Tracking perception of the sounds of English

Natasha WarnerJames M. McQueenAnne CutlerBRM

Citation: The Journal of the Acoustical Society of America **135**, 2995 (2014); doi: 10.1121/1.4870486
View online: http://dx.doi.org/10.1121/1.4870486
View Table of Contents: http://asa.scitation.org/toc/jas/135/5
Published by the Acoustical Society of America

Tracking perception of the sounds of English

Natasha Warner[a)]
Department of Linguistics, University of Arizona, Box 210028, Tucson, Arizona 85721-0028

James M. McQueen[b)]
Radboud University Nijmegen, Postbus 9104, 6500 HE Nijmegen, The Netherlands

Anne Cutler[c)]
The MARCS Institute, University of Western Sydney, Locked Bag 1797, Penrith, New South Wales 2751, Australia

(Received 6 August 2013; revised 21 March 2014; accepted 24 March 2014)

Twenty American English listeners identified gated fragments of all 2288 possible English within-word and cross-word diphones, providing a total of 538 560 phoneme categorizations. The results show orderly uptake of acoustic information in the signal and provide a view of where information about segments occurs in time. Information locus depends on each speech sound's identity and phonological features. Affricates and diphthongs have highly localized information so that listeners' perceptual accuracy rises during a confined time range. Stops and sonorants have more distributed and gradually appearing information. The identity and phonological features (e.g., vowel vs consonant) of the neighboring segment also influences when acoustic information about a segment is available. Stressed vowels are perceived significantly more accurately than unstressed vowels, but this effect is greater for lax vowels than for tense vowels or diphthongs. The dataset charts the availability of perceptual cues to segment identity across time for the full phoneme repertoire of English in all attested phonetic contexts.
 [http://dx.doi.org/10.1121/1.4870486]

PACS number(s): 43.71.Es [BRM] Pages: 2995–3006

I. INTRODUCTION

Speech is an efficient information carrier. Speakers typically produce around 11.7 phonemic segments per second (Greenberg, 1999); these form syllables and, in turn, words and sentences to convey meaning. Information about segments overlaps so that listeners can receive information about a given segment not only before all of the previous ones have been heard, but also after the next one has started. Listeners do not wait for all acoustic information relevant to a segment, but interpret the incoming stream of information probabilistically. Our study investigates the timing of this process for all speech sounds of English in all their possible segmental environments. Listeners heard all 2288 legal diphones (two-sound sequences, e.g., /ab/, /si/, /ps/, /ai/) of a variety of English in fragments varying from one-sixth to all of the diphone, and reported the two sounds they thought most likely to be what they heard.

Our study forms part of a long tradition of datasets on perception of speech sounds. Peterson and Barney (1952) performed the classic such study for vowels, and Miller and Nicely (1955) for consonants; the former analyzed data for ten vowels, while the latter (and recent follow-ups by Phatak *et al.*, 2008, and Paláez-Moreno *et al.*, 2010, as well as Wang and Bilger, 1973) examined 16 English consonants under various noise and filtering conditions. The study with children by Nishi *et al.* (2010) concerned 15 English consonants; Hillenbrand and Nearey (1999) tested perception of natural and resynthesized vowels in /hVd/ context; Benkí (2003) studied perception of 120 consonant-vowel-consonant (CVC) strings in noise.

All of the above studies addressed perception of the entire duration of the sound or syllable, degraded in various ways, and none relate to timing of information. Our work differs chiefly in that perception was scored for each fragment in which the diphones were presented, so that our data reveal how information about sounds and phonological distinctions is conveyed over time. There are also some preceding studies, all but one more limited in scope than ours, in which gating (presentation of fragments of gradually increasing duration) was used to study when acoustic cues become available. Furui (1986) examined perception of the 100 CV or C/j/V syllables of Japanese over very fine-grained time steps, and Smits (2000) presented similar data for 51 VCV sequences of British English. Jesse and Massaro (2010) examined the timing of perception of 22 English consonants in a CV environment based on audio, visual, or audiovisual cues. The work most similar to our project was conducted on Dutch by Smits and the present authors (Smits *et al.*, 2003; Warner *et al.*, 2005), with a closely similar design to the present project. It, too, presents an extensive database, on Dutch diphone perception. It is the only preceding study as comprehensive as the present one.

A study such as this provides information that can be compared across any segments, sequences, or words of the language since all diphones are included and the data come from a single consistent task using the same listeners. Our

[a)]Author to whom correspondence should be addressed. Electronic mail: nwarner@u.arizona.edu
[b)]Currently at: Donders Institute for Brain, Cognition and Behaviour, Centre for Cognition, and Behavioural Science Institute, Radboud University Nijmegen. Also at: Max Planck Institute of Psycholinguistics, Postbus 310, 6500 AH Nijmegen, The Netherlands.
[c)]Also at: Max Planck Institute of Psycholinguistics, Postbus 310, 6500 AH Nijmegen, The Netherlands and Donders Institute for Brain, Cognition and Behaviour, Postbus 9104, 6500 HE Nijmegen, The Netherlands.

diphone set comprises all consonants (C) and all vowels (V) of a common variety of American English in all combinations (CV, VC, CC, or VV) that the language allows. Diphones that occur only across word or compound boundaries (e.g., /ʧm/, as in *batch mode*) were included, as well as more typical diphones that occur syllable internally. All vowels appear both as stressed and unstressed (e.g., /bu/ or /æʃ/ each have two diphones with stressed vs unstressed vowel, while /oʲeʲ/ has four, as in *annoy eighteen*, *alloy aging*, *annoy aging*, *alloy eighteen*). Six gates were created for each diphone, presenting the first third, first two-thirds, and entirety of Segment 1, and Segment 1 plus the first third, first two-thirds, and entirety of Segment 2. (Some diphones, however, only have four gates as explained in Sec. II A.) This yielded a total of 13 464 stimuli. These were presented in random order to listeners who then identified the two sounds of each stimulus as best they could. The resulting dataset will enable investigation of perception of consonant place, manner, or voicing, vowel quality, stress, time point within segments, or properties of and interactions with preceding or following segments. So that all researchers can make use of this substantial dataset, our results are publicly available at http://www.u.arizona.edu/~nwarner/WarnerMcQueenCutler.html.

II. METHODS

A. Materials

We compiled a list of all diphones that can occur either within a word or across word boundaries in American English using the segment inventory in Table I. This inventory reflects the system of the electronic dictionary of American English at http://lexicon.arizona.edu/~hammond/newdic.html (accessed 3/5/2014) (related to the dictionary file at http://dingo.sbs.arizona.edu/~hammond/lsasummer11/newdic discussed in Pisoni *et al.*, 1985). However, we treated the flap allophone [ɾ] as a separate segment since its occurrence is not fully conditioned by the diphone environment, we omitted /ɔ/ because it does not occur in the Arizona dialect or in many other parts of the United States, and we merged the unstressed central vowels [ɨ, ə] since many speakers and listeners are unsure what the [ɨ] category represents. To avoid duplication we also omitted diphones with syllabic consonants ([n̩, m̩], etc.), given that non-syllabic sequences of the same segments were already in the corpus ([tn] in *catnip* vs [tn̩] in *button up*). We did not omit [ɾl̩] as in *bottle* because only syllabic [l̩] can follow [ɾ]; non-syllabic [l] cannot.

TABLE I. American English segment inventory for the diphone list. (A) Consonants. (B) Vowels.

(A) Consonants		
	Voiced	Voiceless
Stops/affricates/flap	b, d, g, ʤ, ɾ	p, t, k, ʧ
Fricatives	v, ð, z, ʒ	f, θ, s, ʃ, h
Nasals	m, n, ŋ	
Glides/approximants	j, w, ɹ, l	

(B) Vowels			
	Front	Central	Back
High	i, ɪ		u, ʊ
Mid	eʲ, ɛ	ʌ, ə, ɚ	oʷ
Low	æ		a
Diphthongs	aʲ, oʲ		aʷ

All combinations of two sounds were considered possible unless they did not occur within a word in this dictionary, could not be formed by the end of one word in the dictionary and the beginning of another, and a phonological reason for their impossibility is known. Hence, VV diphones with lax vowels as the first vowel (e.g., /ɛa/) were excluded because lax vowels cannot end a syllable or word. Furthermore, some sequences cannot occur because of vowel mergers before /ɹ, ŋ/ etc. in most varieties of American English (Ladefoged and Johnson, 2015). Thus, we used /eʲɹ/, but not /ɛɹ, /ɪŋ/, but not /iŋ/, etc., with the production representing the speaker's pronunciation of these strings. These constraints led to a list of 2288 diphones out of the 3136 that would occur if every segment in Table I plus syllabic [l̩] could precede and follow every other segment. (Notes on further detailed methods decisions appear online at http://www.u.arizona.edu/~nwarner/ WarnerMcQueenCutler.html.)

The diphones were recorded by a phonetically trained female speaker who had lived almost her entire life in Tucson, Arizona, and who was monolingual in English until her teenage years. The stimuli thus represent the speech of one speaker, but that speaker is highly appropriate for the choice of dialect. As in Smits *et al.* (2003), contexts were appended: (i) a following context for each diphone (/k/, /keʲ/, or /kə/ after vowels and /eʲ/, /ə/, or /a/ after consonants) to avoid final lengthening within diphones; (ii) a preceding context for some diphones (/a/ before C, /b/ or /ab/ before V). Most CC diphones (e.g., /fp/) cannot occur word-initially, but a preceding vowel makes them pronounceable in a natural way. To avoid preceding context signaling particular diphone types, some remaining diphones also received preceding contexts (giving overall 71% of diphones with preceding context). Preceding and following contexts also helped the speaker to pronounce target stress patterns in VV diphones (e.g., /ˈabiukˈeʲ/ for unstressed-unstressed /iu/, /bˈiˈukə/ for stressed-stressed). CV and VC diphones were followed by /(k)ə/ if the diphone's vowel was stressed, /ˈ(k)eʲ/ if unstressed; VV stressed-stressed and unstressed-unstressed diphones had following syllables with opposite stress. The choices of which context to use before and after each diphone were the same as specified in Smits *et al.* (2003).

We then identified the boundary between the two segments of the diphone, as well as between the diphone and any preceding or following context. Separate boundary criteria were applied for voiceless consonant to voiced segment (onset/offset of voicing), voiced obstruent to voiced segment (F2 onset/offset), nasal to vowel or sonorant (sudden change in frequency of energies), /l/ to vowel (most sudden increase in amplitude of formants), glide or /ɹ/ to vowel (midway through duration of F2 or F3 transition, respectively), voiceless consonant to voiceless consonant (onset/cessation of defining features such as closure, burst noise, or frication

noise) and vowel to vowel (beginning of creak or glottal stop if any, midway through F2 transition, otherwise). Boundary position decisions were closely modeled on the methods of the Dutch work (Smits *et al.*, 2003), in order to make the data for the two languages comparable. Additional details about boundary locations appear in Smits *et al.* (2003).

Recordings were final-gated to produce (with one exception, next paragraph) six stimuli per diphone, usually with a gate termination at each third of the way through the first and second segment of the diphone. That is, in the diphone /sa/, the shortest stimulus included the preceding context (if any) through to one third of the way through /s/; Gate 2 included that material and extended to the point two-thirds through /s/; Gate 3 ended at the boundary of /s/ and /a/; Gate 4 went to one-third through /a/, Gate 5 to two-thirds through /a/, and Gate 6 to the end of the diphone. Thus, any preceding context recorded with the diphone was always presented as part of the stimulus, and the following context was never presented as the last gate included the whole diphone but no transition to the following context. Gate endpoints were defined by proportions of duration of segments, rather than by absolute number of milliseconds (e.g., one gate per 20 ms) because this allows one to compare across all segment types how well listeners can perceive sounds by one-third or two-thirds of the way through the segment. Gating at fixed time intervals would make comparison across manners of articulation (e.g., /m/, which is long, vs /d/ or /j/, which are short), or even across individual stimuli, very difficult.

An exception to the equal gate size occurred with stops and affricates following another segment. Here, the second gate point within the segment was just before the beginning of the burst, rather than at two-thirds of the segment's duration. This avoided having some diphones with the burst in the second gate, but others with the burst only in the third gate. Thus, for all stops and affricates, the burst [and voice onset time (VOT)] information was only available to listeners as of the last gate within that segment (Gate 3 if the stop/affricate was Segment 1 of the diphone, Gate 6 if it was Segment 2). The first gate endpoint within these segments was placed at halfway through the duration from segment onset to pre-burst gate point, thus, halfway through the closure. [ɾ] was not treated as a stop since it often has no burst, but rather had its endpoints at one-third and two-thirds through the flap duration. These gate points were thus close in time, but using three equal gates makes the data comparable across diphones.

The exception to the six-gate pattern concerned stops and affricates as Segment 1 of the diphone, recorded without preceding context. Here, the silent closure phase could not be located. For these 132 diphones, only 4 gates were presented: 1 reaching the end of Segment 1, plus the 3 normal end-points for Segment 2. That is, the two gates that would normally end during the stop/affricate closure were simply omitted as they contained only silence.

Each stimulus token was created by extracting the speech from onset of the diphone or preceding context (if any) to the gate point for that stimulus, then ramping the speech to a square wave with f0 = 500 Hz, which continued for 295 ms after the ramp. The amplitude of the speech was ramped down over a 5 ms time window as the amplitude of the square wave was ramped up. These signals were added to produce a smooth transition from speech to square wave (beep). The square wave amplitude was loud enough to convey a clear beep, but quiet enough not to irritate listeners; the square wave f0 was high enough to prevent resemblance to any speech sound. The square wave and ramp were used to prevent the artifactual perception of a labial consonant that can occur with speech cut suddenly to silence (Öhmann, 1966; Pols and Schouten, 1978).

B. Subjects

Twenty-eight listeners (five male) began participation in the experiment; six (two male) did not finish it. All listeners were monolingual in American English until at least their teenage years and had no substantial exposure to other language(s) in childhood, nor more than a few years' classroom study of foreign languages in school. All grew up in the Southwest of the United States (some came from Texas or southern California, but most from Arizona), and all were students at the University of Arizona at the time of the study. Thus, the listeners' dialect was well matched to that of the speaker. The listeners were recruited through the University of Arizona's honors program to select participants most likely to return reliably for the many sessions, and most able to learn the response symbols easily. No listener had any known speech, hearing, or reading problem. Of the six who did not complete the study, three chose to stop, one missed frequent appointments, and two were dropped due to poor performance in practice.

C. Procedures

The stimuli were randomized and grouped in short experimental blocks, expected to take 10–20 min each to complete, so that listeners could complete 3–5 such blocks during each 1 h experimental session. The order of blocks was varied for individual subjects (though not fully randomized). Four practice blocks were also created using actual stimuli from the experiment with disproportionately many stimuli containing segments for which the response symbol was expected to be relatively difficult to learn (e.g., "dh" for /ð/, "g" as /g/ and not /ʤ/, most vowels).

English spelling is too ambiguous to convey responses, but we used response symbols that were based as closely as possible on typical English spellings (e.g., "oy" for /oʲ/, "j" for /ʤ/, "p" for /p/). Listeners were first instructed in these symbols, and then performed practice blocks for ∼45 min (223 or 335 stimuli for most listeners, depending on how many blocks they completed). Data from practice blocks were used only to evaluate listeners' ability to do the task, and all stimuli presented in the practice sessions appeared again during actual experiments. (Because of the very large number of stimuli, many similar, this is not problematic. A listener is unlikely to recall having heard one of the 335 practice stimuli when hearing it again among 13 464 experimental stimuli.) Two listeners scored <50% correct on both Segments 1 and 2 even at Gate 6 (when both sounds should

be relatively perceptible) during the practice, and had random, perceptually unmotivated error patterns, so their participation was discontinued.

For experimental sessions, listeners sat in an individual sound-protected booth and heard stimuli presented over headphones. Each stimulus was accompanied by a display on a computer screen showing all response alternatives as buttons on the left half of the screen, and the same alternatives on the right half of the screen, with a dividing line between the halves. Listeners used a mouse to click first on the left half of the screen on the first sound they thought they heard, then on the right half on the second sound they heard (or that might have come next). The response options were the same as the inventory of segments (Table I), except that [ɾ] and syllabic [l̩] were not given as options since English listeners consider these types of /t, d, l/. Listeners were also not asked to distinguish /ə/ from /ʌ/, but were trained to use "uh" for both, and to use "er" for both stressed and unstressed /ɚ, ɝ/ (separate symbols in the dictionary used; N.B.: identifying stress was not part of the listening task).

If the diphone was recorded with a preceding context, the context was displayed on the left of the screen in the spelling system of the responses to indicate that the sounds of the preceding context were not part of what the listener should respond to. Thus, for the diphone /iu/ (both unstressed), recorded in /ˈabiukˈeˌ/, "ahb" was shown to the left of the response buttons.

Listeners returned to the lab for multiple one-hour sessions, completing as many experimental blocks as they could per visit (with a brief break between each). Listeners took an average of 32.73 sessions to complete the experiment (range: 28–39). They received a small monetary compensation for each visit, and a bonus equal to five sessions' compensation on completion. After most listeners had finished the experiment, we realized that we had erroneously omitted 25 stimuli. These 25 were randomized with 55 fillers (stimuli from other diphones that had already been presented), and subjects returned to complete these stimuli; responses to fillers in this session were not analyzed.

III. RESULTS

Percent correct responses and type of incorrect responses were computed for each segment of each diphone. The proportion correct averaged over all diphones containing a given segment as Segment 1 (or 2) was then calculated (stressed and unstressed vowels counted separately). Thus, Subject 1's proportion correct for stressed /a/ at Gate 1 represents how often Subject 1 correctly chose /a/ as Segment 1 for all 101 Gate 1 stimuli with /a/ as Segment 1, regardless of Segment 2 identity. Tables II and III present confusion matrices, respectively, for consonants and vowels.

In Secs. III A–III D, we present statistical comparisons analyzing several of the most salient differences within each manner class. The choice of which comparisons to present is also informed by the analyses included in Smits *et al.* (2003). All statistical analyses below were conducted with subject as random factor on proportions correct out of all diphones with the same Segment 1 or Segment 2. Before statistical analysis, proportions were converted to Rationalized Arcsine Transformed Units (RAU), using Eqs. (2) and (3) in Sherbecoe and Studebaker (2004), which adjust for proportions calculated over <150 stimuli. The analyses of variance reported in Tables IV–VII are within-subjects pairwise comparisons of related sound types at each gate, using RAU proportions over all diphones in the relevant category as the dependent variable (e.g., all diphones with /d/ or /t/ as Segment 1 for the comparison of Segment 1 /d/ to /t/). Initial analysis of the data showed that the accuracy of two listeners (one male) was more than 3 standard deviations below the average of all other listeners for perception of either Segment 1 or 2 at three or more gates. They differed in which information they failed to use (in which gates for which segment), but both were clear outliers. These two listeners' data were excluded as not representative, so all figures and tables show the data of the 20 remaining listeners. No other listeners differed markedly from the rest of the group at multiple gates. Figure 1 shows overall accuracy for all consonants, all vowels, and all segments, and clearly reveals listeners' increasing uptake of information as the acoustic signal proceeds.

A. Stops, affricates, and flap

Figure 2 shows percent correct results for stops, affricates, and flap [Fig. 2(A) as Segment 1, Fig. 2(B) as Segment 2). Results for Segment 1 are presented separately for diphones with only four gates (stops, affricates without preceding context) vs with six gates. Gate 3 of diphones with four gates includes only the release burst and any aspiration noise, so for voiceless unaspirated /b, d, g, ʤ/, this gate is short, thus lowering accuracy.

Several overall patterns are evident. Phonemically voiceless stops (/p, t, k/) are usually identified better than their voiced counterparts (/b, d, g/) early in the preceding segment [Gates 1–2, Fig. 2(B)] and once the release burst and any aspiration noise have been heard [Gates 3–6, Fig. 2(A) and Gate 6, Fig. 2(B)]. During the closure of the stop itself, however, the voiced segments are usually perceived as well as the voiceless segments or even more accurately, and this pattern may begin even by the end of the preceding segment [Gates 1–2, Fig. 2(A) and Gates 3–5, Fig. 2(B)]. Statistical results (Table IV) confirm this pattern especially for b/p and g/k (/d∼t/ is discussed below). This suggests that early in the preceding segment, listeners may perceive some place information, but use voiceless as a default choice for voicing. By the end of the preceding segment, longer duration before a voiced stop may be conveying information about voicing. During the stop's closure, voiceless silence conveys no further information, but a voiced closure does, leading to the advantage for voiced segments. Finally, the noisy, longer VOT of /p, t, k/ leads to an increase in perceptibility for these stops during the release.

Figure 2 also shows several individual deviations, for instance, that, relative to other voiceless stops, /t/ is perceived poorly at many gates. The general pattern of better perception of voiceless stops during the preceding segment and once the burst has been heard is shifted by the overall

TABLE II. Confusion matrices for consonants: first segment at Gate 1 (top line of each Stimulus row) and second segment at Gate 4 (bottom line). Responses are summed over subjects and over all diphones containing the consonant. The next-to-rightmost column is the total number of vowel responses to the consonant (any vowel).

	Response																								
Stimuli	p	t	k	b	d	ɡ	ʧ	ʤ	f	θ	s	ʃ	h	v	ð	z	ʒ	m	n	ŋ	l	ɹ	w	j	Vowel	Total
p	**542**	22	40	5	6			1	15	9			20					6			4	1	1		68	740
	482	30	8	95	26	9	2	3	14	21		1	58	14	6	4	2	25	23	5	11	3	13	2	123	980
t	106	**414**	45	6	49		20		4	22			17		2				2		1	7	2		83	780
	9	**333**	6	2	188	11	26	8	8	68	33	5	60	5	8	7	1	4	48	9	5	4	3	9	120	980
k	19	31	**535**	1	3	8			3	10			39		1				1		2	1	4	1	61	720
	22	100	**266**	12	50	48	6	3	7	21	3	1	162	7	5	2	5	5	18	4	4	1	7	16	205	980
b	66		2	**522**	22	11				1			8	29	2			4	1	2	2	2	2		44	720
	98	12		**494**	32	6	5	4	5	13	2	1	41	30	8			34	19	2	15	7	44	2	86	960
d	8	23	1	12	**582**	3		6					3	2	4				7		2	20	1	1	45	720
	13	46	6	21	**499**	8	6	12	9	23	4		38	24	20	2	3	8	51	3	14	10	6	3	131	960
ɾ	4	**15**		3	**207**		1	1					3	2	1						3	45			15	300
	6	**50**	1	8	**333**	1			1	6			15	14	11		2		6		40	23	1	2	60	580
ɡ	1	3	4	2	24	**557**	1	1					4	1		1			3	3	3	5	6	4	57	680
	19	53	39	14	86	**309**	5	10	3	15	4	1	77	24	9	4	4	3	18	4	3	6	12	29	189	940
ʧ	4	326	44	1	53	1	**56**	1		2		2	23						5		2	6		2	52	580
	39	262	13	8	167	7	**64**	19	6	38	6	19	77	14	5	2	1	4	26	3	7	7	12	5	129	940
ʤ		3	1	3	500	5	2	**5**					5					1	5			6			44	580
	6	28	4	22	558	7	9	**20**	2	8	1	1	43	11	13	3	3	5	32	4	8	7	10	11	124	940
f	13	8	6	6	2				**725**	145	5		41	10				3				6			50	1020
	8	5	9	1	3		6	3	**555**	162	11	2	30	75	10	1	1		7		7	4	7		73	980
θ	16	19	3	7					248	**346**	3		19	10	6			3	1			1	1		37	720
	2	3	5	1	2	2	3	1	151	**595**	5	1	38	26	33	2	1	1	4	1	6		1		56	940
s	11	91	3	1		1	1		9	23	**807**	7	22	1	1	12						1	1		28	1020
	4	17	3	3	2		5	3	2	48	**696**	21	10	3	5	62	1	1	1	2	2				49	940
ʃ	5	7	1	2	1		64	7	3	1	10	**611**	9			2	4					4	1		8	740
		14	3		8	5	39	33	6	15	6	**659**	21	2	3	2	42		4		3	5		1	69	940
h	14	5	11	4			1	1	11	7		4	**502**	2	1		1	1	2		1	3	4	1	24	600
	16	39	3	5	10	4	7	7	30	41	7	2	**346**	9	7	1	6	7	23	5	6	5	8	4	282	880
v	8	3		15	2	1			29	16	1		17	**743**	15	2	1	16	6		15	28	16		86	1020
	4	10		16	5	6	1	5	65	31	4	3	29	**598**	20	3	7	15	14	5	8	2	23		66	940
ð	7	1		84	28	2			9	37			20	454	**117**		1	25	17		26	22	28		102	980
	6	6	1	8	31	5	5	2	6	176	4	4	19	224	**230**	1	3	3	15	4	24	11	3	1	88	880
z	1	11		5	5				2	7	4		4	15	6	**877**	2	1	1		24	3	2		50	1020
	3	3	4	2	13	1	3	6		17	54		9	2	4	**763**	9		3		6	1			37	940
ʒ				3	9	8	1	164		1	1	7	3	3		9	**624**	1	11		26	113	1	5	30	1020
	1	1		2	14	10	17	133	1	2	2	40	14	2	5	10	**569**	1	7	1	3	3	4	4	74	920
m	15	2	2	32					2				15	8				**693**	94	13	20	3	19	2	80	1000
	9	11	2	9	6	1	3	2	4	6	2	1	17	19			1	**644**	112	6	8	3	26	3	85	980
n	13	1		34	5				2	1		1	12	2				78	**736**	4	8	3	58	2	60	1020
	2	17	9	5	38	8	4	5		15	2		23	13	12	2	3	44	**629**	3	29	5	11	4	97	980
ŋ	1		1		1	6			1				8			1		23	108	**659**	25	6	8	3	129	980
		2				3							3					9	32	**137**	2		1	2	9	200
l	42	4	3	60	4	4			2	1	1		14	4	1			13	9	6	**655**	11	102		84	1020
	7	37	1	12	18	3	1	2	3	9			28	14	7	1	1	21	15	3	**543**	5	27	3	219	980
ɹ	33	10	14	66	18	6			7	3			13	7				22	12		23	**689**	51	1	65	1040
	16	18		9	20	5		1	2	1	1		17	15	2		4	13	13	1	21	**370**	58	2	211	800
w	6			56									9	1				7	4		39	6	**361**		71	560
	8	47	2	9	14	6	4	2	5	7		2	29	12		1	2	21	26	8	45	15	**349**	1	245	860
j	3	4		22	13	4		1	2	1			12	4	1	1		13	7		8	1	6	**192**	185	480
	3	33	4	9	18	2	1	8		2			21	9	1			3	17	2	8		3	**195**	441	780
Total	938	1003	716	952	1534	617	147	112	188	1074	633	832	632	842	1298	158	905	633	910	1032	687	889	993	675	214	
	783	1177	389	767	2141	467	222	777	292	885	1340	847	764	1225	1166	424	873	671	871	1163	212	828	497	629	299	

weak perception of /t/, but there is an indication of the same pattern. Further, /t, k, ʧ/ fail to show improvement in accuracy from Gate 4 to 5 for Segment 2, as does /p/ to a lesser extent, and these four segments also show little or no improvement from Gate 1 to 2 for Segment 1. These time periods cover the second half of the silent closure. Listeners

TABLE III. Confusion matrices for vowels: first segment at Gate 1 (top line of each Stimulus row) and second segment at Gate 4 (bottom line). Responses are summed over subjects and over all diphones containing the vowel. The next-to-rightmost column is the total number of consonant responses given. Syllabic [ɻ̩] is only unstressed and appears only as Segment 2.

	Response to stressed stimulus vowels																Response to unstressed stimulus vowels													
Stimuli	i	ɪ	e^j	ɛ	æ	a	o^w	u	ʊ	ʌ	ɚ	a^j	o^j	a^w	Consonant	Total	i	ɪ	e^j	ɛ	æ	a	o^w	u	ʊ	ə	ɚ	a^j	o^j	a^w	Consonant	Total
i	**946**	12		1	1		1	3		1	1				14	980	**916**	3	1	4	1	1			1	1	2		1		29	960
	784	34		1	1	1		16		1					22	860	**724**	53	2	5	1		3	18	1	14		2			57	880
ɪ	1	**294**	60	84	1	1		1		2		1			35	480	1	**214**	107	144	2		1	1		1	1	1			7	480
	12	**311**	217	180	5	10	10	2	7	63	2	3		2	56	880	18	**274**	270	192	4	7	18	7	12	53	4	6	2	3	50	920
e^j	3	171	**570**	231	4	6			1	1		10	1		22	1020	5	109	**608**	229	4	3				1	1	8			12	980
	3	197	**477**	168	2	8	2	1	3	5	1	5		1	27	900	5	126	**465**	236	11	8	1	1		15		4	1	2	25	900
ɛ	2	10	15	**233**	73	39	2			17		1		39	9	440	2	12	67	**277**	27	21				12		1	1	5	15	440
	4	3	75	**307**	151	178	3	2	1	109	4	14		33	16	900	1	40	152	**294**	86	137	10	1	1	97	3	10	1	14	53	900
æ			1	9	**167**	122	1			4			1	106	9	420		4	8	60	**167**	102			1	10	1			53	14	420
	1		8	52	**365**	299	1			58	1	15	3	79	18	900	2	5	38	117	**272**	276	4		1	84	2	19		53	27	900
a			6	1	21	**851**	1			53	1	27	1	28	30	1020			9	3	42	**772**	1		2	69	3	18	1	57	23	1000
			3	1	37	**704**				102	2	29		14	28	920		4	4	7	34	**573**	14		2	189	3	28		21	21	900
o^w	2	2	1	5	1	16	**528**	3	31	331	8		12	17	43	1000	1	1	1	4	1	12	**575**	3	45	282	15		22	21	17	1000
	2	4	2	8	1	74	**360**	9	21	301	6	3	20	11	38	860	1	16	8	18		40	**362**	5	36	290	26	1	20	10	47	880
u		9	1			1	4	**923**	25	6		1		2	8	980	1	3		1			5	**890**	20	11	4				25	960
	39	63	2	5	1	3	6	**600**	40	41	6		1	1	52	860	54	72	9	7		3	15	**502**	39	54	6		2		97	860
ʊ		1		4	2	12	97	45	**122**	232	44	1	11	5	44	620	1	5	1	2	2	11	78	35	**121**	319	12		5	8	20	620
	6	13	2	5		26	49	14	**42**	221	4	2	6	5	25	420	2	40	7	13		17	60	10	**35**	212	8	2	3	5	26	440
ʌ/ə	1		3	2	3	187	7		3	**148**		6	1	20	19	400	2	29	17	87	27	211	43	1	26	**447**	44	6	2	35	63	1040
		3	5	10	10	349	30	4	6	**384**	4	15	2	14	24	860	4	82	94	132	16	117	48	2	25	**329**	15	7	4	12	93	980
ɚ	1	5		6	1	4	1	2		37	**692**			1	50	800	1	3		2	2	2	7		5	16	**893**			1	68	1000
		17	6	15	4	11	47	8	19	92	**459**	1	5	1	75	760	3	32	19	25		9	39	11	22	96	**483**	1	6	2	112	860
a^j		1	16	4	128	634	1		1	43		**84**		74	14	1000			13	16	176	502	1	1	1	57	1	**88**	2	132	10	1000
	2	2	32	69	121	331	6			112	2	**109**		38	16	840		3	28	91	140	313	3	1		129	2	**79**		37	14	840
o^j	1	1			1	6	268	8	15	47	3		**571**	8	91	1020				1		4	209	7	12	45	2	2	**599**	5	94	980
		4	2	1	2	17	356	6	22	100	5	1	**224**	15	45	800	1	4	6	7		17	385	3	22	139	10		**149**	10	47	800
a^w	1		1	9	316	300	2		1	22	1	1	5	**336**	25	1020	4	2	2	30	240	252	3		1	37		6	3	**372**	28	980
		2	6	21	221	384	2			79	3	21		**103**	18	860		5	17	75	154	352	10		1	128	1	13	1	**85**	18	860
ɻ̩																		1					6			2				**11**	20	
Total	958	506	674	589	719	2179	913	985	199	944	750	132	603	636	413		934	385	834	860	691	1893	923	938	235	1308	979	130	636	689	425	
	853	653	837	843	921	2395	872	662	161	1668	499	218	261	317	460		815	757	1119	1219	718	1869	978	561	197	1831	563	172	189	254	698	

TABLE IV. F ratios (Bonferroni-corrected significant comparisons only) for stop and affricate voicing with rationalized arcsine transformed proportions. Cells show which phoneme of a pair was perceived more accurately (in heading if effect direction is consistent). Significance level: $p < 0.00208$ with Bonferroni correction for 24 comparisons; df: (1,19) for each comparison. For Segment 1 Gate 3, diphones with preceding context (6-gate diphones) are used. No differences for /b/ vs /p/ as Segment 1 were significant.

	Segment 1			Segment 2			
Gate	d > t	g/k	ʧ > ʤ	p > b	d/t	g/k	ʧ > ʤ
1	33.12		21.34	15.61			
2	118.55	g 13.49		16.77			
3	32.17		28.73				
4	12.75	k 32.04	32.36		d 25.62		56.64
5	18.63	k 25.02	19.67		d 33.13	g 22.93	30.81
6		k 23.70	19.81	500.28	t 168.37	k 157.14	38.84

appear to gain little information from continued silence and cannot improve accuracy further until the stop's release. The voiced stops /b, d, g/ as Segment 2 show a lesser reduction in slope from Gate 4 to 5. Thus, continuation of voiced closure provides little new information, in general, but more information than continued voiceless closure.

Affricates show a distinct pattern from stops: they are perceived poorly until the affrication has been heard (Gate 3 for Segment 1, Gate 6 for Segment 2), but show sudden improvement at that point. This reflects listeners' general processing of incomplete acoustic information. When no unambiguous information is available, listeners delay responding, but if a segment can be identified, listeners do not delay in case further acoustic cues might arrive (and potentially alter the identification). Listeners will not add missing acoustic cues to something that can be perceived as a segment (Ohala and Ohala, 1995; McQueen, 1995). Thus, until the end of the closure of an affricate, listeners assume it is a stop, and this percept only changes once the affrication itself becomes available. For example, at Gate 2 (before the burst) /ʧ/ as Segment 1 is identified as /t/ in 60.0% of responses, as /d/ in 4.8%, and correctly as /ʧ/ in only 8.4%, and /ʤ/ is identified as /d/ in 82.8% of responses, and correctly as /ʤ/ in only 1.0%.

The flap [ɾ] seems to be well perceived very early, even during the preceding segment, but shows little improvement

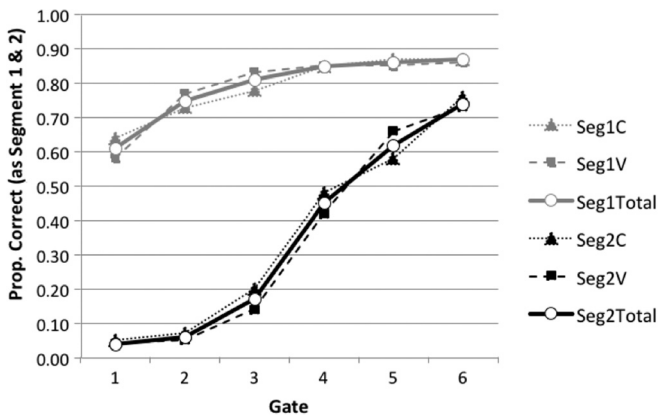

FIG. 1. Proportion correct as Segment 1 of the diphone (top set of lines) and Segment 2 of the diphone (lower set of lines), over time (gate end point 1–6). Average for all consonants, all vowels, and all segments.

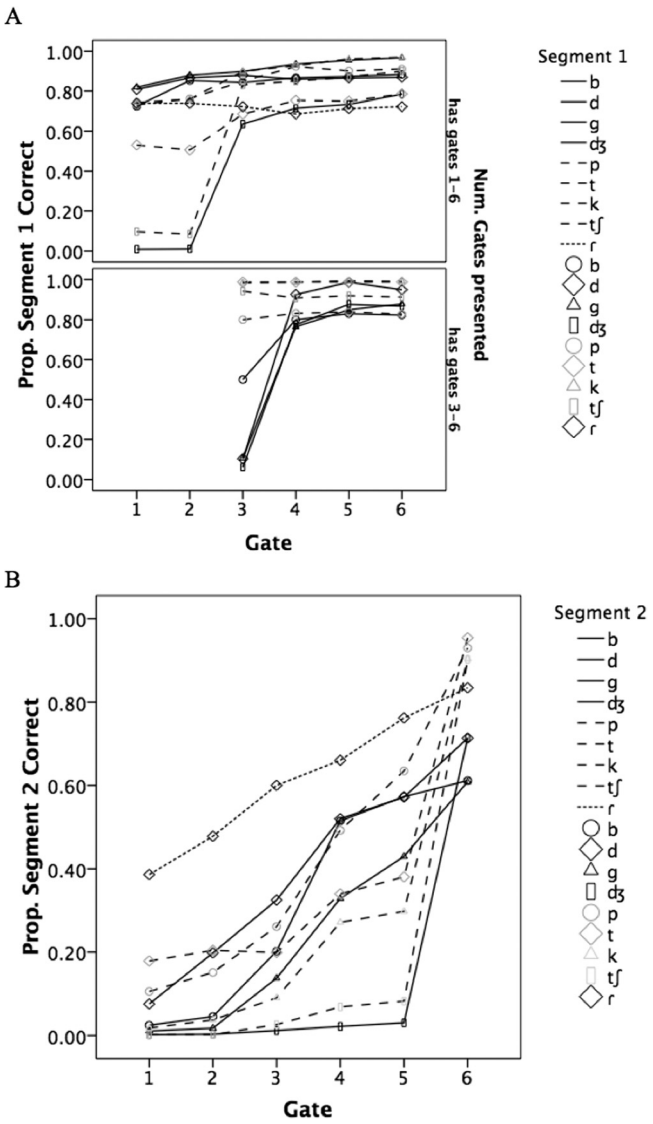

FIG. 2. Proportion correct over gate point for stops, affricates, and flap. (A) As Segment 1 (for 6-gate and 4-gate diphones separately). (B) As Segment 2.

over time and no such sudden improvement suggesting concentration of information. The flat improvement curve is likely to reflect the unusually short durations separating gate points during the flap itself, given the inherent short duration of [ɾ], and the high early accuracy is due to both "t" and "d" being counted as correct responses for flap. Preceding context for flaps was also more informative than average, since all diphones with flap as Segment 1 had to be preceded by /a/, and all flaps as Segment 2 had to follow a vowel or /ɹ/.

B. Fricatives

Results for fricatives appear in Fig. 3. Here, segment-specific effects outweigh any general effect of voicing. The segments /ð, θ, ʒ/ are all poorly perceived. Table V shows that for Segment 1, /ð/ is perceived worse than /θ/ (the next lowest fricative); /θ/ itself is perceived somewhat worse than /ʒ/ (the next lowest fricative, differences not significant with Bonferroni correction), and /ʒ/ is perceived worse than its voiceless counterpart /ʃ/ at all but one gate. The differences between /ð, θ/ and /ʒ, ʃ/ are far larger than the differences

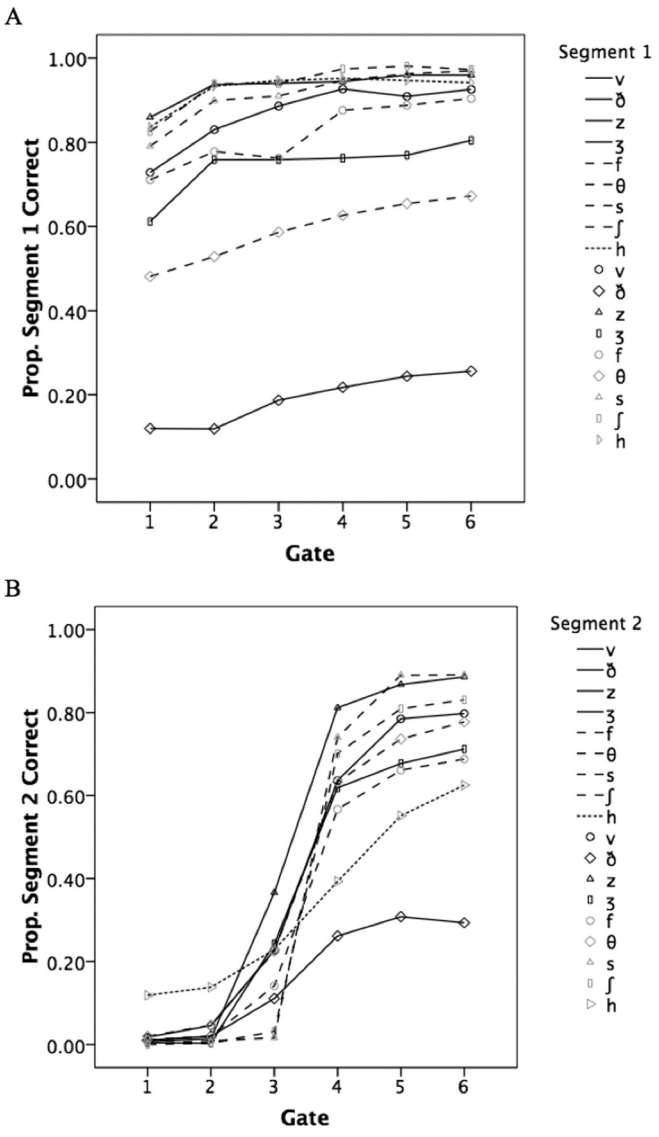

FIG. 3. Proportion correct over gate point for fricatives. (A) As Segment 1. (B) As Segment 2.

between voiced/voiceless pairs for stops, suggesting that segment-specific factors, rather than perception of voicing, dominate perception of these fricatives.

The interdentals are most often misidentified as labiodentals (/ð/ as /v/, /θ/ as /f/), far more often than the reverse confusion: Segment 1 /ð/ received 19% correct responses, 57.1% /v/ and 9.7% /θ/ (all other responses <3%), while Segment 1 /v/ was 86.8% correct, with /ð/ as the next most common response at 2.3%. This shows that the low accuracy for interdentals is not a spelling confusion stemming from English orthography (both interdentals written as "th"), but is a perceptual effect. Our finding here matches earlier perceptual results (Jongman *et al.*, 2000, 2003), and confirms that interdentals are far less perceptible than other English segments.

The other poorly perceived fricative, /ʒ/, has low frequency across the lexicon and has no standard grapheme, both of which most likely contribute to its low identification accuracy.

Besides the above fricative cases, only /f, v; s, z/ are distinct in voicing. In both pairs, the voiced member was better perceived at a few gates (Table V). Figure 3(B) shows a steep slope for most fricatives from Gate 3 to 4, and little further improvement beyond that, indicating that most perceptual information about these segments occurs in the first third of the frication noise.

TABLE V. F ratios (Bonferroni-corrected significant comparisons only) for fricative comparisons with rationalized arcsine transformed proportions. Direction of effect in column headers: significance level: $p < 0.00167$ for Segment 1, and $p < 0.00417$ for Segment 2 (Bonferroni correction for 30 comparisons for Segment 1 and 12 for Segment 2), df: (1,19) for each comparison. /ʒ/ vs /θ/ and /z/ vs /s/ (both Segment 1) were also tested, but were not significant.

	Segment 1			Segment 2	
Gate	θ > ð	ʃ > ʒ	v > f	z > s	v > f
1	110.28	17.48			
2	110.70	16.88			
3	82.53		18.53	748.11	94.91
4	69.69	21.75			
5	69.26	20.25			32.12
6	68.04	15.30			22.49

The most unusual pattern among the fricatives is for /h/: high accuracy throughout the diphone when it appears as Segment 1, and also during the preceding segment when it appears as Segment 2, but poor perception during the segment itself as Segment 2. The high correct response rates partly reflect some listeners' choice to use "h" as a default response when they could not identify a segment well at all and had to guess. The high accuracy as Segment 1 could further be due to the frication noise being more acoustically distinct than that for some fricatives (e.g., /f, θ/), and to the absence of a voiced competitor segment. Finally, poor perception as Segment 2 may be a phonotactic effect: English has no coda /h/, so listeners were less willing to choose it as a response after a vowel.

C. Sonorants

Figure 4 shows that among sonorant consonants, nasals are most easily perceived, while glides, /j/, in particular, are poorly perceived, often identified as vowels (e.g., /j/ as Segment 1 at Gate 2 is identified as /aʲ/ in 7.1% of responses and as /i/ in 11.5%, correctly perceived in 68.1%). Identifications of glides as diphthongs such as /aʲ, aʷ/ may stem from the glide following /a/ as a preceding context. /w/ was sometimes also misperceived as /b/ (3.8% of responses for Segment 1, Gate 2) and /l/ (3.4%), reflecting its labial constriction and the similarity of dark [ɬ] to a back glide or vowel. Syllabic [l̩] (used only in the diphone [ɾl̩], as discussed above) shows very similar accuracy to non-syllabic [l] (not tested statistically because there is only one diphone with syllabic [l̩], so proportion correct cannot be used). Overall, accuracy of sonorant perception, particularly as Segment 2, mirrors how consonantal a given sonorant is, in the sense of how distinct its acoustic boundaries are. Nasals (with the most discrete boundaries) are most perceptible, then liquids, then glides (the most vowel-like, and the most

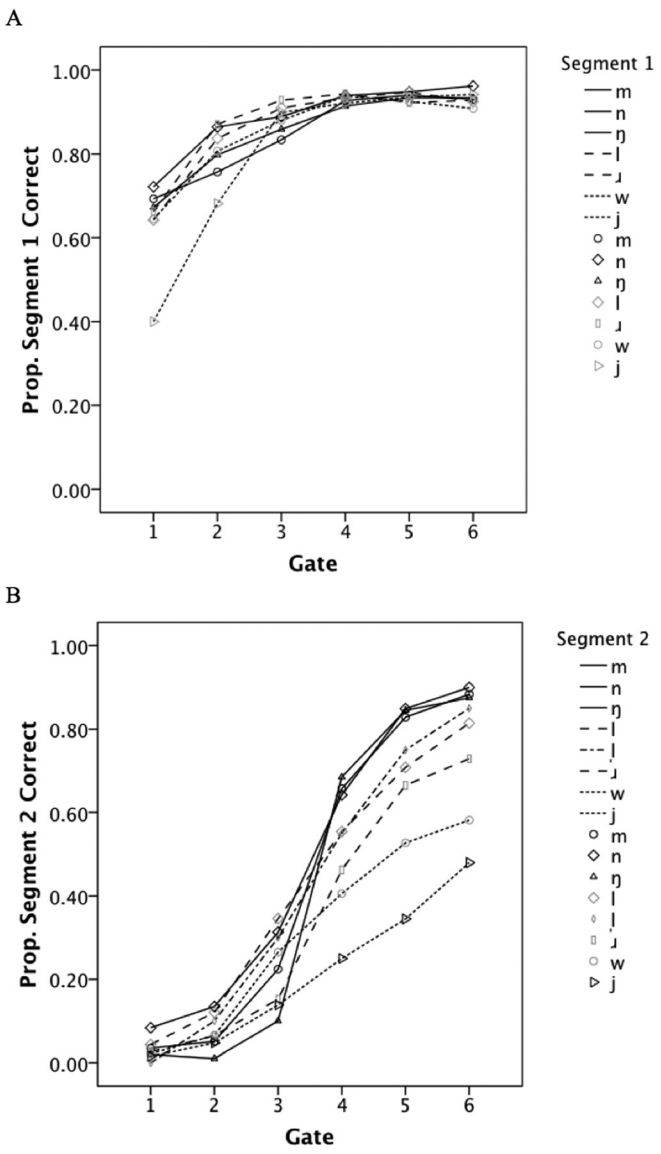

FIG. 4. Proportion correct over gate point for sonorants. (A) As Segment 1. (B) As Segment 2.

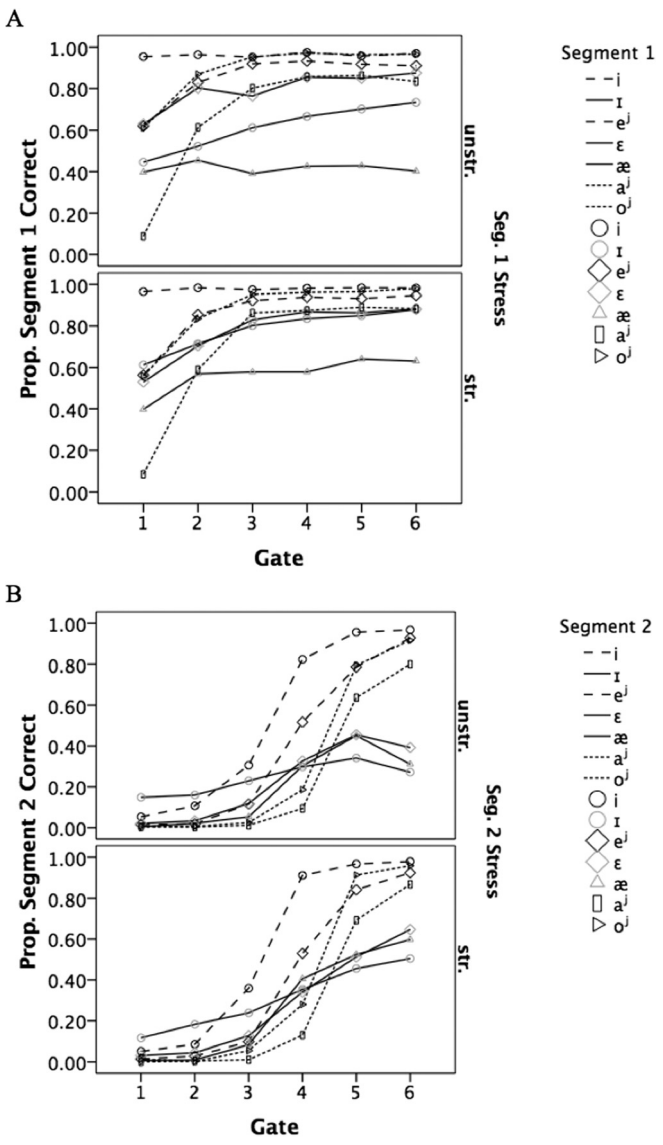

FIG. 5. Proportion correct over gate point for front vowels and front-ending diphthongs. (A) As Segment 1. (B) As Segment 2.

difficult consonants for which to identify acoustic boundaries).

D. Vowels

Results for vowels appear in Figs. 5 (front vowels and front-ending diphthongs) and 6 (back and central vowels), plotted by stress of the target segment as well as by position in diphone. In all cases, high tense vowels /i, u/ are identified accurately from early on. Tense vowels are generally perceived more accurately than their corresponding lax vowels (Table VI). It can be seen from the Segment 2 data that this effect develops at the end of the preceding segment or during the second segment (Gate 3 for /u, ʊ/, during Segment 2 for other pairs) because neither tense nor lax vowels are perceived well early in the preceding segment. /ʌ/ and /ə/ are sometimes perceived non-significantly better than /a/ during the preceding segment, but this reflects an early default bias toward the ʌ/ə response.

Among tense vowels, the more diphthongized a vowel is, the later accuracy improves. Thus, nearly steady-state /i, u, a/ as Segment 1 begin at approximately their maximum accuracy at Gate 1, and as Segment 2, they show improvement in accuracy before any other vowels, with the steepest increase usually from Gate 3 to 4. The diphthongized mid tense vowels /e^{j}, o^{w}/ improve next, with the most increase in accuracy from Gate 1 to 2 for Segment 1 and Gate 3 to 5 for Segment 2. Diphthongs (/a^{j}, a^{w}, o^{j}/) show the latest improvement in accuracy, primarily from Gate 1 to 3 as Segment 1 and Gate 4 to 5 or 4 to 6 as Segment 2. This is consistent with the findings for affricates above: listeners will not hypothesize that additional perceptual cues to the segment they are currently perceiving might yet happen. (For example, /a^{j}/ as Segment 1 at Gate 1 is misperceived as /a/ in 63.4% of responses, and as /æ/ in 12.8%, and correctly perceived in only 8.4% of responses.) However, once the cues that distinguish inherently changing segments such as diphthongs or affricates from more stable segments become available, perception shifts rapidly to the changing segment.

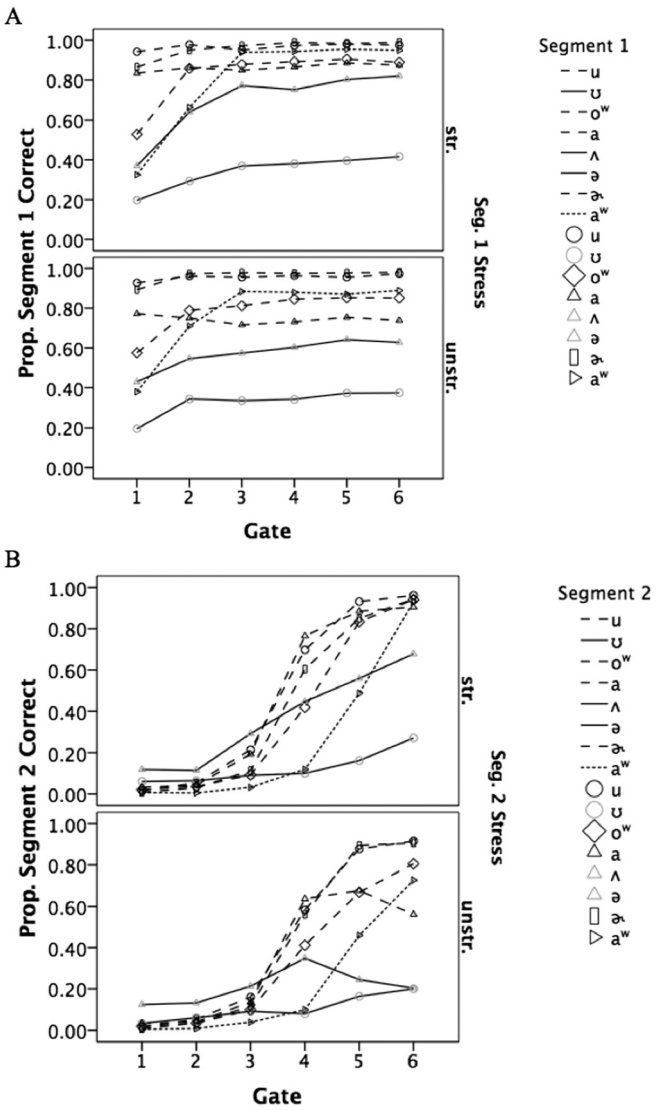

FIG. 6. Proportion correct over gate point for back vowels and back-ending diphthongs. (A) As Segment 1. (B) As Segment 2.

What is perhaps less expected is that the three degrees of diphthongization in English are distinguished by when the improvement occurs during the segment, rather than only true diphthongs differing from other vowels.

Figure 7 displays the effect of stress, averaged within tense vowels (/i, e^j, a, o^w, u, ɚ/), lax vowels (/ɪ, ɛ, æ, ʊ, ʌ, ə/), and diphthongs (including only /aj, a^w, o^j/). Stressed vowels are generally perceived more accurately than unstressed (Table VII), with this effect becoming significant during the last third of the vowel for Segment 1 (where there is often no preceding context), and during the last third of the preceding segment for Segment 2. The effect of stress develops slowly and is greater for lax than for tense vowels or diphthongs. The size of the effect increases at the end of the vowel (Gate 3 or 6), then remains stable throughout the following segment (for vowels as Segment 1). In all cases, stressed and unstressed vowels are perceived equally well (or poorly) early on when little information is available, but perception improves more for stressed than for unstressed vowels. Unstressed vowels in English are centralized

TABLE VI. F ratios (Bonferroni-corrected significant comparisons only) for vowel tenseness with rationalized arcsine transformed proportions. Direction of effect in column headers: significance level: $p < 0.00208$ (Bonferroni correction with 24 comparisons), df: (1,19) for each comparison.

	Segment 1				Segment 2			
Gate Stressed	i > ɪ	$e^j > ɛ$	u > ʊ	a > ʌ	i > ɪ	$e^j > ɛ$	u > ʊ	a > ʌ
1	143.25		208.51	67.98				
2	180.82	17.92	172.41	39.54				
3	131.20	16.36	106.09				17.28	
4	113.90	16.98	116.09		110.48		158.75	25.26
5	57.71	14.88	116.01		81.17	24.78	373.74	24.81
6	55.32	15.90	81.31		82.63	35.89	205.08	16.98
Unstressed	i > ɪ	$e^j > ɛ$	u > ʊ	a > ə	i > ɪ	$e^j > ɛ$	u > ʊ	a > ə
1	107.78		200.92	48.66				
2	92.13		99.59	22.25				
3	115.70	36.66	131.14	19.66				
4	127.97	43.02	96.54		99.05		188.46	22.81
5	120.30	17.04	68.18		119.99	17.18	234.61	97.70
6	302.71		83.32		219.44	87.41	192.60	114.59

(Fourakis, 1991), making them all less distinct, especially the already quite central lax vowels. However, the timing of the stress effect shows that stressed and unstressed vowels begin with equal acoustic information, but stressed vowels add more information later in the vowel.

Accuracy for some unstressed vowels as Segment 2 even decreases from Gate 5 to 6, at the end of the vowel. Since all diphones were recorded with following context to prevent final lengthening, but this final context was always removed, all Segment 2 vowels contain coarticulatory cues to the absent following sound. The increase in the size of the stress effect at Gate 6 indicates that, in an unstressed vowel, listeners find it very difficult to separate perceptual cues to a vowel from those to a following consonant, particularly when the following consonant is not available to disambiguate the cues.

Unstressed lax vowels were difficult to perceive. Even by Gate 6, unstressed Segment 1 lax vowels average <60% correct responses, with /æ, ʊ/ perceived especially poorly, and only /ɪ, ɛ/ over 65% correct. Misidentifications usually involved another lax vowel nearby in the vowel space, less often the corresponding tense vowel (e.g., /ʊ/ as Segment 1 at Gate 6 was reported as /ʌ/ in 51.0% of responses, as /u/ in 3.9%, and as /o^w/ in 3.4%, and was correctly perceived in 37.4% of responses).

IV. DISCUSSION

The transmission of information about segments over time differs across segment types. For some segments, especially affricates and diphthongs, information is highly localized so that listeners show a sudden improvement in perceptual accuracy during a particular time range, but little improvement before or after it. For other segments, most clearly stops and sonorants, relatively more information spreads into the preceding segment, or the information

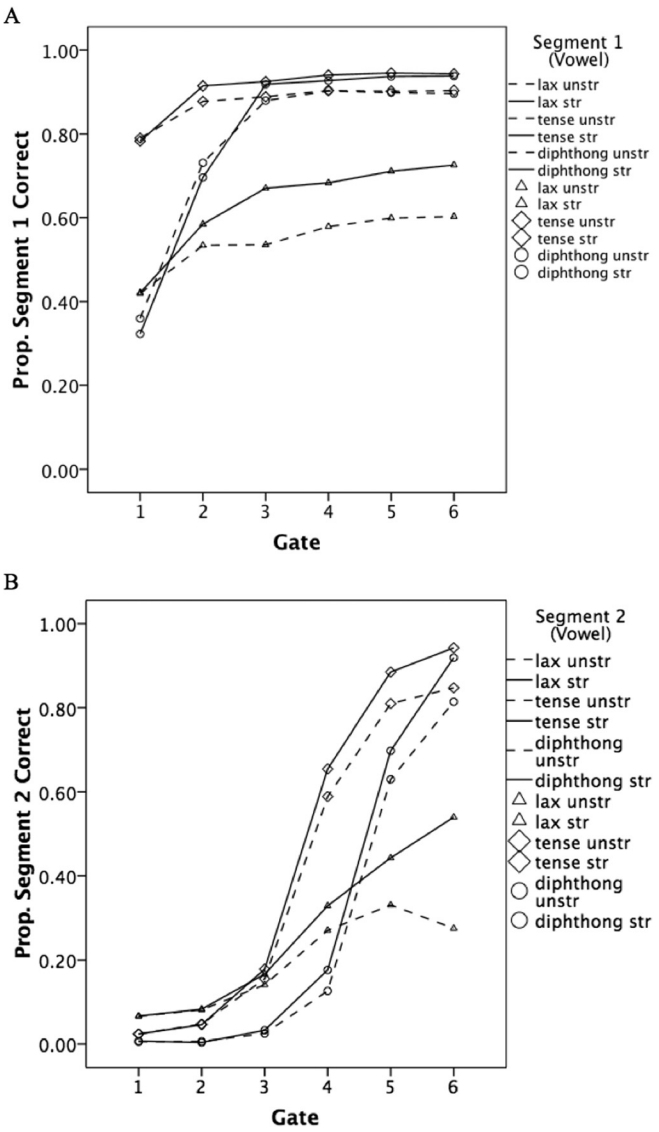

FIG. 7. Proportion correct over gate point for stressed and unstressed vowels, for tense vowels, lax vowels, and diphthongs. (A) As Segment 1. (B) As Segment 2.

becomes available gradually throughout the segment itself. For many segments, perception continues to improve slightly during the following segment, indicating that some additional perceptual cues become available late.

Thus, information about particular speech sounds differs both in its degree of localization and its spread across time. While it is generally true that information in speech is spread across segments, the amount of overlap is known to vary as a function of segment identity (Furui, 1986; Smits *et al.*, 2003). Our analyses show also how the degree of localization of perceptual information depends on each speech sound's identity and phonological class. The phonological class of a neighboring segment also clearly influences how much acoustic information can spread into it (so far more information about a /b/ is available by early gates of the diphone /ab/ than of the diphone /fb/).

Even within phonological class, however, listeners' ability to perceive a segment sometimes depends on segment-specific properties. For example, lax vowels, particularly

TABLE VII. F ratios (Bonferroni-corrected significant comparisons only) for stress (stressed more accurate than unstressed) with rationalized arcsine transformed proportions. Significance level: $p < 0.00278$ (Bonferroni correction with 18 comparisons), df: (1,19) for each comparison.

	Segment 1			Segment 2		
Gate	Tense	Lax	Diphthongs	Tense	Lax	Diphthongs
1						
2	48.41					
3	101.84	117.50	26.31	23.15	14.31	
4	56.15	97.17	13.67	74.93	14.86	32.58
5	68.31	87.12	39.67	78.55	54.99	37.44
6	75.39	253.66	53.03	181.35	145.55	217.44

when unstressed, are, in general, poorly perceived, but /ʊ, æ/ are perceived very poorly even within this group, while in contrast, /ɪ, ɛ/ are perceived relatively well. Among fricatives, /ð/ is perceived badly, but /θ, ʒ/ are perceived less poorly. Among tense vowels, /i, u/ are perceived well, but /a/ is not. The poor perception of particular segments could have a number of different causes, such as a marginal or borrowed status in the phonemic inventory (a possible explanation in the case of /ʒ/), low frequency or limited distribution of the phoneme in the lexicon overall (/ʊ, ð/), lack of an orthographic convention for the sound (/ʊ, ʒ/), or just acoustic similarity to nearby sounds (/æ/, given /ɛ/). Only further research can provide the correct explanation for these patterns, and indeed for many other patterns in our data.

Despite this evidence of segment-specificity, many phonological class effects appear. The highly localized information for affricates and diphthongs vs more gradual information for monophthongs, stops, and sonorants is a class-general effect. Another is the timing of perception of voiced vs voiceless stops, with better perception of voiceless stops during the preceding segment and once the release burst has been heard, but better perception of voiced stops during the closure. The plateau in accuracy improvement that voiceless stops show during the two gates of the stop closure is another feature of a whole phonological class, and the fact that voiced stops show a lessening of slope of improvement at the same time (a less severe form of the plateau) shows how the pattern among voiceless stops has a related pattern within the broader phonological class of stops. Similarly, despite the idiosyncratic behavior of the lax vowels /ʊ, æ/, there is also a general pattern of lax vowels being perceived less well than their tense counterparts. Accuracy of perception of a given sound thus stems from both general patterns over entire phonological classes and segment-specific effects.

The strong effect of stress in English was also evident in our data. English unstressed vowels show reduction even when they are full vowels rather than schwa (Fourakis, 1991; Fear *et al.*, 1995); as our results show, this reduction has a significant impact on perceptual accuracy.

A final phonological generalization concerns a consonant-vowel difference. Both consonants and vowels were perceived more accurately when adjacent to a vowel rather than a consonant. For perception of Segment 1 even after hearing the following segment (at Gate 6), the initial

consonant is perceived correctly in 90% of CV stimuli, but only 84% of CC stimuli, and an initial stressed vowel is perceived correctly in 93% of VV stimuli, but only 86% of VC stimuli (all vowels stressed). So, in general, coarticulatory information about consonants in vowels helps consonant perception, but coarticulation with consonants hinders vowel perception (a pattern held to underlie listeners' greater willingness to alter initial decisions about vowels than about consonants; Van Ooijen, 1996).

V. CONCLUSIONS

The dataset described here shows how listeners perceive speech sounds over time for all sounds of American English in all possible environments. Acoustic information, and hence perceptual cues, are shown to be distributed through the speech signal differentially over time, with the precise timing of the distribution depending on phonological categories, specific segment identities, and stress.

The present work, and the associated publicly available complete dataset (http://www.u.arizona.edu/~nwarner/WarnerMcQueenCutler.html), allows comparison of the timing of perception for any English segment preceded or followed by any other possible English segment. All diphones were tested with the same experimental methods, pronounced by the same speaker, and heard by the same listeners. This degree of comparability across a whole language repertoire could never be reached by meta-analyses of studies of a specific set of segments or diphones. The scale and comparability of this dataset thus allows current and future researchers to answer a wide variety of questions about speech perception. It also allows modeling of spoken word recognition in English with probabilistic data about how likely listeners are to think they are hearing a given sound at a given point in time, not just with a "toy" lexicon, but with the entire English lexicon. This use of the dataset will be implemented in a forthcoming release for English of the Bayesian probabilistic model of continuous speech recognition Shortlist B (Norris and McQueen, 2008, currently implemented for Dutch using the dataset of Smits *et al.*, 2003). Of course, the data could equally well be used as input to other models. To conclude, the dataset provides a way for researchers to answer questions about both spoken word recognition and speech perception in English, without the need to collect large sets of new data for each question.

ACKNOWLEDGMENTS

The project was supported by a special grant from the Max Planck Society. The authors thank Priscilla Shin and Maureen Hoffmann for their work on this project.

Benkí, J. (**2003**). "Analysis of English nonsense syllable recognition in noise," Phonetica **60**, 129–157.

Fear, B. D., Cutler, A., and Butterfield, S. (**1995**). "The strong/weak syllable distinction in English," J. Acoust. Soc. Am. **97**, 1893–1904.

Fourakis, M. (**1991**). "Tempo, stress, and vowel reduction in American English," J. Acoust. Soc. Am. **90**, 1816–1827.

Furui, S. (**1986**). "On the role of spectral transition for speech perception," J. Acoust. Soc. Am. **80**, 1016–1025.

Greenberg, S. (**1999**). "Speaking in shorthand—A syllable-centric perspective for understanding pronunciation variation," Speech Commun. **29**, 159–176.

Hillenbrand, J. M., and Nearey, T. M. (**1999**). "Identification of resynthesized /hVd/ utterances: Effects of formant contour," J. Acoust. Soc. Am. **105**, 3509–3523.

Jesse, A., and Massaro, D. W. (**2010**). "The temporal distribution of information in audiovisual spokenword identification," Atten. Percept. Psychophys. **72**, 209–225.

Jongman, A., Wang, Y., and Kim, B. (**2003**). "Contributions of sentential and facial information to perception of fricatives," J. Speech Lang. Hear. Res. **46**, 1367–1377.

Jongman, A., Wayland, R., and Wong, S. (**2000**). "Acoustic characteristics of English fricatives," J. Acoust. Soc. Am. **108**, 1252–1263.

Ladefoged, P., and Johnson, K. (**2015**). *A Course in Phonetics* (Cengage Learning, Stamford, CT), Chap. 4, pp. 89–114.

McQueen, J. M. (**1995**). "Processing versus representation: Comments on Ohala and Ohala," in *Phonology and Phonetic Evidence. Papers in Laboratory Phonology IV*, edited by B. Connell and A. Arvaniti (Cambridge University Press, London), pp. 61–67.

Miller, G. A., and Nicely, P. E. (**1955**). "An analysis of perceptual confusions among some English consonants," J. Acoust. Soc. Am. **27**, 338–352.

Nishi, K., Lewis, D. E., Hoover, B. M., Choi, S., and Stelmachowicz, P. G. (**2010**). "Children's recognition of American English consonants in noise," J. Acoust. Soc. Am. **127**, 3177–3188.

Norris, D., and McQueen, J. M. (**2008**). "Shortlist B: A Bayesian model of continuous speech recognition," Psych. Rev. **115**, 357–395.

Ohala, J. J., and Ohala, M. (**1995**). "Speech perception and lexical representation: The role of vowel nasalization in Hindi and English," in *Phonology and Phonetic Evidence. Papers in Laboratory Phonology IV*, edited by B. Connell and A. Arvaniti (Cambridge University Press, London), pp. 41–60.

Öhman, S. E. (**1966**). "Perception of segments of VCCV utterances," J. Acoust. Soc. Am. **40**, 979–988.

Peláez-Moreno, C., García-Moral, A. I., and Valverde-Albacete, F. J. (**2010**). "Analyzing phonetic confusions using formal concept analysis," J. Acoust. Soc. Am. **128**, 1377–1390.

Peterson, G. E., and Barney, H. L. (**1952**). "Control methods used in a study of the vowels," J. Acoust. Soc. Am. **24**, 175–184.

Phatak, S. A., Lovitt, A., and Allen, J. B. (**2008**). "Consonant confusions in white noise," J. Acoust. Soc. Am. **124**, 1220–1233.

Pisoni, D. B., Nusbaum, H. C., Luce, P. A., and Slowiaczek, L. M. (**1985**). "Speech perception, word recognition and the structure of the lexicon," Speech Commun. **4**, 75–95.

Pols, L. C. W., and Schouten, M. E. H. (**1978**). "Identification of deleted consonants," J. Acoust. Soc. Am. **64**, 1333–1337.

Sherbecoe, R. L., and Studebaker, G. A. (**2004**). "Supplementary formulas and tables for calculating and interconverting speech recognition scores in transformed arcsine units," Int. J. Audiol. **43**, 442–448.

Smits, R. (**2000**). "Temporal distribution of information for human consonant recognition in VCV utterances," J. Phonetics **28**, 111–135.

Smits, R., Warner, N., McQueen, J. M., and Cutler, A. (**2003**). "Unfolding of phonetic information over time: A database of Dutch diphone perception," J. Acoust. Soc. Am. **113**, 563–574.

Van Ooijen, B. (**1996**). "Vowel mutability and lexical selection in English: Evidence from a word reconstruction task," Mem. Cognit. **24**, 573–583.

Wang, M. D., and Bilger, R. C. (**1973**). "Consonant confusions in noise: A study of perceptual features," J. Acoust. Soc. Am. **54**, 1248–1266.

Warner, N., Smits, R., McQueen, J. M., and Cutler, A. (**2005**). "Phonological and statistical effects on timing of speech perception: A database of Dutch diphone perception," Speech Commun. **46**, 53–72.