ANOVA revealed a significant effect of prime duration on priming magnitude [ $F_{(6,63)}=3.45$ ,  $p=0.005$ ]. Polynomial trend analysis of the speech-prime conditions found significant or nearly-significant linear [ $F_{(1,63)}=10.0$ ,  $p=0.002$ ], quadratic [ $F_{(1,63)}=2.81$ ,  $p=0.098$ ], and cubic trends [ $F_{(1,63)}=7.02$ ,  $p=0.010$ ], without significant fourth-order [ $F_{(1,63)}=0.312$ ,  $p=0.579$ ] or

fifth-order [ $F_{(1,63)}=0.13$ ,  $p=0.133$ ] trends. As in experiment 1, the nature of this trend was further explored by testing contrasts between short (25 and 50 ms), medium (100 and 150 ms), and long (200 and 300 ms) prime durations. In experiment 1, short primes produced as large an effect as medium, and a greater effect than long primes. In contrast, the medium-duration primes of experiment 2 produced a larger effect than short [ $F_{(1,63)}=4.2$ ,  $p=0.045$ ] or long [ $F_{(1,63)}=7.2$ ,  $p=0.009$ ] primes, while short and long primes did not differ significantly [ $F_{(1,63)}=0.41$ ,  $p=0.52$ ].

Finally, the results of experiments 1 and 2 were directly compared in a two-factor (prime duration  $\times$  prime type) ANOVA, using the results of experiment 1 as a "nonspeech" prime type condition and the results of experiment 2 as a speech prime type condition. Results indicated significant main effects of prime duration [ $F_{(6,126)}=5.18$ ,  $p<0.001$ ] and prime type [speech vs nonspeech;  $F_{(1,126)}=67.4$ ,  $p<0.001$ ], and an interaction that approached significance [ $F_{(6,126)}=2.07$ ,  $p=0.061$ ]. This interaction was consistent with the different patterns observed in experiments 1 and 2.

## C. Discussion

Experiment 2 found a significantly greater priming effect for speech primes than for nonspeech. This is neither surprising nor particularly interesting given that the primes of experiment 2 were acoustically closer to the targets than were the nonspeech primes of experiment 1, and matched the targets phonetically as well. Of more interest is the effect of prime duration on the speech and nonspeech prime conditions. While experiment 1 found the greatest priming at the shortest prime durations, experiment 2 found priming to be greatest with the medium-duration primes of 100 and 150 ms. These two patterns likely reflect a difference in the processing mechanisms for the two kinds of stimuli. Experiment 1, using nonspeech primes, taps primarily auditory levels of processing, whereas experiment 2, using vowel primes, taps both auditory and phonetic levels. Thus, the data from experiment 2 suggest that phonetic processes operate over a timescale of approximately 150 ms, consistent with the interpretation of backward masking results by [Dorman et al. \(1977\)](#) ([Massaro, 1972](#); [Pisoni, 1972](#)) in the perception of vowels as reflecting higher-level processing occurring over 100–200 ms.

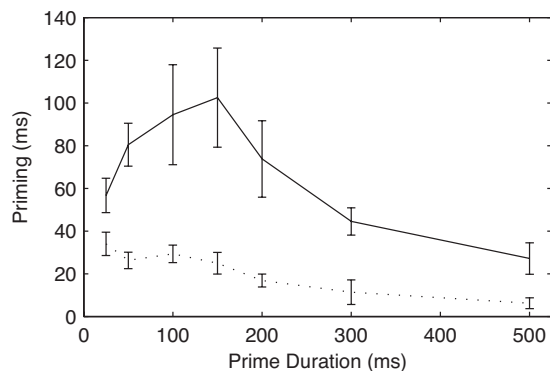


FIG. 3. Results of experiment 2. Solid line shows the magnitude of priming by prime duration for vowel primes and vowel targets. Dotted line shows magnitude of priming by duration for nonspeech primes and vowel targets (data from experiment 1). Error bars indicate 1 standard error for cross-subject mean.

## IV. GENERAL DISCUSSION

The current experiments explored the temporal processing of vowels by manipulating the duration of both non-speech (experiment 1) and speech (experiment 2) primes, which preceded vowel targets in a vowel-identification task. In both experiments, matched primes facilitated the identification of acoustically similar targets. In experiment 1, the magnitude of priming showed a linear decline across the duration intervals; priming magnitudes were highest at the shortest prime durations, remained level at durations of 100–150 ms, and declined at longer durations. In experiment 2, priming magnitudes showed an initial increase with the greatest priming at medium durations of 100–150 ms followed by a subsequent decline. These results support a model in which vowels undergo an initial auditory analysis within a short temporal window. The output of this analysis is then mapped onto phonetic perceptual mechanisms where longer-scale temporal dynamics play a role.

At the auditory level, the current results are compatible with a range of window durations, including Poeppel's (2003) proposal of 20–40 ms, but they are not compatible with the exclusive involvement of a medium or long window of analysis for vowel quality perception. Instead, the current studies suggest that vowel quality perception draws on the output of a high-temporal-resolution, low-spectral-resolution auditory processing channel. It is possible that the temporal window for the auditory analysis of vowels may be shorter than 25 ms and possibly shorter than 10 ms as proposed by Viemeister (1979); however, given that primes of this duration were not used this hypothesis would need to be tested directly.

Determining the temporal and spectral parameters of this auditory channel will help explain the vowel perception process as a whole. For example, models of vowel perception have tended to avoid direct claims about auditory processing. Syrdal and Gopal (1986), for example, described a model of vowel identification from formant frequencies but did not specify how these frequencies are obtained [they used Peterson and Barney's (1952) measurements]. Molis (2005) contrasted a number of models of vowel perception. Although some of these incorporate knowledge of peripheral auditory processing (i.e., critical-band filtering), none includes the possibility that central auditory processing might introduce different channels with different spectral and temporal resolutions.

The current study has proceeded from the assumption of multiple temporal windows, but there are other potential explanations for the differences that have been found in the temporal resolution of different psychoacoustic tasks. One of these Viemeister and Wakefield's (1991) "multiple looks" model (cf. also Moore, 2003, 2008) proposes that incoming sounds are analyzed within a single, short temporal window, but that specific sounds are identified on the basis of templates, which can span several windows (i.e., multiple looks). Hillenbrand and Houde's (2003) model of vowel perception takes a similar approach in the frequency domain, arguing that spectral smoothing to obscure a vowel's harmonics is not necessary, because a high-spectral-resolution representa-

tion of the vowel can instead be compared directly to a smooth template. Experimentally distinguishing between a "multiple windows" model and a "multiple looks" model is rather difficult, and the results of the current study are compatible with both models. Within a multiple looks framework, the current study suggests that steady-state vowels are identified on the basis of short, low-spectral-resolution templates.

Much remains to be learned about the temporal and spectral parameters of the auditory processing of vowels, as well as other speech sounds. Future studies can explore these areas using the basic methodological framework presented in the current paper. Perceptual interactions between speech and nonspeech sounds, because they are likely to occur at relatively early auditory stages of the speech-processing stream, provide a unique window into this level of processing. By illuminating such mechanisms, this line of research will also shed light on the building blocks of higher-level processes.

## ACKNOWLEDGMENTS

This work was supported by NIH Grant Nos. DC000314 and DC006220 to Brown University.

- Assmann, P., and Katz, W. (2000). "Time-varying spectral change in the vowels of children and adults," *J. Acoust. Soc. Am.* **108**, 1856–1866.
- Belin, P., Zilbovicius, M., Crozier, S., Thivard, L., Fontaine, A., Masure, M.-C., and Samson, Y. (1998). "Lateralization of speech and auditory temporal processing," *J. Cogn Neurosci.* **10**, 536–540.
- Blumstein, S. E. (1995). "The neurobiology of the sound structure of language," in *The Cognitive Neurosciences*, edited by M. Gazzaniga (MIT, Cambridge, MA), 915–929.
- Boemio, A., Fromm, S., Braun, A., and Poeppel, D. (2005). "Hierarchical and asymmetric temporal sensitivity in human auditory cortices," *Nat. Neurosci.* **8**, 389–396.
- Boersma, P., and Weenink, D. (2007). "Praat: Doing phonetics by computer, version 4.5.16," computer program, <http://www.praat.org/> (last viewed February, 2007).
- Britton, B., Blumstein, S. E., Myers, E. B., and Grindrod, C. (2009). "The role of spectral and durational properties on hemispheric asymmetries in vowel perception," *Neuropsychologia* (in press).
- Carlson, R., Granström, B., and Fant, G. (1970). "Some studies concerning perception of isolated vowels," Speech Transmission Laboratory Quarterly Progress Report No. 2-3, KTH, Stockholm.
- Cheveigné, A., and Kawahara, H. (1999). "Missing-data model of vowel identification," *J. Acoust. Soc. Am.* **105**, 3497–3508.
- Chistovich, L. (1985). "Central auditory processing of peripheral vowel spectra," *J. Acoust. Soc. Am.* **77**, 789–805.
- Chistovich, L., and Lublinskaya, V. (1979). "The 'center of gravity' effect in vowel spectra and critical distance between the formants: Psychoacoustical study of the perception of vowel-like stimuli," *Hear. Res.* **1**, 185–195.
- Chistovich, L. A., and Sheikin, R. L. (1979). "Relevant parameters of steady-state vowel spectra," in *Sensornye Systemy*, edited by G. V. Gershuni (Nauka, Leningrad), pp. 116–129 (in Russian).
- Delattre, P. C., Liberman, A. M., Cooper, F. S., and Gerstman, L. J. (1952). "An experimental study of the acoustic determinants of vowel colour; observations on one- and two-formant vowels synthesized from spectrographic patterns," *Word* **8**, 195–210.
- Diehl, R., Lindblom, B., Hoemeke, K., and Fahey, R. (1996). "On explaining certain male-female differences in the phonetic realization of vowel categories," *J. Phonetics* **24**, 187–208.
- Dorman, M., Kewley-Port, D., Brady, S., and Turvey, M. (1977). "Vowel recognition: Inferences from studies of forward and backward masking," *Q. J. Exp. Psychol.* **29**, 483–497.
- Fant, G. (1956). "On the predictability of formant levels and spectrum envelopes from formant frequencies," in *For Roman Jakobson*, edited by M. Halle, H. Lunt, and H. Maclean (Mouton, The Hague), pp. 109–120.
- Fitzgibbons, P. (1983). "Temporal gap detection in noise as a function of frequency, bandwidth, and level," *J. Acoust. Soc. Am.* **74**, 67–72.

- Furukawa, T., and Ishii, Y. (1967). "Neurophysiological studies of hearing in goldfish," *J. Neurophysiol.* **30**, 1377–1403.
- Giraud, A.-L., Kleinschmidt, A., Poeppel, D., Lund, T. E., Frackowiak, R. S. J., and Laufs, H. (2007). "Endogenous cortical rhythms determine cerebral specialization for speech perception and production," *Neuron* **56**, 1127–1134.
- Giraud, A.-L., Lorenzi, C., Ashburner, J., Wable, J., Johnsrude, I., Frackowiak, R., and Kleinschmidt, A. (2000). "Representation of the temporal envelope of sounds in the human brain," *J. Neurophysiol.* **84**, 1588–1598.
- Hesling, I., Dilharreguy, B., Clement, S., Bordesoules, M., and Allard, M. (2005). "Cerebral mechanisms of prosodic sensory integration using low-frequency bands of connected speech," *Hum. Brain Mapp* **26**, 157–169.
- Hillenbrand, J., and Houde, R. (2003). "A narrow band pattern-matching model of vowel perception," *J. Acoust. Soc. Am.* **113**, 1044–1065.
- Hirsh, I. J. (1959). "Auditory perception of temporal order," *J. Acoust. Soc. Am.* **31**, 759–767.
- Holt, L. L., Lotto, A. J., and Kluender, K. R. (2000). "Neighboring spectral content influences vowel identification," *J. Acoust. Soc. Am.* **108**, 710–722.
- Jamison, H., Watkins, K., Bishop, D., and Matthews, P. (2006). "Hemispheric specialization for processing auditory nonspeech stimuli," *Cereb. Cortex* **16**, 1266–1275.
- Johnsrude, I., Zatorre, R., Milner, B., and Evans, A. (1997). "Left-hemisphere specialization for the processing of acoustic transients," *NeuroReport* **8**, 1761–1765.
- Joos, M. (1948). "Acoustic phonetics," *Language* **24**, 5–136.
- LaRivière, C., Winitz, H., and Herriman, E. (1975). "Vocalic transitions in the perception of voiceless initial stops," *J. Acoust. Soc. Am.* **57**, 470–475.
- Lifshitz, S. (1933). "Two integral laws of sound perception relating loudness and apparent duration of sound impulses," *J. Acoust. Soc. Am.* **5**, 31–33.
- Lotto, A. J., and Kluender, K. R. (1998). "General contrast effects in speech perception: Effect of preceding liquid on stop consonant identification," *Percept. Psychophys.* **60**, 602–619.
- Luo, H., and Poeppel, D. (2007). "Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex," *Neuron* **54**, 1001–1010.
- Luo, H., Husain, F. T., Horwitz, B., and Poeppel, D. (2005). "Discrimination and categorization of speech and non-speech sounds in an MEG delayed-match-to-sample study," *Neuroimage* **28**, 59–71.
- Massaro, D. (1972). "Preperceptual images, processing time, and perceptual units in auditory perception," *Psychol. Rev.* **79**, 124–145.
- Mann, V. A. (1980). "Influence of preceding liquid on stop-consonant perception," *Percept. Psychophys.* **28**, 407–412.
- McClelland, J. L., and Elman, J. L. (1986). "The trace model of speech perception," *Cogn. Psychol.* **18**, 1–86.
- Molis, M. R. (2005). "Evaluating models of vowel perception," *J. Acoust. Soc. Am.* **118**, 1062–1071.
- Moore, B. (2003). "Temporal integration and context effects in hearing," *J. Phonetics* **31**, 563–574.
- Moore, B. C. J. (2008). "Basic auditory processes involved in the analysis of speech sounds," *Philos. Trans. R. Soc. London, Ser. B* **363**, 947–963.
- Moore, B., Glasberg, B., Plack, C., and Biswas, A. (1988). "The shape of the ear's temporal window," *J. Acoust. Soc. Am.* **83**, 1102–1116.
- Moulines, E., and Laroche, J. (1995). "Non-parametric techniques for pitch-scale and time-scale modification of speech," *Speech Commun.* **16**, 175–205.
- Munson, W. (1947). "The growth of auditory sensation," *J. Acoust. Soc. Am.* **19**, 734–735.
- Nearey, T., and Assmann, P. (1986). "Modeling the role of inherent spectral change in vowel identification," *J. Acoust. Soc. Am.* **80**, 1297–1308.
- Obleser, J., Zimmermann, J., Van Meter, J., and Rauschecker, J. (2007). "Multiple stages of auditory speech perception reflected in event-related fMRI," *Cereb. Cortex* **17**, 2251–2257.
- Penner, M. (1977). "Detection of temporal gaps in noise as a measure of the decay of auditory sensation," *J. Acoust. Soc. Am.* **61**, 552–557.
- Penner, M. (1978). "A power law transformation resulting in a class of short-term integrators that produce time-intensity trades for noise bursts," *J. Acoust. Soc. Am.* **63**, 195–202.
- Penner, M., and Cudahy, E. (1973). "Critical masking interval: A temporal analog of the critical band," *J. Acoust. Soc. Am.* **54**, 1530–1534.
- Peterson, G. E., and Barney, H. L. (1952). "Control methods used in a study of vowels," *J. Acoust. Soc. Am.* **24**, 175–184.
- Peterson, G., and Lehiste, I. (1960). "Duration of syllable nuclei in English," *J. Acoust. Soc. Am.* **32**, 693–703.
- Pfeiffer, R. R. (1966). "Classification of response patterns of spike discharges for units in the cochlear nucleus: Tone-burst stimulation," *Exp. Brain Res.* **1**, 220–235.
- Pisoni, D. B. (1972). "Perceptual processing time for consonants and vowels," *Haskins Laboratories Status Report on Speech Research No. SR-31/32*, Haskins Laboratories, New Haven, CT.
- Pisoni, D. (1977). "Identification and discrimination of the relative onset time of two component tones: Implications for voicing perception in stops," *J. Acoust. Soc. Am.* **61**, 1352–1361.
- Plomp, R. (1964). "Rate of decay of auditory sensation," *J. Acoust. Soc. Am.* **36**, 277–282.
- Plomp, R., and Bouman, M. (1959). "Relation between hearing threshold and duration for tone pulses," *J. Acoust. Soc. Am.* **31**, 749–758.
- Poeppel, D. (2003). "The analysis of speech in different temporal integration windows: Cerebral lateralization as 'asymmetric sampling in time'," *Speech Commun.* **41**, 245–255.
- Poeppel, D., Guillemin, A., Thompson, J., Fritz, J., Bavelier, D., and Braun, A. R. (2004). "Auditory lexical decision, categorical perception, and FM direction discrimination differentially engage left and right auditory cortex," *Neuropsychologia* **42**, 183–200.
- Rhode, W., and Smith, P. (1985). "Characteristics of tone-pip response patterns in relationship to spontaneous rate in cat auditory nerve fibers," *Hear. Res.* **18**, 159–168.
- Rhode, W. S., Smith, P. H., and Oertel, D. (1983). "Physiological response properties of cells labelled intracellularly with horseradish peroxidase in cat dorsal cochlear nucleus," *J. Comp. Neurol.* **213**, 426–447.
- Ryalls, J. H., and Lieberman, P. (1982). "Fundamental frequency and vowel perception," *J. Acoust. Soc. Am.* **72**, 1631–1634.
- Samson, S., and Zatorre, R. (1994). "Contribution of the right temporal lobe to musical timbre discrimination," *Neuropsychologia* **32**, 231–240.
- Shailer, M., and Moore, B. (1983). "Gap detection as a function of frequency, bandwidth, and level," *J. Acoust. Soc. Am.* **74**, 467–473.
- Shailer, M., and Moore, B. (1987). "Gap detection and the auditory filter: Phase effects using sinusoidal stimuli," *J. Acoust. Soc. Am.* **81**, 1110–1117.
- Smith, R. (1977). "Short-term adaptation in single auditory nerve fibers: Some poststimulatory effects," *J. Neurophysiol.* **40**, 1098–1111.
- Stevens, K. (1972). "The quantal nature of speech: Evidence from articulatory-acoustic data," in *Human Communication, A Unified View*, edited by E. E. David, and P. B. Denes (McGraw-Hill, New York), pp. 51–66.
- Stevens, K. (1989). "On the quantal nature of speech," *J. Phonetics* **17**, 3–45.
- Strange, W. (1989). "Dynamic specification of coarticulated vowels spoken in sentence context," *J. Acoust. Soc. Am.* **85**, 2135–2153.
- Syrdal, A., and Gopal, H. (1986). "A perceptual model of vowel recognition based on the auditory representation of American English vowels," *J. Acoust. Soc. Am.* **79**, 1086–1100.
- Versnel, H., and Shamma, S. A. (1998). "Spectral-ripple representation of steady-state vowels in primary auditory cortex," *J. Acoust. Soc. Am.* **103**, 2502–2514.
- Viemeister, N. (1979). "Temporal modulation transfer functions based upon modulation thresholds," *J. Acoust. Soc. Am.* **66**, 1364–1380.
- Viemeister, N., and Wakefield, G. (1991). "Temporal integration and multiple looks," *J. Acoust. Soc. Am.* **90**, 858–865.
- Wallace, A. B., and Blumstein, S. E. (2006). "Nonspeech sounds prime acoustically similar words," *J. Acoust. Soc. Am.* **119**, 3245.
- Wartenburger, I., Steinbrink, J., Telkemeyer, S., Friedrich, M., Friederici, A., and Obrig, H. (2007). "The processing of prosody: Evidence of inter-hemispheric specialization at the age of four," *Neuroimage* **34**, 416–425.
- Zaehle, T., Wüstenberg, T., Meyer, M., and Jäncke, L. (2004). "Evidence for rapid auditory perception as the foundation of speech processing: A sparse temporal sampling fMRI study," *Eur. J. Neurosci.* **20**, 2447–2456.
- Zatorre, R. J., and Belin, P. (2001). "Spectral and temporal processing in human auditory cortex," *Cereb. Cortex* **11**, 946–953.
- Zatorre, R., Belin, P., and Penhune, V. (2002). "Structure and function of auditory cortex: Music and speech," *Trends Cogn. Sci.* **6**, 37–46.
- Zwicker, E. (1961). "Subdivision of the audible frequency range into critical bands (frequenzgruppen)," *J. Acoust. Soc. Am.* **33**, 248.
- Zwicker, E., and Terhardt, E. (1980). "Analytical expressions for critical-band rate and critical bandwidth as a function of frequency," *J. Acoust. Soc. Am.* **68**, 1523–1525.
- Zwislocki, J. (1960). "Theory of temporal auditory summation," *J. Acoust. Soc. Am.* **32**, 1046–1060.