Journal of Experimental Psychology: General
2005, Vol. 134, No. 4, 477–500

Copyright 2005 by the American Psychological Association
0096-3445/05/$12.00 DOI: 10.1037/0096-3445.134.4.477

Integration of Multiple Speech Segmentation Cues: A Hierarchical Framework

Sven L. Mattys, Laurence White, and James F. Melhorn
University of Bristol

A central question in psycholinguistic research is how listeners isolate words from connected speech despite the paucity of clear word-boundary cues in the signal. A large body of empirical evidence indicates that word segmentation is promoted by both lexical (knowledge-derived) and sublexical (signal-derived) cues. However, an account of how these cues operate in combination or in conflict is lacking. The present study fills this gap by assessing speech segmentation when cues are systematically pitted against each other. The results demonstrate that listeners do not assign the same power to all segmentation cues; rather, cues are hierarchically integrated, with descending weights allocated to lexical, segmental, and prosodic cues. Lower level cues drive segmentation when the interpretive conditions are altered by a lack of contextual and lexical information or by white noise. Taken together, the results call for an integrated, hierarchical, and signal-contingent approach to speech segmentation.

Keywords: speech segmentation, word recognition, rhythm, phonotactics, coarticulation

The inspection of a speech waveform does not reveal clear correlates to what the human ear perceives as word boundaries (Klatt, 1980). The acoustic signal of a typical utterance contains few noticeable interruptions, and these are far less reliable indicators of word boundaries than are blank spaces in the written language. Yet the continuous nature of speech hardly seems to pose a problem for everyday listening, as the subjective experience of speech is not of continuity but of discreteness—that is, a string of words. The passage from continuity to discreteness constitutes a well-known challenge to psychologists and computer scientists alike. It also reaches far beyond the speech domain, as segmentation problems have been amply documented for nonspeech auditory stream segregation (e.g., Bregman, 1999) and visual scene analysis (e.g., Palmer, 1999). Likewise, within the speech sciences, segmentation constrains not only word recognition but also sentential parsing and, ultimately, discourse comprehension.

At the word level, the focus of this article, scientists have traditionally tackled speech segmentation from one of two standpoints. According to one position, segmentation is based on multiple sublexical cues probabilistically associated with word boundaries (e.g., Christiansen, Allen, & Seidenberg, 1998). Cues that have been shown to be successfully exploited by listeners include metrical stress, phonotactic regularities, and acoustic–phonetic variants (for a review, see, e.g., Davis, Marslen-Wilson, & Gaskell, 2002).

An alternative stance posits segmentation as the product of word recognition rather than one of its prerequisites (e.g., McClelland & Elman, 1986; Norris, 1994). With multiple lexical activation as a core principle, lexically driven segmentation is achieved when competition between candidates—activated sequentially (e.g., Cole & Jakimik, 1980; Grosjean & Gee, 1987; Luce & Lyons, 1999; Marslen-Wilson, 1984) or in parallel (e.g., Frauenfelder & Peeters, 1990; McClelland & Elman, 1986; Norris, 1994)—settles on a lexically acceptable parsing solution. In this view, for instance, the string /hikɔldɪmidɪətlɪ/ is parsed as *he called immediately*, not because of the intervention of sublexical cues but because this parsing solution is the only one that does not leave any fragments lexically unaccounted for.

Taken individually, both approaches have limitations. Sublexical views usually fail to take advantage of lexical and sentential contributions to segmentation and rarely have a provision for conflicting or incorrect sublexical segmentation cues. For example, phonotactic regularities make it difficult for a high-probability within-word diphone like /st/ to be interpreted as a word boundary in /gæstænk/, even though lexical information makes *gas tank* the only acceptable segmentation solution. Likewise, the acoustic–phonetic realization of /nɑɪtreɪt/ as *nitrate* versus *night rate* is likely to be subordinate to the semantic context in which the stimulus is heard (e.g., in a pharmacy vs. a parking garage).

The limitations of lexically driven segmentation are in the opposite direction. First, lexically driven segmentation produces a great deal of superfluous lexical activity (e.g., embedded words, boundary-straddling candidates) because it usually overlooks the potential of naturally occurring statistical regularities about word boundaries. Second, word knowledge can prove inefficient in cases of ambiguous lexical interpretations (e.g., *known ocean* vs.

Sven L. Mattys, Laurence White, and James F. Melhorn, Department of Experimental Psychology, University of Bristol, Bristol, England.

This research was supported by Biotechnology and Biological Sciences Research Council Grant 7/S18783 to Sven L. Mattys. We thank Jeff Bowers, Markus Damian, Colin Davis, and Clive Frankish for their comments on drafts; Vivienne Long for recording the stimuli of Experiment 3; and Clive Frankish for recording the utterances of Experiments 1 and 2. All other stimuli were recorded by James F. Melhorn. We also thank Anne Cutler and Marilyn Vihman for their comments on drafts of the article. Special thanks are due to Julie Liss for providing the impetus for this work.

Correspondence concerning this article should be addressed to Sven L. Mattys, Department of Experimental Psychology, University of Bristol, 8 Woodland Road, Bristol BS8 1TN, United Kingdom. E-mail: sven.mattys@bris.ac.uk

no notion) or lexical embeddedness (e.g., *sat in* vs. *satin*), even if we apply the minimal accretion principle (i.e., a bias for longer lexical candidates; e.g., Swinney, 1981). Finally, lexically driven segmentation breaks down when confronted with speech materials that do not map onto stored representations, which makes it a suboptimal principle for language acquisition.

Although the need for a compromise between the two approaches is generally acknowledged, models that have outlined the details of such a compromise are scarce. To date, one of the most successful, empirically supported models merges multiple lexical activation (Norris, 1994) and stress-based segmentation (Cutler & Norris, 1988) by favoring stress-initial candidates in the competition process (McQueen, Norris, & Cutler, 1994; Norris, McQueen, & Cutler, 1995). Similarly, in their good start model, Gow and Gordon (1995) posited that the discovery of word boundaries is primarily lexically driven but that sublexical specifications such as allophones and stress can facilitate—but not inhibit—lexical activation by making some locations of the signal more salient. Such a facilitatory influence gets lexical processing off to a "good start," not unlike the lexical boost provided by stress in Norris's model. Independently, Sanders and Neville (2000) noted trading relations among syntactic, lexical, and metrical segmentation cues as a function of their availability in the signal. Stress tended to have a stronger impact when other cues were absent, although it could not completely compensate for the lack of lexical and syntactic word-boundary information. Compromises between lexically driven and cue-driven segmentation have also been documented in computer simulations (e.g., Grossberg & Myers, 2000; Norris, McQueen, Cutler, & Butterfield, 1997).

Segmentation cues have often been investigated individually, with any competing or complementary cues either controlled or randomized. Given that the perceptual system is shown to capitalize on any functional contingencies present in the environment (e.g., Gómez, 2002; Seligman, 1970), even after only brief exposure (Chambers, Onishi, & Fisher, 2003), it is not surprising to find evidence for an increasingly long list of word-boundary cues when these are tested in isolation and in controlled laboratory conditions. In natural speech settings, however, the unique contribution of each cue is less clear, as all sources of information usually converge toward a single interpretation. Moreover, in the same way that Gestalt principles studied in isolation may sometimes misrepresent the complexity of real-life perception (Palmer, 1999; see also Vecera, Vogel, & Woodman, 2002), a cue-by-cue approach to speech segmentation is likely to circumscribe our understanding of the phenomenon. A more complete model should specify the strategies that tend to dominate when multiple cues are available in the signal. Finally, segmentation accounts have often been restricted to listening conditions in which the acoustic details of the signal are fully intelligible. However, Mattys (2004) showed that the introduction of a noisy background can have dramatic consequences on the relative reliance on individual cues. Because such degraded conditions are often encountered in natural speech environments, segmentation models would also improve their validity if they accounted for segmentation behavior in noise.

We start our investigation with an analysis of how one of the most widely documented sublexical cues, word stress, fares against alternative sources of information—coarticulation, phonotactics, and lexical knowledge—and of how background noise alters segmentation strategies.

The Case of Stress

Segmentation based on metrical prosody has been documented in a large number of studies (see Cutler, Dahan, & van Donselaar, 1997, for a review). The basic principle is that languages with a significant bias in the distribution of word stress provide listeners with a powerful segmentation tool. In English and Dutch, for instance, most content words are stress initial (Cutler & Carter, 1987; Vroomen & de Gelder, 1995); thus, treating strong syllables as word onsets is a potentially efficient segmentation heuristic.

Even in such languages, however, the direct contribution of word stress to segmentation and lexical activation remains unclear. In a study by Cutler and Norris (1988), word spotting (e.g., *mint*) was found to be slowed down if the target word overlapped with a subsequent strong syllable (e.g., *mintayve*) compared with a weak one (e.g., *mintesh*). Thus, the presence of a strong syllable interfered with the detection of the utterance-initial word. However, whether this effect can be related to the initiation of lexical access is not entirely clear, because the segmentation point was the offset of the test word; its onset was always confounded with that of the utterance. A later experiment (Norris et al., 1995) suggested that strong syllables did indeed involve lexical initiation because the interference found in the Cutler and Norris study was modulated by the number of potential competitors beginning with the strong syllable, but the evidence was, again, gathered from the spotting latencies of the utterance-initial word.

More direct evidence for stress-based segmentation has been obtained in impoverished interpretive conditions, such as faint or noisy speech. For example, when listeners were presented with spoken sentences played in noise (Smith, Cutler, Butterfield, & Nimmo-Smith, 1989) or at reduced intensity (Cutler & Butterfield, 1992), their lexical misperceptions involved primarily stress-initial interpretations (e.g., *achieve* being misperceived as *a cheap*). Hypokinetic dysarthria, a speech disorder resulting in imprecise, low-intelligibility articulation, has also been shown to elicit stress-based missegmentation in normal listeners (Liss, Spitzer, Caviness, Adler, & Edwards, 1998), with the magnitude of stress-based patterns a function of strength contrastivity (Liss, Spitzer, Caviness, Adler, & Edwards, 2000). These studies, along with computer simulations (e.g., Harrington, Watson, & Cooper, 1989), suggest that stress is especially useful when acoustic–phonetic information is impoverished.

Direct comparisons between stress and other segmentation cues are scarce. Those few studies that have considered stress alongside segmental cues—for example, phonotactics and vowel harmony—have not provided a consistent picture. In some cases, stress has been described as a secondary cue or a cue emerging from phonotactic regularities (e.g., Cairns, Shillcock, Chater, & Levy, 1997; McQueen, 1998); in others, it has been described as equipotent or dominant (e.g., Norris et al., 1997; Suomi, McQueen, & Cutler, 1997; Vitevitch, Luce, Charles-Luce, & Kemmerer, 1997; Vroomen, Tuomainen, & de Gelder, 1998). No firm conclusion can be drawn, because none of these studies was explicitly designed to evaluate the relative impact of stress on segmentation. In the following experiments, the role of stress on segmentation is evaluated when it is pitted against acoustic–phonetic cues (Experiment 1A), phonotactics (Experiment 2), and lexicality (Experiment 3).

Experiment 1A: Stress Versus Acoustic–Phonetic Cues

In this experiment, we assess the contribution of acoustic–phonetic cues to segmentation by manipulating the degree of coarticulation at potential word boundaries. Fougeron and Keating (1997) found that segments at the edges of prosodic domains (e.g., words and phrases) have more extreme lingual articulation and exhibit less overlap with adjacent segments than those within domains. Listeners exploit these associations to segment words from fluent speech (Mattys, 2004), and such sensitivity is already noticeable in early infancy. Johnson and Jusczyk (2001), using concatenation to simulate low coarticulation, found that 8-month-olds treated concatenation points as word onsets, even when these cues conflicted with distributional information. Similarly, Mattys (2004) noticed that, when coarticulation and stress were pitted against each other, adult listeners relied more heavily on coarticulatory discontinuities than strong syllables to initiate lexical activation. When the signal was played in a background of noise, however, stress-initial primes were more efficiently segmented than stress-noninitial primes, despite the conflicting coarticulation. Experiment 1A is an attempt to replicate and extend the stress-versus-coarticulation result using a fully orthogonal design applicable not only to coarticulation but also to phonotactics (Experiment 2) and lexicality (Experiment 3). All three experiments are based on a cross-modal fragment priming paradigm (e.g., Mattys, 2004; van Donselaar, Koster, & Cutler, 2005). Participants performed a lexical-decision task on a visual target displayed after the playback of a nonsense utterance. On test trials, the later portion of the utterance served as a phonological prime to the target (e.g., the utterance /revə‘mærə/, with the priming fragment underlined, and the target *marathon*). The extent to which the stress pattern of the prime and coarticulatory, phonotactic, and lexical cues in the utterance affected the degree of priming was estimated relative to a baseline condition. Stimuli were played in their intact format and in noise.

Method

Participants. The participants in all of the experiments reported in this study were native speakers of British English and were undergraduate or graduate students at the University of Bristol. They received course credit or a small honorarium for their participation in the experiments. None reported a history of speech or hearing difficulties. For Experiment 1A, 61 participants were randomly assigned to the intact (n = 30) or noise (n = 31) condition.

Materials. Forty trisyllabic words were chosen; 20 of them had primary stress on the initial syllable, the other 20 on the second syllable. The two sets of words were matched pairwise on their initial phoneme and frequency of occurrence. With their last syllable removed (see *Design and procedure* section), these words constituted 20 strong–weak (SW) primes (e.g., /’mærə/, from *marathon*) and 20 weak–strong (WS) primes (e.g., /mə‘tɪ/, from *material*). None of the primes was a word. Many of these bisyllables reached their uniqueness point before their offset. The average cohort size at the offset of the primes was 2.00 (SD = 2.64) for the SW primes and 0.80 (SD = 1.47) for the WS primes, $t(19) = 1.72$, $p = .10$. The average number of words containing the strong syllable in any position was 31.00 (SD = 15.83) for the SW primes and 40.00 (SD = 36.98) for the WS primes, $t(19) = -1.28$, $p = .22$. Average neighborhood density, estimated as the number of words departing from the prime by a one-phoneme substitution, deletion, or addition in any position (Luce & Pisoni, 1998), was 1.00 (SD = 1.65) for the SW primes and 0.50 (SD = 1.10) for the WS primes, $t(19) = 1.60$, $p = .13$. All lexical statistics were estimated from CELEX (Baayen, Piepenbrock, & Gulikers, 1995).

Each test prime was embedded at the end of a nonsense utterance, for example, /revə‘mærə/. The speech fragment preceding the prime, called *context*, was a nonsense disyllable with an SW or WS stress pattern. To ensure phonotactic legality, many of the disyllables were fragments of existing words. To minimize context–prime pairing idiosyncrasies, we created two sets of utterances. In one set, a given prime was preceded by an SW context. In the other, the same prime was preceded by a WS context. Participants were randomly assigned to one set or the other (see Appendix A).

The onset of the prime was either decoarticulated (i.e., through concatenation) or coarticulated with the end of the context. Thus, the acoustic–phonetic cues in the utterances either favored or disfavored segmentation before the prime’s onset. To balance overall coarticulation in the two conditions, the “unfavorable” acoustic–phonetic cues condition—in which the prime onset was coarticulated with the prior context—contained a concatenation point between the first and second syllables of the context (see an illustration in Table 1).

A further set of utterances served as baseline for the estimation of priming effects. We created each baseline utterance from a test utterance by replacing the prime (i.e., the last two syllables) with distorted speech. To prevent coarticulatory information on the onset of the prime from being available at the end of the context, we ensured that the context of the baseline utterances was that of the concatenated condition. We created the distorted speech by digitally superimposing several disyllabic SW and WS primes. The resulting fragment sounded like scrambled speech, with no identifiable segmental or suprasegmental characteristics. The purpose of the baseline was to offset any lexical-decision idiosyncrasies from the initial-stress and noninitial-stress visual words, hence making priming results comparable in the SW and WS conditions.

To vary the position of the prime within the utterances, we created 40 filler utterances in which the first two syllables or middle syllables constituted the prime. Like the test utterances, the filler utterances came in

Table 1
Examples of Test Utterances in Experiments 1, 2, and 3

	Stress pattern of the prime	
	SW prime	WS prime
Experiment 1A: Stress Versus Acoustic-Phonetic Cues		
Target	marathon	material
Fav. acoustics	/revə-'mærə/	/revə-mə'tɪ/
Unfav. acoustics	/re-və'mærə/	/re-və'mætɪ/
Experiment 2: Stress Versus Phonotactics		
Target	customer	cathedral
Fav. phonotac.	/gɑstem'kʌstə/	/gɑstemkə'θi/
Unfav. phonotac.	/gɑsteŋ'kʌstə/	/gɑsteŋkə'θi/
Experiment 3: Stress Versus Lexicality		
Target	versatile	victorian
Word context	/ɪ'nɔrməs'vɜsə/	/kɑg'nɪʃənvɪk'tɔ/
Nonword context	/ə'reɪməs'vɜsə/	/əm'bəɪʃənvɪk'tɔ/

Note. For Experiment 1A, decoarticulation points are represented by dashes. For Experiment 2, critical diphones are underlined (e.g., /mk/, low frequency; /ŋk/, high frequency). For Experiment 3, the word contexts in the example are *enormous* and *cognition.* SW = strong–weak; WS = weak–strong; Fav. = favorable; Unfav. = Unfavorable; phonotac. = phonotactics.

three versions: One contained a concatenation point at the start of the prime, the second contained a concatenation point in another location, and the third featured the prime in a distorted format (similar to the baseline test utterances). The stress pattern of the prime was balanced across all types of fillers.

Design and procedure. On all trials, the visual target appeared on a computer monitor 100 ms after the utterance offset. The 100-ms delay was motivated by prior results (Mattys, 2004) showing maximum form priming effects for that delay compared with earlier alignments. Each of the 40 test target words was presented in three utterance conditions: favorable acoustic–phonetic cues, unfavorable, and baseline, for a total of 120 test trials. A similar breakdown applied to the filler utterances. All test and filler utterances were also followed by visual target nonwords. These were trisyllabic orthotactically legal letter strings matched with the word targets on their average number of letters. Thus, the total number of trials was 480. Sixteen original utterances were prepared for practice, representing the various trial conditions.

The utterances were recorded in a sound-attenuated booth by a male native speaker of southern British English. These recordings included the entire trisyllabic words instead of the disyllabic primes. The speaker was instructed to pronounce the two fragments one after the other: for example, /revə/ and /’mærəΘən/ (concatenated condition) or /re/ and /və‘mærəΘən/ (coarticulated condition), with a brief pause between them. The speaker closed his mouth during the pause between the two fragments to eliminate coarticulation (see Appendix B for acoustic measurements). We concatenated the two fragments by editing out the pause. In an attempt to minimize any obvious artifacts associated with splicing (e.g., clipping, rhythm or pitch discontinuities), we recorded several renditions of the fragments and only kept those resulting in natural-sounding utterances, as judged by Sven L. Mattys and James F. Melhorn. Average intensity of the primes was 64 dB and 65 dB for the concatenated and coarticulated conditions, respectively, for the SW primes and 66 dB and 66 dB, respectively, for the WS primes. Mattys (2004) showed that the acoustic properties of stimuli created with the above decoarticulation–concatenation procedure did not notably depart from the properties of those containing naturally decoarticulated boundaries. Once the concatenated utterances were assembled, the last syllable of the trisyllabic word was removed. We created the baseline utterances by concatenating the context (e.g., /’revə/) and the scrambled disyllable.

The utterances were digitized (16 bit A/D) at 32 kHz. On output, the utterances were converted to analog form (16 bit D/A, 32 kHz) and delivered over good-quality headphones. The average intensity level of the utterances was 65 dB. The utterances were presented either intact or in noise to two different sets of participants. We created the noise condition by playing the utterances in a background of noise, generated with a spectral frequency of 1/f within the bandwidth 0–15 kHz, with a signal-to-noise ratio (SNR) of approximately −5 dB (the noise intensity was 70 dB), measured against the average signal intensity of the nonsilent portions of the signal. As a reference point, an SNR of 0 dB is known to reduce the intelligibility of isolated words to about 50% (Acton, 1970) and one of −10 dB to prevent any segmental identification beyond a mere speech–nonspeech discrimination (Erber, 1971). There was no noise between utterances.

Trials were pseudorandomized; two identical utterances, primes, or targets in a row were presented, and the number of consecutive SW or WS primes was limited to three. Participants were tested individually in a quiet room, seated in front of a computer monitor and wearing headphones. They were told that an utterance would be played over the headphones on each trial and a string of letters would be presented on the computer monitor after the utterance. They were instructed to decide whether the letter string was a word, using a two-button response box. To those participants assigned to the noise condition, a warning was made about the relative unintelligibility of the utterances.

On each trial, an utterance was played, and a visual target appeared in the center of a computer monitor 100 ms after utterance offset. Participants had 3 s, measured from target presentation onset, to give their response. Following the response or at the end of the 3-s response, there was a 1.5-s pause before the onset of the next utterance.

Results and Discussion

Lexical decision latencies were measured from the onset of visual target presentation. Incorrect responses to word targets and correct responses two standard deviations from the mean (computed separately for each participant) were discarded. Altogether, the discarded responses amounted to 10.0% of the test trials in the intact condition (6.1% incorrect) and 7.5% in the noise condition (3.7% incorrect). Mean lexical decision latencies and accuracy levels are reported in Table 2. Priming effects, plotted in Figure 1, were calculated as the difference between the baseline and test conditions.

The data show a sharp contrast between the intact and noise conditions. In intact speech, priming was facilitated by favorable

Table 2
Lexical-Decision Latencies and Percentage Correct in Experiments 1, 2, and 3

Experiment and condition	Intact				Noise			
	SW prime		WS prime		SW prime		WS prime	
	Latency	% correct	Latency	% correct	Latency	% correct	Latency	% correct
Experiment 1								
Fav. acoustics	513	95	525	95	523	97	564	97
Unfav. acoustics	529	95	545	93	526	98	560	96
Baseline	562	94	577	92	551	95	559	96
Experiment 2								
Fav. phonotac.	470	95	483	94	529	95	576	89
Unfav. phonotac.	490	96	499	95	536	95	576	91
Baseline	556	92	569	88	564	94	577	90
Experiment 3								
Word context	521	97	517	98	528	97	544	95
Nonword context	540	97	542	98	531	97	554	96
Baseline	586	96	585	96	560	95	547	94

Note. SW = strong–weak; WS = weak–strong; Fav. = favorable; Unfav. = unfavorable; phonotac. = phonotactics.

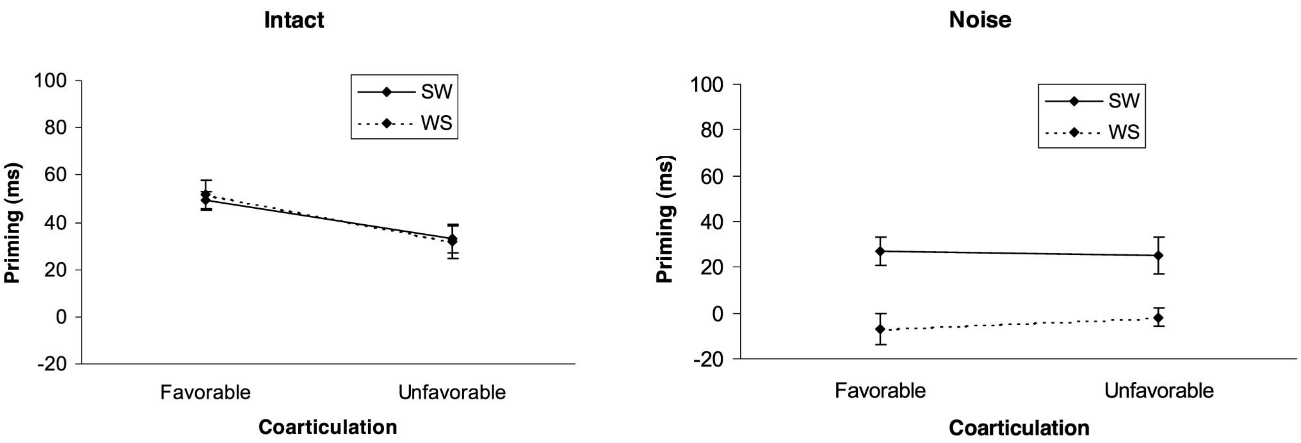


Figure 1. Experiment 1A: Priming effects (and standard-error bars) for SW and WS primes in acoustically favorable and unfavorable conditions. SW = strong–weak; WS = weak–strong.

acoustic–phonetic cues, regardless of the stress pattern of the prime. In noisy speech, priming was facilitated by stress, regardless of the acoustic–phonetic cues. This finding was confirmed in an analysis of variance by subjects factoring interpretive condition (intact vs. noise), acoustic–phonetic cues (favorable vs. unfavorable), and stress (SW vs. WS).[1] Of particular interest were an interaction between interpretive condition and acoustic–phonetic cues, $F(1, 59) = 10.01$, $p < .01$, $MSE = 532.82$, $\eta^2 = .14$, and an interaction between interpretive condition and stress, $F(1, 59) = 8.67$, $p < .01$, $MSE = 1{,}600.64$, $\eta^2 = .13$. In intact speech, favorable acoustic–phonetic cues caused more priming than unfavorable ones, $F(1, 29) = 15.28$, $p < .001$, $MSE = 1{,}157.17$, $\eta^2 = .34$, but there was neither a stress effect nor an interaction between acoustic–phonetic cues and stress, $F(1, 29) < 1$, in both cases. In contrast, in the noise condition, SW primes generated more priming than WS ones, $F(1, 30) = 12.55$, $p < .001$, $MSE = 604.52$, $\eta^2 = .29$, but there was neither an effect of acoustic–phonetic cues nor an interaction between acoustic–phonetic cues and stress, $F(1, 30) < 1$, in both cases.[2]

These results replicate those of Mattys (2004), showing that reliance on acoustic–phonetic cues and stress varies as a function of signal quality, with stress-based segmentation operating better in conditions of acoustic–phonetic uncertainty. The absence of an effect of acoustic–phonetic cues in noisy speech can almost certainly be attributed to noise-induced degradation of acoustic–phonetic information. The absence of a stress effect in clear speech is consistent with metrical prosody having a secondary role when alternative cues are available. The possibility that the stress effect in noise was solely the consequence of greater intelligibility for strong than for weak syllables was discounted in a subsequent control experiment that kept the need for segmentation to a minimum. In this control experiment, the SW and WS primes from Experiment 1A (in noise) were spliced out of their carrier utterances and presented in isolation.

Experiment 1B: Isolated Primes

Method

The SW and WS primes from the concatenated condition of Experiment 1A (in noise) were spliced out of their carrier sentences. The baseline utterance was shortened such that it now only consisted of the scrambled-speech disyllabic fragment. We also used a second type of baseline to assess how our scrambled-speech baseline compared with a more conventional baseline, namely, an unrelated prime (i.e., one of the other primes, e.g., /ˈmærə/, paired with the target *lamenting*). The unrelated primes were chosen to be phonologically and semantically distinct from the target with which they were paired. The stress pattern of the unrelated primes was counterbalanced across targets and participants. The three instances of a target (primed condition, scrambled baseline, and unrelated baseline) were presented in different blocks. The three conditions were randomly assigned to the three blocks, individually for each target. This assignment was counterbalanced across participants and targets. This allowed us to enter block as a dummy variable in the analyses to check whether the repetition

[1] For editorial concision, analyses by items, which were peer reviewed, are not reported in this article. Apart from a few cases of marginal discrepancy, these were all consistent with the analyses by subjects. Similarly, statistical analyses on the accuracy data for all experiments either mirrored latency differences or were nonsignificant. There were no significant instances of speed–accuracy trade-off. Both analyses by items and accuracy analyses can be obtained from Sven L. Mattys on request.

[2] Although the context fragments were chosen to be as phonotactically permissible as possible, their component syllables greatly varied in their probability of being found within English words and, in particular, at the end of English words. In an attempt to evaluate the effect that such variability might have had on the results, we calculated the frequency of each preprime syllable (i.e., the second syllable of the context). To be consistent with the rationale of the experiment, we limited the frequency count to the number of words containing the syllable in a word-final position (estimates from CELEX). Because we counterbalanced the stress pattern of the context for each prime, half the critical syllables were weak and half were strong (e.g., /və/ and /ˈkeɪ/; see Appendix A). Syllable frequency was 148.00 ($SD = 278.03$) for the weak syllables of the SW contexts and 213.00 ($SD = 671.41$) for the strong syllables of the WS contexts, $t(19) = -0.55$, $p = .59$. We then reran the analysis of variance (by items), with syllable frequency as a covariate. None of the main patterns of results was significantly modulated by the introduction of the covariate. Thus, the frequency of the syllable preceding the onset of the prime did not affect the trade-off between coarticulatory and stress cues in clear and noisy speech.

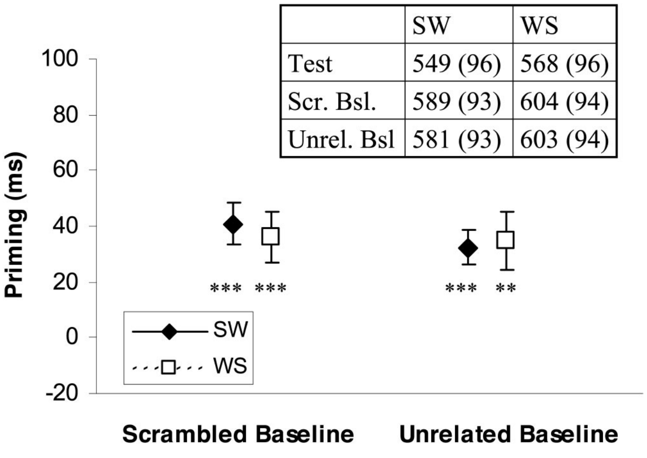
feature introduced any unwanted bias. The rest of the design and procedure were the same as in Experiment 1A ($N = 30$ participants).

Results and Discussion

As can be seen in Figure 2, all conditions produced significant priming. An analysis of variance with stress (SW vs. WS), baseline type (scrambled vs. unrelated), and repetition (Block 1, 2, 3) showed no effect of stress, $F(1, 29) < 1$; baseline type, $F(1, 29) = 1.59$, $p = .22$; or repetition, $F(2, 58) = 1.97$, $p = .15$; or interaction among any of the variables (all *p*s $> .10$). Planned comparisons confirmed that there was no stress effect in either the scrambled-baseline condition, $F(1, 29) < 1$, or the unrelated-baseline condition, $F(1, 29) = 1.16$, $p = .29$. Three conclusions can be drawn from these results. First, the fact that SW and WS disyllables were equally efficient primes when excised from continuous speech suggests that the stress effect in Experiment 1A cannot be reduced to a simple intelligibility difference, whereby strong syllables initiate lexical access solely because they are the only intelligible portions of the signal. The contrasted results between Experiment 1A and this control experiment reinforce the conclusion that reliance on strong syllables to access lexical information is contingent on the segmentation need imposed by the longer utterances. Second, the absence of a difference between the two types of baseline confirms that our scrambled baseline, while allowing greater control over stimulus comparisons and less repetition of priming material, did not over- or underestimate priming magnitude compared with an unrelated baseline. Third, the nonsignificant impact of the repetition factor suggests that the results were relatively stable across repetitions rather than emerging as their consequence.

Experiment 2: Stress Versus Phonotactics

The next question is whether stress shows similar weighting relative to other cues at the segmental level, for example, phonotactic probabilities. Research has shown that low-probability diphones (phonotactic troughs) are generally treated as possible word boundaries (e.g., Luce & Large, 2001; McQueen, 1998; Vitevitch & Luce, 1999). Although the effectiveness of phonotactic constraints for speech segmentation is thought to be computationally lower than that of stress (Norris et al., 1997), this claim has not been supported by behavioral evidence. In fact, McQueen (1998) found that listeners tended to rely more strongly on phonotactics than on metrical prosody when both cues were present. The next experiment is similar in design and procedure to Experiment 1A. As before, SW or WS primes were embedded at the end of nonsense utterances, but here we made the onset of the prime phonotactically favorable or unfavorable for segmentation by manipulating the probability of the diphone straddling the context–prime boundary.



	SW	WS
Test	549 (96)	568 (96)
Scr. Bsl.	589 (93)	604 (94)
Unrel. Bsl	581 (93)	603 (94)

Figure 2. Priming effects and standard-error bars for the SW and WS primes of Experiment 1B (noise condition) presented in isolation, as a function of the type of baseline. The inset table displays the lexical-decision latencies (and percentage correct) for the main conditions. Significance levels of priming effects (i.e., difference between baseline utterances and test utterances) are indicated for each condition: ** $p < .005$, *** $p < .001$ (*t* tests by subjects, $df = 29$). SW = strong–weak; WS = weak–strong; Scr. = scrambled; Bsl. = baseline; Unrel. = unrelated.

Method

Participants and materials. Sixty-nine participants were randomly assigned to the intact ($n = 34$) or noise ($n = 35$) condition. As the selection of the stimuli followed the same constraints as in Experiment 1A, only the critical information and differences are reported here. Thirty trisyllabic words were chosen, of which 15 had initial primary stress and 15 had medial primary stress. Their first two syllables were used as SW primes (e.g., /kʌtə/, from *customer*) and WS primes (e.g., kə'θi/, from *cathedral*). The average cohort size at the offset of the primes was 1.10 ($SD = 1.22$) for the SW primes and 0.60 ($SD = 0.82$) for the WS primes, $t(14) = 0.94$, $p = .36$. The average number of words containing the strong syllable in any position was 27.00 ($SD = 29.25$) for the SW primes and 23.00 ($SD = 27.52$) for the WS primes, $t(19) = 0.47$, $p = .65$. Average neighborhood density was 0.60 ($SD = 0.99$) for the SW primes and 0.30 ($SD = 0.46$) for the WS primes, $t(19) = 1.16$, $p = .26$. Primes were embedded at the end of nonsense utterances, for example, /gɑstem'kʌstə/. The within-word frequency of the diphone straddling the boundary between context and prime (e.g., /mk/ in /gɑstem'kʌstə/) was either low or high. Because low-frequency diphones are unlikely to be found inside words, they should be phonotactically favorable for segmentation compared with the high-frequency diphones (see Table 1 for an example and Appendix C). The diphones were as follows (low/high): mk/ŋk, mt/nt, ʒp/sp, mg/ŋg. Absolute CELEX occurrences averaged across the four sets were as follows (low/high): anywhere in a word (10/1,768), at syllable boundaries (10/884). Conditional probabilities, estimated as the probability of the second segment of the diphone occurring within a word given the first one, showed a similar contrast: .001/.261.

Baseline utterances were created as in Experiment 1A, the context of half being that of the favorable condition and the context of the other half that of the unfavorable condition. The filler utterances were similar to those in Experiment 1A, with no phonotactic constraints on phoneme sequences.

Design and procedure. These were the same as in Experiment 1A, except for the recording procedure (the speaker pronounced each utterance without interruption) and the total number of trials. Average intensity of the primes was 66 dB and 65 dB for the phonotactically favorable and unfavorable conditions, respectively, for the SW primes, and 67 dB and 66 dB, respectively, for the WS primes. There were 90 test utterances and 90 filler utterances, each presented with a target word and nonword, for a total of 360 trials.

Results and Discussion

Incorrect lexical-decision responses and responses beyond cutoff amounted to 10.5% of the test trials in the intact condition (6.7% incorrect) and 12.2% in the noise condition (7.8% incorrect). Figure 3 shows a similar overall pattern of results to Exper-

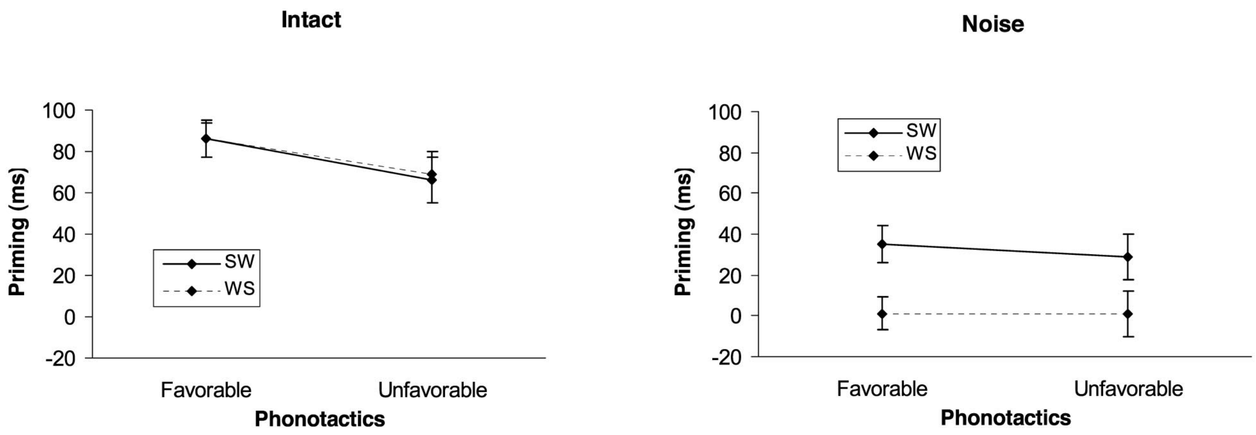


Figure 3. Experiment 2: Priming effects (and standard-error bars) for SW and WS primes in phonotactically favorable and unfavorable conditions (and standard-error bars). SW = strong–weak; WS = weak–strong.

iment 1A (see also Table 2): When the signal was intact, phonotactic constraints affected priming of both SW and WS primes, but stress, not phonotactics, promoted priming in the signal-degraded condition. An analysis of variance confirmed the interaction between interpretive condition and phonotactics, $F(1, 67) = 5.37$, $p < .05$, $MSE = 604.52$, $\eta^2 = .07$, and between interpretive condition and stress, $F(1, 67) = 11.06$, $p < .005$, $MSE = 1{,}610.42$, $\eta^2 = .14$. Analyses focusing on the intact condition showed a phonotactic effect, $F(1, 33) = 17.71$, $p < .001$, $MSE = 1{,}235.87$, $\eta^2 = .35$, but no stress effect nor interaction between phonotactics and stress, $F(1, 33) < 1$, in both cases. Similar analyses on the noise condition showed a stress effect, $F(1, 34) = 20.94$, $p < .001$, $MSE = 3{,}147.30$, $\eta^2 = .38$, but no phonotactic effect nor interaction, $F(1, 34) < 1$, in both cases.

Thus, in line with McQueen's (1998) results, we found a strong effect of phonotactic probabilities on the segmentation of words in clear speech. More critical, phonotactic regularities outweighed stress in phonetically clear listening conditions. That is, low-probability diphones tended to be interpreted as word boundaries, regardless of the stress level of the subsequent syllable. Likewise, high-probability diphones discouraged segmentation even if the subsequent syllable was strong. However, as in Experiment 1A, stress had a substantial impact when the signal was degraded.

Experiment 3: Stress Versus Lexicality

The rationale behind lexically driven segmentation is that the speech system favors segmentation solutions that are lexically plausible and disfavors those that are not. Segmentation strategies based on extracting familiar words from continuous speech have been shown to efficiently bootstrap lexical development (Brent, 1999; Dahan & Brent, 1999). Evidence from word spotting and priming experiments also suggests that lexical knowledge plays a significant role in online segmentation (e.g., Gow & Gordon, 1995; Norris et al., 1995, 1997). Even though the latter studies are more computationally specific about the mechanisms underlying lexically driven segmentation (e.g., by looking at how multiple alignment can solve the problem of lexical embeddedness and competition), their views and ours converge in describing segmentation as a product of lexical access.

Computational and behavioral evidence suggests that stress has a supporting function for lexical access, not a leading one (Norris et al., 1995; Vroomen & de Gelder, 1995). For instance, Norris et al. (1995) saw stress operating as a bias in the selection of already activated candidates, with stress-initial words being given an activation boost. However, no research has pitted lexical and metrical segmentation against each other. The following experiment fills this gap using the design of Experiments 1–2.

Method

Participants and materials. Sixty-eight participants were randomly assigned to the intact ($n = 34$) or noise ($n = 34$) condition. Stimulus selection was similar to that in Experiments 1–2. Only the main differences are reported here. The first two syllables of 30 trisyllabic words (15 initial stress, 15 medial stress) were used as SW and WS primes (e.g., /'vɜsə/, from *versatile*, and /vɪk'tɔ/, from *victorian*). The average cohort size at the offset of the primes was 1.60 ($SD = 3.07$) for the SW primes and 0.50 ($SD = 0.64$) for the WS primes, $t(14) = 1.64$, $p = .12$. The average number of words containing the strong syllable in any position was 25.00 ($SD = 17.14$) for the SW primes and 30.00 ($SD = 37.72$) for the WS primes, $t(19) = -1.00$, $p = .33$. Average neighborhood density was 0.80 ($SD = 1.28$) for the SW primes and 0.20 ($SD = 0.35$) for the WS primes, $t(19) = 1.65$, $p = .12$.

As illustrated in Table 1, each prime followed either a word (e.g., /ɪ‘nɔ məs'vɜsə/, *enormous versa*) or a nonword (e.g., /ə‘reɪməs'vɜsə/, *eraymous versa*). In contrast to the context fragments used in Experiments 1–2, the context words and nonwords in this experiment were three syllables long. This feature was meant to maximize the contrast between word and nonword contexts and minimize the chances of viable lexical parses other than those intended by the design. Word and nonword contexts were matched, pairwise, on stress pattern and average diphone frequency and shared the rime—or more—of their final syllable. Most words and nonwords reached their uniqueness or deviation point before their offset. Primes and contexts can be seen in Appendix D. Half the context stimuli had initial primary stress, and the other half had medial primary stress. Context stimuli were randomly paired with the primes, and pairings in which context words and primes were semantically associated were avoided. To minimize any unwanted context–prime pairing idiosyncrasies, we created two blocks of utterances, each with a different context–prime assignment. For each prime, the stress pattern of the context was counterbalanced across the two blocks. Participants were randomly assigned to

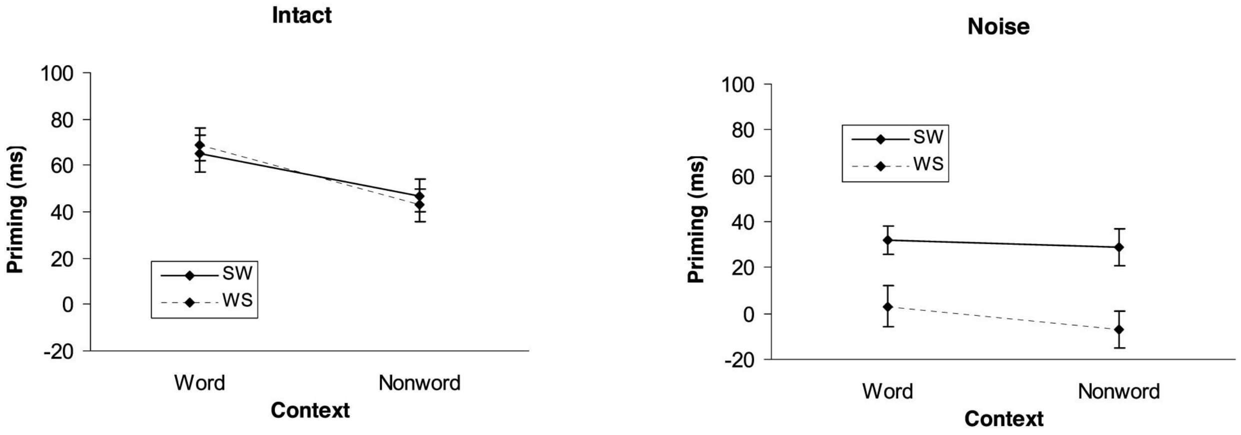


Figure 4. Experiment 3: Priming effects (and standard-error bars) for SW and WS primes in word and nonword contexts (and standard-error bars). SW = strong–weak; WS = weak–strong.

Block 1 or Block 2 (e.g., Block 1 participants heard /ɪ‘nɔ məs'vɜsə/, whereas Block 2 participants heard /’telɪskəp'vɜsə/).

Unlike the previous experiments, it was difficult to create a baseline utterance for each of the 30 stimulus sets because the preceding context differed segmentally within the sets. Therefore, we created and used a single baseline utterance for all the sets. We created the baseline utterance by digitally superimposing several test utterances. The resulting utterance sounded like babbling noise, the duration of which was roughly five syllables. The filler utterances were similar to those in Experiment 1A, with the targets overlapping various portions of the context of the utterance. Half the filler utterances started with a word and the other half with a nonword.

Design and procedure. These were similar to those of Experiment 2, with the reader, a female native speaker of southern British English, asked to pronounce each utterance without interruption. To minimize acoustic idiosyncrasies between the word and nonword context conditions, we used the same recording of the prime for both the word and the nonword conditions and as much as possible of the last syllable of the context. This common portion originated from the word-context recording and the nonword-context recording an equal number of times. It was then spliced to the initial part of the context. Average intensity of the SW and WS primes was 64 dB and 65 dB, respectively.

Results and Discussion

Incorrect lexical-decision responses and responses beyond cutoff amounted to 7.1% of the test trials in the intact condition (3.1% incorrect) and 8.5% in the noise condition (4.3% incorrect).

Figure 4 (see also Table 2) indicates that lexically driven segmentation prevailed in intact speech, whereas metrical prosody took over in noise. An analysis of variance revealed an interaction between listening condition and lexicality, $F(1, 66) = 5.74$, $p < .02$, $MSE = 3{,}184.28$, $\eta^2 = .08$, and between listening condition and stress, $F(1, 66) = 6.13$, $p < .02$, $MSE = 698.51$, $\eta^2 = .07$. Analyses focusing on the intact condition showed a lexical effect, $F(1, 33) = 20.79$, $p < .001$, $MSE = 775.68$, $\eta^2 = .39$, but no stress effect nor interaction between lexicality and stress, $F(1, 33) < 1$, in both cases. Similar analyses on the noise condition showed a stress effect, $F(1, 33) = 11.01$, $p < .005$, $MSE = 2{,}999.92$, $\eta^2 = .25$, but no lexical effect, $F(1, 33) = 2.05$, $p = .16$, nor Lexicality × Stress interaction, $F(1, 33) < 1$.[3] These results indicate that, consistent with a lexical approach to speech segmentation, the identification of known words in the input provides an efficient segmentation frame, independent of metrical cues. Metrical cues are confined to a modulatory role, contingent on signal quality.

Segmental Cues Versus Lexical Constraints

The first three experiments indicate that reliance on word stress is subsumed to segmental cues and lexical knowledge, with the term *segmental cues* used here to embrace both phonotactics and acoustic–phonetic variations in the realization of particular segments. A fuller account of the relation between segmentation strategies also requires a comparison between segmental cues and lexical knowledge. In Experiment 4, the absence of stress among the variables of interest enables us to use an alternative testing method, namely, word monitoring, in which participants monitor the presence or absence of a prespecified target word in a subsequent utterance. Although monitoring latencies are known to be affected by the lexical status of the preceding context and the transitional probabilities leading to the target (e.g., Foss & Blank, 1980), it is not known how these high-level sources of information fare against conflicting sublexical cues. Segmental cues are pitted

[3] A potential weakness in our analyses of Experiments 1–3 is that they involved repeated presentations of the target (baseline, favorable segmentation condition, unfavorable segmentation condition). Although intended to increase the power of the statistical tests, repetition could have caused some of the patterns of results to emerge because of the repetition feature rather than because of genuine segmentation strategies. Therefore, we reran the critical analyses of variance for Experiments 1–3, with order of presentation (first, second, third occurrence) as a control variable. Despite the very unequal cell sizes occasioned by this post hoc design, almost all the main results survived the repetition factor, and, more important, order of presentation did not interact with any of them. The stress effect in the noise condition of Experiment 3 only reached a significance value of .08, $F(1, 2) = 3.37$, $p = .08$, $MSE = 13{,}308.62$, $\eta^2 = .14$. However, it did not significantly interact with order of presentation, $F(2, 54) = 1.17$, $p = .32$, which indicates that the mild stress effect was probably due to the loss of power imposed by the post hoc design. Thus, overall, there was no indication that target repetition significantly contributed to the main patterns of results.

against lexical knowledge in Experiment 4 and against lexical–semantic information in Experiment 5.

Experiment 4: Segmental Cues Versus Lexicality

Method

Participants and materials. Fifty participants were randomly assigned to one of two conditions, intact ($n = 25$) and truncated ($n = 25$). Each of 28 monosyllabic target words (e.g., *male*) was embedded at the end of two tetrasyllabic utterances. In one utterance, the context was a trisyllabic word with no semantic link with the target (e.g., *calculus male*); in the other, it was a nonword (e.g., *baltuluf male*). Word and nonword contexts were matched, pairwise, on the nucleus of their third syllable and their average diphone probability, together with stress pattern, which varied between sets (see Table 3). In the word-context condition, the diphone straddling the end of the context and the beginning of the target was high frequency in English words (e.g., /sm/, in *calculus male*). In contrast, the diphone straddling the end of the nonword context and the target was low frequency (e.g., /fm/, in *baltuluf male*). In addition, the boundary between the context and the target was coarticulated in the word condition and decoarticulated in the nonword condition. The phonotactic and coarticulatory contrast between the word and nonword conditions was intended to put segmental cues and lexical ones in conflict. In the word condition, target word segmentation was lexically favored but segmentally disfavored. In the nonword condition, the pattern was the opposite. The test utterances can be seen in Appendix E. The straddling diphones were as follows (high/low): sm/fm, sn/fn, fl/ʒl, fr/mr, sl/ʒl, sm/ŋm, sf/zf. Occurrences in CELEX, averaged across all sets, were as follows (high/low): anywhere in a word (538/9), at syllable boundaries (208/9). Conditional probabilities, estimated as the probability of the second segment of the diphone occurring given the first one, were as follows: .049/.002.

Target-present fillers consisted of the same targets to detect in a different set of utterances. In some of these utterances, the first two syllables were a word or a nonword, the third syllable was the target, and the final syllable was a nonword (e.g., *plastic line min* or *placon line rin*, with *line* as the target). In others, the first syllable was a nonword, the second syllable was the target, and the last two syllables were a word or a nonword (e.g., *mell nose police* or *rell nose petring*, with *nose* as the target). Finally, there were two sets of target-absent trials. In one set, the trials mimicked the test trials in that the utterances started with a trisyllabic word or nonword and ended with a monosyllabic word. The targets were those of the test trials. None of the utterances' syllables corresponded to the targets. The other target-absent trials mimicked the structure of the filler utterances, with different words and nonwords and the same targets as before. In all, there were 56 test trials, 56 target-present filler trials, and 112 target-absent trials.

Table 3
Examples of Test Utterances in Experiments 4 and 5

Target: *male*	Intact	Truncated
Target: *male*		
Experiment 4: Segmental cues versus lexicality		
Word context	calculus male	culus male
Nonword context	baltuluf-male	tuluf-male
Target: *gap*		
Experiment 5: Segmental cues versus lexical semantics		
Congruent context	deepening gap	pening gap
Incongruent context	pseudonym-gap	donym-gap

Note. In Experiment 4, in the word condition, the onset of the target was coarticulated, and the straddling diphone was high frequency. In the nonword condition, the onset of the target was decoarticulated, and the straddling diphone was low frequency. Decoarticulation points are represented by dashes, and critical diphones are underlined. In Experiment 5, in the semantically congruent condition, the onset of the target was coarticulated, and the straddling diphone was high frequency. In the semantically incongruent condition, the onset of the target was decoarticulated, and the straddling diphone was low frequency.

Design and procedure. All the utterances were recorded by a male native speaker of southern British English (different from the speaker in Experiments 1–2). For the utterances involving a decoarticulation point—that is, those in the nonword-context condition—the speaker was instructed to pronounce the two fragments of each utterance one after the other (e.g., *baltuluf* and *male*), with a brief pause and mouth closure between them. The pause was then edited out and the two fragments concatenated following the procedure described in Experiment 1A. To prevent monitoring latencies from being affected by duration differences between the two versions of each target, we used a single rime in both context conditions. This common section of speech originated from the waveform of each condition an equal number of times. In many instances, however, the high degree of coarticulation between the onset and the rime of the target, together with differences in onset duration, meant that some of a target's onset was also common to the two conditions. As a consequence, the acoustic–phonetic contrast between conditions was not as marked as in Experiment 1A and was predominantly manifested prior to the onset of the target.

All utterances were either left intact or truncated. In the truncated condition, the first syllable of the utterances was excised with a speech editor (*calculus male* became *culus male*; *baltuluf male* became *tuluf male*). When one or both truncated contexts of a pair still constituted a word, the two contexts were truncated one segment further. Average diphone probability and deviation point of the truncated contexts were similar in the word and nonword conditions. The goal of the truncated condition was to neutralize the lexical distinction provided by the context while leaving the segmental distinction intact.

Participants were told that, on each trial, they would first see a word in the center of a computer monitor and then hear an utterance over the headphones. They were instructed to press a key, labeled *yes*, as soon as they heard the target word in the utterance or press another key, labeled *no*, if they did not hear the target in the utterance. Both speed and accuracy were emphasized. On each trial, a target word appeared in the center of a monitor for 1 s, followed immediately by the utterance. The participants had 3 s from the utterance offset to press a key. At the end of the 3-s response window or when the participant pushed a button, there was a 1-s pause before the next target.

Results and Discussion

Word-monitoring latencies were measured from the onset of the spoken target words. Incorrect responses to word targets and correct responses two standard deviations from the mean were discarded. In the intact condition, 12.3% of the test trials were discarded (8.4% incorrect), and 10.4% were discarded in the truncated condition (6.0% incorrect). Figure 5 indicates that, when present (intact condition), lexicality provided a stronger segmentation cue than segmental information. When lexicality was neutralized through truncation, the segmental cues took over. An analysis of variance factoring condition (intact vs. truncated) and cue (lexical vs. segmental) showed an interaction between the two factors, $F(1, 48) = 24.44$, $p < .001$, $MSE = 1{,}129.80$, $\eta^2 = .34$, in addition to somewhat faster latencies in the intact than in the truncated condition, $F(1, 48) = 3.03$, $p = .09$, $MSE = 13{,}653.60$,

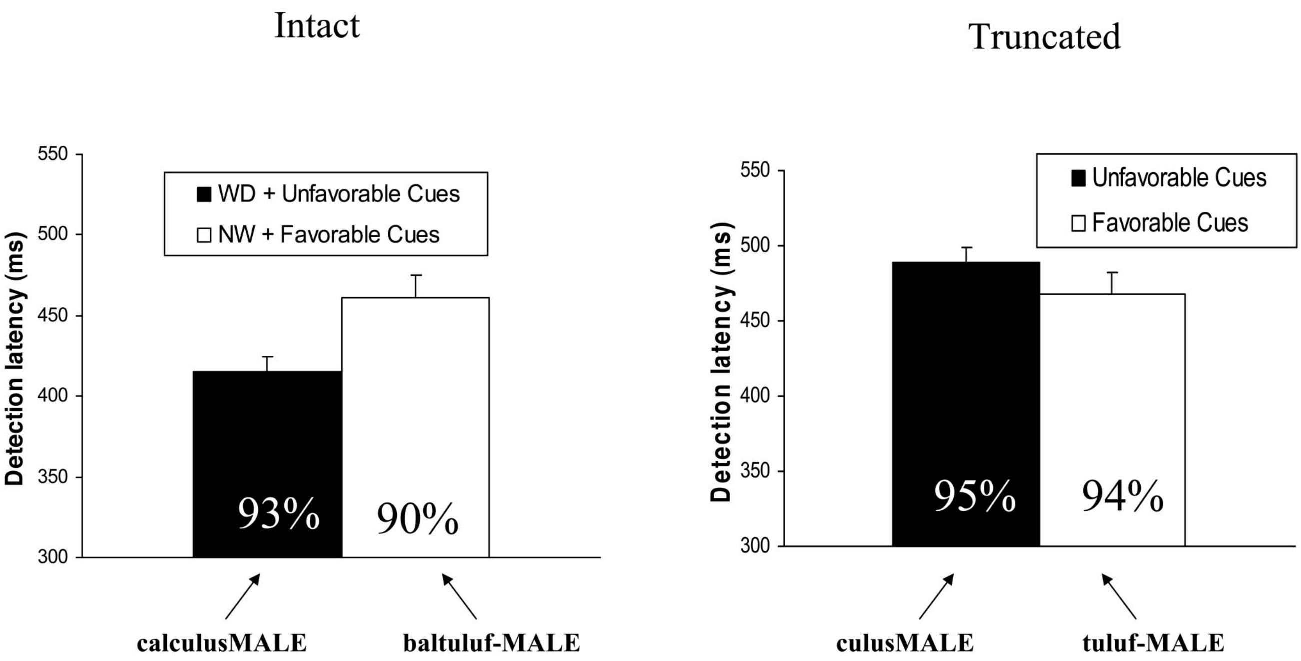


Figure 5. Experiment 4: Word-monitoring latency (and percentage correct). Segmental cues versus lexicality. Sublexically marked boundaries are represented by dashes (-).WD = word; NW = nonword.

$\eta^2 = .06$. An analysis of simple effects revealed a lexical advantage in the intact condition, $F(1, 24) = 16.32$, $p < .001$, $MSE = 1{,}619.85$, $\eta^2 = .40$, and a segmental advantage in the truncated condition, $F(1, 24) = 8.20$, $p < .01$, $MSE = 639.74$, $\eta^2 = .25$.

The results clearly demonstrate that lexical and segmental cues are not on the same footing during speech segmentation. Consistent with segmentation models that ascribe word recognition a central place in speech segmentation (e.g., Dahan & Brent, 1999; Norris et al., 1995, 1997), our finding suggests that a segmentation solution promoted by a lexically plausible parse is favored over one promoted by segmental cues, even when the latter conflict with lexical information (see also Gow & Gordon, 1995; Tabossi, Burani, & Scott, 1995).

Experiment 5: Segmental Cues Versus Lexical Semantics

This experiment tests the possibility that the predominance of lexically driven segmentation is promoted not only by lexical knowledge per se but also by lexical–semantic constraints, that is, the semantic relevance of a word in a given lexical context (e.g., Blank & Foss, 1978; Tyler & Wessels, 1983). In this experiment, lexical information provided by the context was supplemented by semantic information. Target word segmentation was either favored by semantic cues and disfavored by segmental cues or favored by segmental cues and disfavored by semantic cues.

Method

Participants and materials. Fifty participants were randomly assigned to the intact ($n = 25$) or truncated ($n = 25$) condition. Thirty monosyllabic target words were embedded at the end of two polysyllabic utterances each. In one utterance, the context was a word with a semantic link to the target (e.g., *deepening gap*); in the other, the context was a word with an incongruent semantic relation to the target (e.g., *pseudonym gap*). The context word was two or three syllables long, with varying stress patterns. As in Experiment 4, the two types of context were matched, pairwise, on number of syllables, stress pattern, nucleus of the final syllable, and average diphone probability as well as uniqueness point. In the congruent-context condition, the diphone straddling the end of the context and the beginning of the target was high frequency (e.g., /ŋg/, in *deepening gap*) and coarticulated, whereas, in the incongruent-context condition, the diphone was low frequency (e.g., /mg/, in *pseudonym gap*) and decoarticulated (see Table 3). The test utterances are listed in Appendix F. The diphones were as follows (high/low): ŋg/mg, sm/fm, sn/fn, fl/ʒl, sl/ʒl, sf/zf, ŋk/mk, nt/mt, sp/ʒp, sk/zk. Occurrences in CELEX, averaged across all sets, were as follows (high/low): anywhere in a word (1,202/6), at syllable boundaries (509/6). Conditional probabilities, estimated as the probability of the second segment of the diphone occurring given the first one, were as follows: .123/.002.

Target-present fillers and target-absent trials were of similar design to Experiment 4, except that the context portions of the utterances were always words. Their semantic relation to the targets varied from congruent to incongruent. In all, there were 60 test trials, 60 target-present filler trials, and 120 target-absent trials.

Design and procedure. These were the same as in Experiment 4. Average diphone probability and deviation point of the truncated contexts were similar in the congruent and incongruent conditions.

Results and Discussion

Word-monitoring latencies were measured from the onset of the spoken target words. Discarded data amounted to 8.5% of the test trials in the intact condition (4.9% incorrect) and 7.7% in the truncated condition (3.7% incorrect). In line with Experiments 3 and 4, the data pointed to greater reliance on high-order information (lexical semantics) than on segmental cues (see Figure 6). An analysis of variance showed an interaction between condition

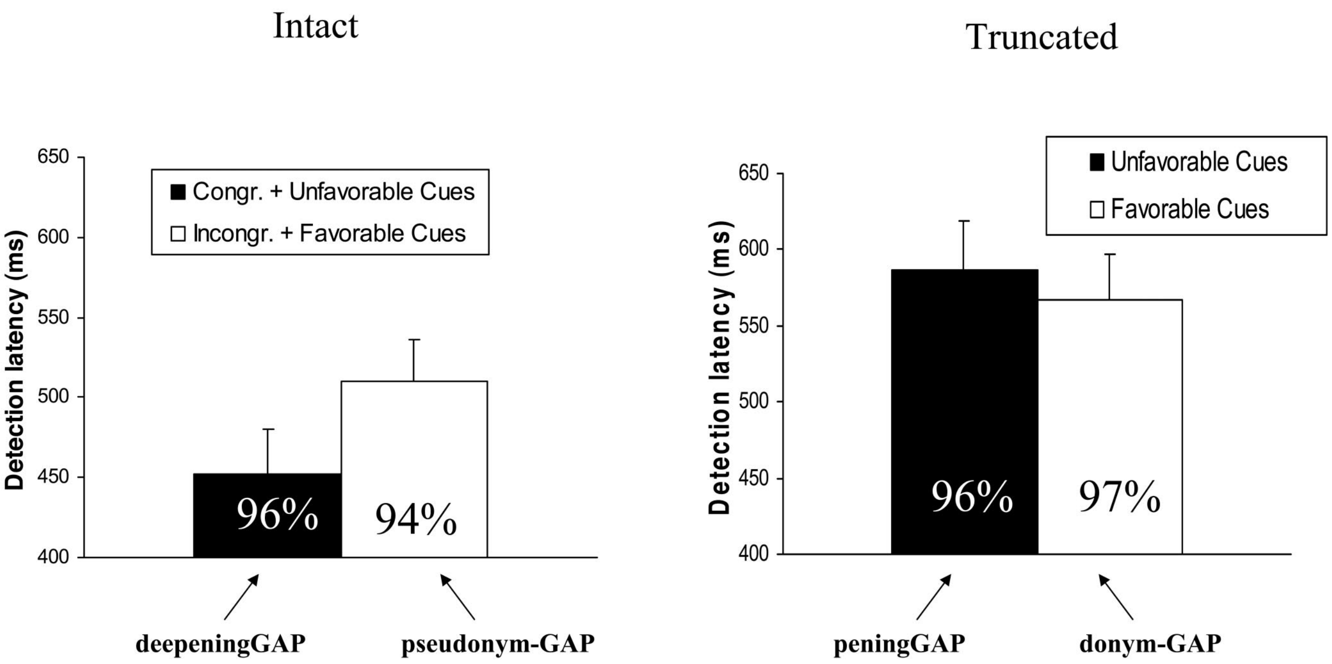


Figure 6. Experiment 5: Word-monitoring latency (and percentage correct). Segmental cues versus lexical semantics (congruent vs. incongruent). Sublexically marked boundaries are represented by dashes. Congr. = congruent; Incongr. = incongruent.

(intact vs. truncated) and cue (lexical–semantic vs. segmental), $F(1, 48) = 67.19$, $p < .001$, $MSE = 565.32$, $\eta^2 = .58$, in addition to a main effect of listening condition, $F(1, 48) = 5.52$, $p < .05$, $MSE = 42{,}022.39$, $\eta^2 = .10$. An analysis of simple effects revealed a lexical–semantic advantage in the intact condition, $F(1, 24) = 50.39$, $p < .001$, $MSE = 834.42$, $\eta^2 = .68$, and a segmental advantage in the truncated condition, $F(1, 24) = 16.81$, $p < .001$, $MSE = 296.23$, $\eta^2 = .41$. Thus, boundary cues in the signal were outweighed not only by lexical information (Experiment 4) but also by the semantic context that words provide. Greater reliance on semantic rather than acoustic cues is consistent not only with the ultimate purpose of the speech act (i.e., the conveyance of information) but also with behavioral and electrophysiological evidence showing that the detection of semantic incongruities in speech is more automatized (i.e., independent from attention) than that of phonological incongruencies (Perrin & García-Larrea, 2003; see also Radeau, Besson, Fonteneau, & Castro, 1998).

Hierarchically Organized Constraints

Taken together, the first five experiments suggest that listeners' segmentation strategies operate within a rank-ordered structure. In particular, Experiments 3–5 demonstrate that lexically–semantically mediated segmentation takes precedence over both segmental and metrical prosodic cues. With lexical–semantic information removed or neutralized, segmentation falls back on segmental information in preference to stress (Experiments 1–2). Metrical prosody acts as a last resort segmentation heuristic when alternative cues are obscured by noise. These contingencies are illustrated in Figure 7. Although not an exhaustive description of the variables and processes involved in speech segmentation, it is the first empirically supported account of how several previously documented cues interact and trade off as a function of information level and signal quality. The core claim is that there is an underlying hierarchy of weights, whereby reliance on some cues is intrinsically greater than reliance on others. Furthermore, although the weights of the segmentation cues are fixed, cues at lower levels of the hierarchy tend to manifest themselves when the interpretive conditions make those at higher levels unavailable or inefficient.

On the basis of our results, we have grouped segmentation cues in three tiers of importance. When all cues are optimally available, speech segmentation is lexically driven (Tier I), even in the presence of discrepant sublexical cues. Although we do not directly test the concept in the present experiments, we expect that, in many instances, the semantic and syntactic content of an utterance contributes to lexically driven segmentation by favoring those words most likely given a particular context. Sublexical cues are called on when lexical information is unavailable, impoverished, or ambiguous. A further distinction is made between segmental information (Tier II) and metrical prosody (Tier III), with the former outweighing the latter. Metrical prosody, or word stress, appears to best induce segmentation when word boundaries cannot be inferred using segmental cues. The subordinate status of stress is consistent with its partial reliability as a segmentation heuristic. An indiscriminate use of stress-based segmentation not only misses the onset of all noninitial-stress words but also missegments a majority of polysyllabic words (i.e., those containing more than one strong syllable). By contrast, segmental regularities offer a higher degree of reliability, because acoustic–phonetic cues are

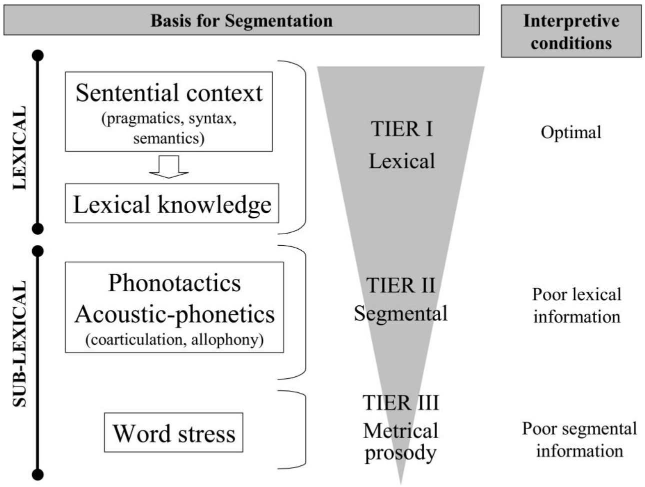


Figure 7. Sketch of the hierarchical approach to speech segmentation. The relative weights of the segmentation cues are illustrated by the width of the gray triangle.

conditioned by word boundaries and phonotactic patterns are influenced by lexical constraints on phoneme succession.

Multiple Cues: Stress, Sentential Context, and Coarticulation in Naturalistic Sentences

The previous experiments might offer a somewhat simplified account of speech segmentation. First, everyday utterances tend to include redundant cues rather than conflicting ones and are, in general, produced as meaningful sentences. Moreover, the experiments do not reveal—in some cases, were not designed to reveal—additive effects between cues. For instance, none of the experiments investigating metrical prosody (Experiments 1–3) found that stress reinforced segmentation on the basis of phonotactic, coarticulatory, or lexical cues. This could indicate that segmentation cues are used in an all-or-nothing fashion, with lower order cues being disregarded if higher order cues are present. However, the controlled designs also could have masked the graded interaction between cues that would typically be observable in richer listening environments.

Experiment 6A

In this experiment, we attempt to further specify the hierarchical constraints on cue reliance by (a) considering segmentation within a sentential context and (b) using intermediate levels of signal degradation to investigate the possibility that cue modulation is graded rather than all or nothing. This experiment is based on a cross-modal identity priming paradigm. Participants heard spoken sentences containing a WS test word in their later half. The strong syllable (S) of the WS word was itself a word (e.g., *creMATE*). The sentential context was semantically consistent with the WS word but not with the S word (e.g., *An alternative to traditional burial is to creMATE the dead*). Thus, the segmentation of the WS word was favored by the lexical and semantic context but disfavored metrically. In contrast, the S word was disfavored by the context but favored metrically. Segmentation was estimated by the amount of priming on the WS or S words visually displayed at the end of the WS word.

The sentences were played in four levels of signal quality: intact, mild noise, moderate noise, and severe noise (the severe condition was equivalent to the noise condition in the earlier experiments). Under the assumption that sentential context is a stronger cue than stress in optimal listening conditions, the WS words should show more priming than the S words when speech is intact. However, under the assumption that stress is a robust segmentation cue in acoustically degraded conditions, the S words should be less affected by noise than the WS words, and, hence, segmentation of S words should be relatively easier in severe noise. The intermediate noise levels will provide an insight into the relative recruitment of contextual and metrical information in cases of interpretive uncertainty. On the one hand, a strict hierarchical approach predicts that listeners should hold on to lexical–contextual cues as long as the signal quality permits it and then switch to a metrically dominated strategy. On the other hand, intermediate degradation levels could be conducive to more graded contributions from both sources of information, with a gradual rather than a discrete transition from contextual dominance in intact speech to metrical dominance in severely degraded speech.

Method

Participants and materials. Forty-eight participants were randomly assigned to the intact, mild noise, moderate noise, and severe noise conditions ($n = 12$ in each group). Forty test sentences were chosen, which provided a congruent semantic context for a late-occurring iambic (WS) word (e.g., *An alternative to traditional burial is to cremate the dead*, with the WS word underlined). The strong syllable of the WS word was itself a word (e.g., *mate*) but the WS and the S words were semantically unrelated. The WS and S words were matched on their frequency of occurrence. We held phonotactic cues constant by matching the probability of the diphone straddling the onset of the WS word with that of the diphone straddling the onset of the S word (e.g., . . . *to cremate* . . ., in which the two critical diphones are underlined). The test sentences can be seen in Appendix G. Another set of 40 sentences, matched pairwise with the 40 test sentences on their total duration, provided a neutral context for both the WS and the S words. These baseline sentences did not contain the WS or S words or any words phonologically overlapping with them. As in the previous cross-modal experiments (Experiments 1–3), the purpose of the baseline was to factor out any intrinsic lexical-decision differences between the WS and S words. For instance, the baseline sentence for the targets *cremate* and *mate* was *Waiting for the world leaders outside the conference were thousands* of protesters*, with the asterisk indicating roughly when *cremate* or *mate* was visually presented. Forty filler sentences contained other combinations of semantic or phonological overlap among SW, WS, and S spoken words and mono- and disyllabic visual target words. Finally, a separate set of 100 sentences were paired with mono- or disyllabic target nonwords whose phonological overlap with the preceding spoken word was full, partial, or nil. For practice, 16 original sentences were created, in which the breakdown of conditions was similar to that of the test set.

Design and procedure. The sentences, recorded by the speaker of Experiments 4 and 5, were either left acoustically intact or degraded by white noise. The SNR in the noise conditions, measured against the average intensity of the nonsilent portions of the utterances (~ 65 dB), was 5 dB (mild), 0 dB (moderate), and −5 dB (severe), with the noise level for the severe condition similar to that of Experiments 1–3. Participants were randomly assigned to one of the four noise conditions (intact, mild, moderate, and severe). To prevent any repetition of sentences for a given participant, we further arranged the participants in each noise condition into two subgroups. For one subgroup, half of the 40 test sentences were presented with the WS visual target and the other half with the S visual target. The other subgroup received the opposite assignment.

Participants were told that sentences would be played over the headphones and that a letter string would appear on a computer monitor in front of them at some point during playback. They were instructed to decide whether the letter string was a word, using a two-button response box. Both speed and accuracy were emphasized. To those participants assigned to the noise conditions, a warning was made about the relative unintelligibility of the sentences. In the test trials, the target appeared 100 ms after the boundary between the WS word and the next word of the sentence. The target remained on the screen until the participants gave their response or until 4 s had elapsed. There was then a 1-s pause before the onset of the next utterance.

Results

Lexical-decision latencies were measured from the onset of presentation of the visual target. Incorrect responses and latencies beyond the two standard deviation cut-off amounted to 11.4%, 10.4%, 7.8%, and 8.5% of the test trials in the intact, mild-noise, moderate-noise, and severe-noise conditions, respectively (7.5%, 6.1%, 4.2%, and 3.5%, respectively, were incorrect responses). Average lexical-decision latencies and accuracy are reported in Table 4. Priming effects, which are plotted in Figure 8, were calculated as the difference between the lexical-decision latencies in the baseline and test conditions.

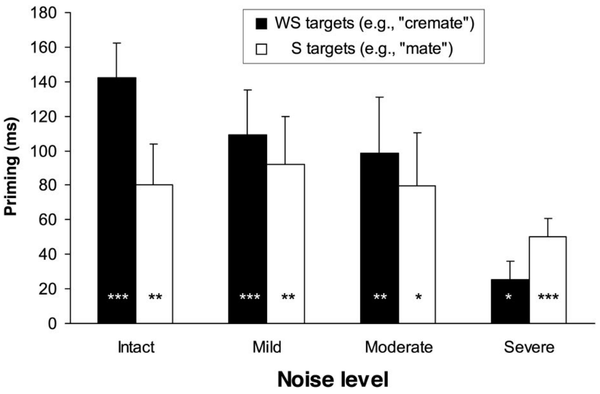


Figure 8. Experiment 6A: Priming effects (and standard-error bars) for WS and S targets as a function of degree of signal degradation. Significance levels of priming effects (i.e., difference between baseline utterances and test utterances) are indicated for each condition: * $p < .05$, ** $p < .01$, *** $p < .001$ (t tests by subjects, $df = 10$). WS = weak–strong; S = strong.

Overall, the results showed greater priming of WS than S words in intact speech, suggesting dominance of contextual information over stress, and greater priming of S than WS words in severe noise, suggesting the opposite pattern. In mild and moderate noise conditions, context-based priming was slightly greater than stress-based priming, but this effect was not statistically reliable. These results were confirmed in an analysis of variance factoring source of segmentation (context [WS] vs. stress [S]) and noise level (intact, mild, moderate, severe). Of primary interest was a significant interaction between source of segmentation and noise level, $F(3, 44) = 4.64$, $p < .01$, $MSE = 1{,}626.65$, $\eta^2 = .24$. In intact

Table 4
Lexical-Decision Latencies and Percentage Correct for Experiments 6A and 6B

	Intact		Mild		Moderate		Severe	
Noise level	Latency	% correct	Latency	% correct	Latency	% correct	Latency	% correct
				Experiment 6A: Stress in sentential context				
WS target								
Test sentence	472	96	604	98	561	99	542	98
Baseline sentence	615	93	713	95	659	96	568	97
S target								
Test sentence	529	92	615	94	571	95	537	96
Baseline sentence	609	89	707	89	651	93	588	95
				Experiment 6B: Stress in sentential context (plus decoarticulation)				
W#S decoarticulation								
WS target								
Test sentence	546	100	527	97	588	98	621	96
Baseline sentence	678	93	594	94	649	95	657	95
S target								
Test sentence	570	97	539	96	593	94	602	95
Baseline sentence	658	93	604	90	645	93	667	89
#WS decoarticulation								
WS target								
Test sentence	512	98	508	96	586	98	622	96
Baseline sentence	660	93	602	92	656	92	638	90
S target								
Test sentence	559	94	547	95	595	97	601	95
Baseline sentence	640	94	602	93	640	94	656	93

Note. WS = weak–strong; S = strong; W#S and #WS = weak–strong test words, with # indicating the decoarticulation point.

speech, context-based priming was greater than stress-based priming, $F(1, 11) = 19.68$, $p < .001$, $MSE = 2{,}368.90$, $\eta^2 = .64$. In mild and moderate noise conditions, the source of segmentation effect was not significant, $F(1, 11) < 1$, in both cases. In severe noise, however, stress-based priming was greater than context-based priming, $F(1, 11) = 4.93$, $p < .05$, $MSE = 1{,}473.17$, $\eta^2 = .31$. A separate set of simple-effect analyses revealed that noise level affected the contribution of contextual information (i.e., priming of the WS words), $F(3, 44) = 4.32$, $p < .01$, $MSE = 6{,}734.63$, $\eta^2 = .23$, but not that of stress (i.e., priming of the S words), $F(3, 44) < 1$.

Discussion

These results suggest two conclusions. First, the reversal of segmentation preference between the intact and severe noise conditions confirms the dominance of higher order information when the signal is acoustically clear and the reliance on stress in degraded listening conditions (−5 dB SNR). In this respect, the data are consistent with the trade-off between stress and lexical information found in Experiment 3. Note that the reversal in cue reliance in severe noise was mostly due to an attenuation of the effect of contextual cues relative to stress rather than to an increase of stress reliance per se. Thus, as mentioned earlier, the notion of stress as a last resort segmentation cue (cf. Liss et al., 1998, 2000; Mattys, 2004) should be seen not as a change in the absolute weight of stress in degraded listening conditions but rather as the result of the greater tolerance of stress to signal degradation. In addition, the data suggest a more flexible conceptualization of hierarchically organized cues than was apparent in Experiment 3. Indeed, in intact speech, although context-based segmentation elicited more priming than did stress-based segmentation, the latter elicited significant priming as well (see Figure 8 for significance levels). Comparable contributions were recorded in intermediate noise conditions. In severe noise, stress-based segmentation elicited greater priming, but context-based segmentation elicited significant, though reduced, priming. These results suggest that the hierarchy should provide for graded rather than all-or-none cue recruitment in natural speech environments.

The extent to which the priming effect found for the S words reflects their activation level during sentence processing is not entirely clear, however. Prior experiments using similar methodologies and stimuli have provided mixed results. On the one hand, a number of studies using *associative* cross-modal priming (i.e., featuring a semantic relation between prime and target) have indeed shown that late-embedded words (e.g., *bone* in *trombone*) are activated during playback of the carrier word (e.g., Luce & Cluff, 1998; Shillcock, 1990), at least when their onset is aligned with sublexical word-boundary cues such as stress (Vroomen & de Gelder, 1997), boundary-appropriate allophones (Gow & Gordon, 1995), or syllable boundaries (Isel & Bacri, 1999). On the other hand, the few studies that have used *identity* priming, as we did, have either failed to show activation of embedded words (Marslen-Wilson & Warren, 1994) or found inhibition (Norris, Cutler, McQueen, & Butterfield, 2005). The disparity between those findings and ours might originate from differences in experimental design, especially with regard to the construction of baseline sentences. Norris et al.'s (2005) test sentences were reused as baseline sentences for other targets across different participant groups. In our design, test and baseline sentences were distinct: Thus, participants heard each sentence only once but saw a given target twice, once in the baseline condition and once in the test condition. Target repetition was a compromise aimed to maximize the number of items per condition, given the already design-taxing noise-level variable (between subjects) and source-of-segmentation variable (latticed across subjects and items). Target repetition might have inflated priming estimates for both target types.

However, this design consideration does not alter the conclusion that signal degradation affects the relative contribution of contextual information and stress and that this interaction is better described as graded rather than all or nothing. It is somewhat surprising, however, that the advantage of contextual cues over stress dissipated with noise levels as low as 5 dB SNR. Given the documented facilitatory effects of sentential context on word recognition in moderate noise (e.g., Kalikow, Stevens, & Elliott, 1977), one might have expected a more persistent contribution of contextual information in the intermediate noise conditions. In Experiment 6B, we investigate whether the inclusion of an acoustic segmentation cue, such as decoarticulation, affects the pattern of results found in Experiment 6A.

Experiment 6B

This experiment is similar to Experiment 6A, except that the test sentences were recorded in two new versions. In one version (labeled W#S), the onset of the S word was decoarticulated from the preceding weak syllable (e.g., *cre#MATE*). In the other version (labeled #WS), the onset of the WS word was decoarticulated from the preceding material (e.g., *#creMATE*). Thus, the W#S condition provided acoustic–phonetic cues that favored segmentation of the S word, whereas the #WS condition provided acoustic–phonetic cues that favored segmentation of the WS word.

Method

Participants and materials. Ninety-six participants were randomly assigned to the intact, mild noise, moderate noise, and severe noise conditions ($n = 24$ in each group). The materials were those of Experiment 6A, except that each test sentence was rerecorded to include a decoarticulation point.

Design and procedure. The speaker of Experiment 6A was asked to read the test sentences of Experiment 6A with a break in speech either before the strong syllable of the WS test word (e.g., *An alternative to traditional burial is to <u>cre#mate</u> the dead*) or before the weak syllable (e.g., *An alternative to traditional burial is to <u>#cremate</u> the dead*). The decoarticulation procedure was the same as that in Experiments 1A, 4, and 5. Of these new recordings, only the WS word and the immediately surrounding syllables were kept. The fragment was then spliced into the original sentence (from Experiment 6A). The number of surrounding syllables kept around the WS word depended on the ease with which the fragment could be spliced into the host sentence. The maximum was three syllables before the WS word and one syllable after it. Thus, the new sentences were identical to the original ones, except for the portion containing the decoarticulation point (acoustic measurements are reported in Appendix H).

As in Experiment 6A, the sentences were presented in four noise levels: intact, 5 dB SNR (mild noise), 0 dB SNR (moderate noise), and −5 dB SNR (severe noise). Participants were randomly assigned to one of the four conditions. In each of these, participants were further arranged into four subgroups such that all of them were presented with both target conditions

(WS and S) and both decoarticulation conditions (W#S and #WS), but on different sets of sentences. The subgroup assignment was based on a breakdown of the 40 test sentences into four sets of 10. The four subgroups within a noise condition rotated through the four sentence sets according to a Latin square design factoring the two target conditions (WS and S) and the two decoarticulation conditions (W#S and #WS). Thus, compared with the participants in Experiment 6A, who received 20 sentences per condition, those in Experiment 6B received only 10. We tried to compensate for this difference by doubling the participant sample size. The rest of the procedure was the same as in Experiment 6A.

Results and Discussion

Lexical-decision latencies were measured from the onset of presentation of the visual target. Incorrect responses and latencies beyond the two standard deviation cut-off amounted to 9.3%, 11.0%, 9.4%, and 10.4% of the test trials in the intact, mild-noise, moderate-noise, and severe-noise conditions, respectively (4.7%, 5.9%, 4.9%, and 6.6%, respectively, were incorrect responses). Average lexical-decision latencies and accuracy are shown in Table 4, and priming effects can be seen in Figure 9. Overall, the priming results replicate those of Experiment 6A, showing a reversal of cue reliance as a function of noise level: Contextual information outweighed stress in intact speech, whereas stress outweighed contextual information in severe noise. Whether the onset of the WS or S word was decoarticulated only had an effect in the mild-noise condition. In this condition, decoarticulating the onset of the WS words caused greater priming of the WS than of the S words. Decoarticulating the onset of the S word caused this effect to disappear. As in Experiment 6 A, the moderate-noise condition did not significantly favor one segmentation cue over the other.

An analysis of variance, with source of segmentation (context [WS] vs. stress [S]), noise level (intact, mild, moderate, severe), and decoarticulation (W#S vs. #WS) as independent variables, showed a main effect of noise level, $F(3, 92) = 5.58$, $p < .01$, $MSE = 15{,}190.87$, $\eta^2 = .15$, with priming decreasing as the noise level increased. A source-of-segmentation effect, $F(1, 92) = 5.88$, $p < .05$, $MSE = 3{,}559.49$, $\eta^2 = .06$, suggested that, overall, priming was slightly greater for WS than for S words. However, this effect interacted with noise level, $F(3, 92) = 9.08$, $p < .001$, $MSE = 3{,}559.49$, $\eta^2 = .23$, confirming the compensatory relation between contextual information and stress. There was no main effect of decoarticulation, $F(1, 92) < 1$.

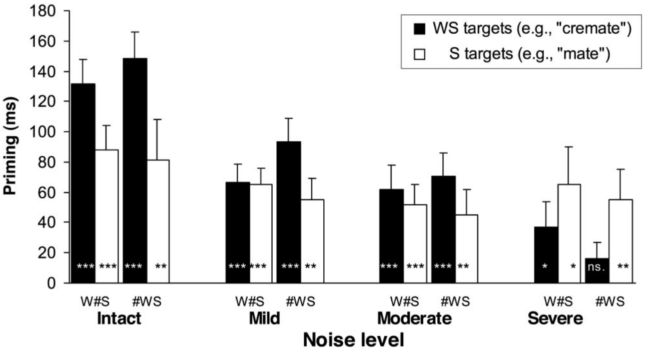


Figure 9. Experiment 6B: Priming effects (and standard-error bars) for WS and S targets as a function of degree of signal degradation and decoarticulation point (W#S [e.g., cre#MATE] vs. #WS [e.g., #cremate], where # indicates a decoarticulation point). Significance levels of priming effects (i.e., difference between baseline utterances and test utterances) are indicated for each condition: * $p < .05$, ** $p < .01$, *** $p < .001$ (t tests by subjects, $df = 22$). # = decoarticulation point.

To further specify the effect of noise and decoarticulation on the use of contextual versus stress cues, we ran a separate analysis of variance for each noise condition. In intact speech, significantly greater priming was found for WS than for S words, $F(1, 23) = 23.63$, $p < .001$, $MSE = 3{,}109.32$, $\eta^2 = .51$. The interaction between source of segmentation and decoarticulation was not significant, $F(1, 23) < 1$. In the mild-noise condition, however, this interaction was significant, $F(1, 23) = 4.61$, $p < .05$, $MSE = 1{,}833.41$, $\eta^2 = .17$: #WS decoarticulation enhanced priming for WS over S words, $F(1, 23) = 5.80$, $p < .05$, $MSE = 3{,}112.04$, $\eta^2 = .20$, whereas W#S decoarticulation suppressed this effect, $F(1, 23) < 1$. In moderate noise, no significant effects or interaction were found. In severe noise, priming was greater for S than for WS words, $F(1, 23) = 7.61$, $p = .01$, $MSE = 3{,}606.78$, $\eta^2 = .25$.

These results reinforce those of Experiment 6A in showing that listeners trade off context-based for stress-based segmentation when the acoustic signal is too degraded for the former to be sufficiently reliable. In addition, they reveal that acoustic–phonetic cues facilitate segmentation when the interpretive conditions (mild noise) provide neither clear contextual aid nor such strong signal degradation that acoustic–phonetic details become inaudible. This finding fits in well with the intermediate tier status of segmental cues outlined in Figure 7. Thus, taken together, Experiments 1A, 2, 4–5, and 6B suggest that segmental cues best contribute to speech segmentation when lexically driven segmentation is weakened by either (a) paucity of lexical and contextual information or (b) reduced availability of such information due to mild signal degradation.

General Discussion

This study shows that a full understanding of how listeners segment the speech stream must go beyond a description of the effects of individual cues. We found that the contribution of each cue or strategy was dependent on its position in the hierarchy of weights illustrated in Figure 7. In particular, stress proved to be a strong cue when word boundaries could not easily be derived from segmental-acoustic or lexical information. Segmental-acoustic cues were, themselves, secondary to lexical and higher level knowledge. The dominance of knowledge-based (Tier I) over sublexical cues (Tiers II–III) undoubtedly relates to the communicative, meaning-oriented nature of speech, as relying on lexical and higher order knowledge to carve meaningful chunks out of the input affords a far greater chance of communicative success than relying on sublexical cues. How the two types of information interact has implications for the following issues.

Hierarchical Constraints and Spoken-Word Recognition

Models of spoken-word recognition that describe the mapping between connected speech and word forms have often placed an emphasis on multiple alignment between signal and word entries and on lexical competition (e.g., Gow & Gordon, 1995; Norris, 1994). The extent to which sublexical boundary cues have an effect on activation and competition has been debated as well. Norris (1994) and Gow and Gordon (1995) ascribed a constraining rather than determining function to sublexical cues in lexical activation. Our results, too, indicate that sublexical cues are subsumed under lexically driven segmentation. We do not rule out the possibility that consistency between lexical and sublexical information can improve segmentation (via some activation-boosting mechanism or alignment benefit). However, in light of our data, we question whether inconsistency could ever result in sublexical dominance, as this would critically undermine the recognition of words containing low-frequency phoneme sequences (e.g., *Baghdad*), unconventionally pronounced words (e.g., unaspirated word-initial /t/ in southern British English *prints talk*), and noninitially stressed words.

Similarly, although our results provide little evidence that sublexical cues actively contribute to segmentation in optimal interpretive conditions, we recognize that those candidates that align with sublexical cues might still receive higher activation than those that do not, as reported by Gow and Gordon (1995). Our results suggest, however, that, relative to the evidence provided by lexical and contextual information, such activity differences would be too small to have a significant impact on segmentation in rich interpretive conditions. In ambiguous parsing conditions (caused by mild noise, e.g.), lexical and postlexical evidence could be sufficiently impoverished to make the activity difference between sublexically aligned and misaligned candidates decisive. Similarly, the hierarchical approach gives stress-initial candidates a measurable boost primarily when both lexical and segmental sources of information fail to provide an unequivocal segmentation solution—for example, ambiguous embeddedness, novel words, and severely impoverished input.

Implementing the three-tier structure into models of spoken-word recognition may involve assigning rank-ordered weights to lexical, segmental, and metrical sources of information. During online processing, lexical candidates compatible with portions of the incoming stream would compete against each other for a suitable match with the input, possibly following the type of sequential activation implemented in Shortlist (Norris, 1994). This lexically driven process would be constrained by sentential information, for example, via a decrease of activity threshold for words consistent with the semantic and syntactic content of the utterance (see, e.g., Gaskell & Marslen-Wilson, 2001). However, whether sentential information affects segmentation via restricted activation or subsequent lexical selection cannot be inferred from the present study. Distinguishing the two accounts requires a finer analysis of the time course of lexical activation around the critical word boundaries. Note that an online implementation of the hierarchy also predicts that the likelihood of engaging sublexical tiers decreases as the utterance unfolds, because of the growing constraints provided by contextual information on lexical activation and selection. Researchers could test this correlate by examining the patterns of cue dominance in an experiment similar to Experiment 6B, in which the location of the probe word in the sentence is manipulated (e.g., early, middle, late) and/or the cloze probability at probe onset systematically varied.

Cross-Linguistic Implications

Because Tier 1 information—lexicality and semantic and syntactic context—is of foremost importance in all languages, it is likely to show little cross-linguistic variation in its position in the hierarchy. In contrast, sublexical cues clearly vary between languages. Each language, for example, has its own inventory of allophonic variations occasioned by word juncture. Relative cue reliability also varies cross-linguistically: As discussed below, cues that are only statistical trends in English may be more reliable in other languages. Where a given sublexical cue is more consistently associated with word boundaries than are others in the same language, the hierarchy of weights may reflect this.

There are a number of potential sources of cross-linguistic difference in sublexical cue weighting. Word-initial stress is a matter of statistical predominance in English, but in some languages, such as Italian, stress placement is more variable (e.g., Vivalda, 1987); in others, stress placement is largely or, indeed, wholly reliable. For example, primary stress is generally word final, with some exceptions, in Turkish (Hallé & Vergnaud, 1987) and always word initial in Hungarian (Hayes, 1995). Thus, native Hungarian or Turkish speakers may have a higher weighting for stress in their segmentation hierarchy relative to other sublexical cues. Likewise, phonotactic patterns show wide cross-linguistic variation in the number and type of consonants allowable in word onsets and codas. Many languages allow fewer distinct patterns of syllable onsets and codas than does English: Finnish lacks complex onsets, and Italian lacks complex codas (e.g., Levelt & van de Vijver, 2004). Indeed, virtually all Italian words end in a vowel. This almost guarantees that any consonant in Italian speech is word initial or word medial, allowing word segmentation to operate almost exclusively following vowels. In such languages, with a much smaller inventory of possible word onsets and/or codas than English, phonotactic cues may be more reliable and thus have a higher weighting than acoustic–phonetic cues, such as allophony and strengthening/decoarticulation.

Various languages manifest lengthening and strengthening at the initial and final edges of words or higher level prosodic constituents (e.g., Keating, Cho, Fougeron, & Hsu, 2003), but the relative weighting of such effects as segmentation cues is likely to depend on the salience and reliability of other sublexical cues. For example, if the whole word-initial syllable is consistently lengthened by virtue of being stressed, greater duration of the onset consonant (the typical locus of word-initial lengthening) is likely to be less salient and thus less important as a cue.

Cross-linguistic differences in sublexical cues to word boundaries and their relative weightings have obvious implications for second-language learning: Do people attend to the cues that they learned in first language acquisition when listening to a second language? Sanders, Neville, and Woldorff (2002) found that the use of a metrical segmentation strategy in English by nonnative speakers was influenced by the stress characteristics of listeners' first language. Similarly, looking at phonotactics, A. Weber (2000) showed that segmentation by proficient second language speakers was influenced by the word onset inventories of both first and second languages (German and English, respectively), with first language phonotactics dominant in cases of conflict. More work is required to determine the degree to which segmentation strategies in the native language can be adapted to second language input, either simply by suppression of the first language strategy (see, e.g., Cutler, Mehler, Norris, & Segui, 1992) or by development of new and separate strategies for the second language.

Developmental Implications

An obvious difference between adult language users and young language learners is that the latter must face the continuous signal with virtually no lexical knowledge. Thus, lexically driven segmentation (Tier I) is unlikely to be a prominent feature of early segmentation. However, an increasingly large literature suggests that many crucial language-specific sublexical regularities (i.e., Tiers II and III) are picked up before 18 months of age. In particular, research shows that familiarity with suprasegmental patterns emerges prior to that with segmental patterns. For example, soon after birth, infants show greater responsiveness to native than to nonnative prosody (e.g., Nazzi, Bertoncini, & Mehler, 1998) and, by 5 months, to the dominant lexical stress pattern in the native language (C. Weber, Hahne, Friedrich, & Friederici, 2004). Evidence of stress-based segmentation in English learners has been noted as early as 7.5 months (Jusczyk, Houston, & Newsome, 1999), but phonotactic and allophonic segmentation has not been observed until 8.5 or 9 months (e.g., Friederici & Wessels, 1993). The lower weighted cues in adult speech segmentation (Tier III) thus seem the earliest and, hence, the most critical ones at the onset of language development. This was confirmed by Mattys, Jusczyk, Luce, and Morgan (1999), who found that, although American 9-month-olds were familiar with both the prosodic and the phonotactic regularities of the language, prosodic well formedness overrode phonotactic cues when these were put in conflict. A similar dominance of word stress was found when pitted against distributional regularities (Johnson & Jusczyk, 2001; but see Thiessen & Saffran, 2003, for finer analyses[4]).

The use of stress in early segmentation is shown to progressively give way to more subtle acoustic–phonetic cues—at least in English learners. Morgan and Saffran (1995) found that, whereas 6-month-olds provided with both stress and distributional information relied primarily on the former in perceptual-grouping tasks, 9-month-olds could integrate both sources of information. The benefits of a shift in cue dominance were clearly demonstrated by Jusczyk et al. (1999): 7.5-month-olds missegmented weak–strong–weak sequences (e.g., *gui 'tar is*) before the stressed syllable (e.g., responding to *taris*, not *guitar*). However, at 10.5 months, infants correctly picked out the WS word (e.g., *guitar*), probably drawing on cues such as phonotactics and coarticulation. Thus, English-learning infants initially use stress information (Tier III) as a coarse segmentation tool, then gradually modulate its use by integrating more subtle and collectively more accurate speech cues (Tier II) at the turn of the 2nd year (see also Vihman, Nakai, DePaolis, & Hallé, 2004).

Sublexical segmentation during early infancy facilitates the rapid development of a lexicon and higher order linguistic knowledge, such as syntax and semantics (i.e., Tier I). In terms of computation, word knowledge itself is a powerful segmentation tool (e.g., see Bortfeld, Morgan, Golinkoff, & Rathbun, 2005, for a striking demonstration with young infants), and it is easy to imagine lexically driven segmentation progressively superseding sublexical cues in the course of language development. For example, Brent (1999) found that a segmentation algorithm based on the extraction of newly discovered words achieved incrementally successful segmentation when applied to a standard child-directed speech corpus. However, although lexical growth promotes improved segmentation accuracy, some of its correlates—increased neighborhood density, embeddedness, homophony—are likely to make the task harder. Ambiguity that cannot be solved at the lexical level has to be handled by other information tiers. Thus, just as the various segmentation cues in the hierarchy are unlikely to be consulted in an all-or-nothing fashion by adults, the development of the hierarchy from bottom to top tiers is unlikely to be a strictly linear process in infancy. Rather, we hypothesize that higher level tiers build on lower tiers, contributing more powerful and heavily weighted segmentation strategies to the existing repertoire while turning initial strategies into optional ones.

[4] Thiessen and Saffran (2003) suggested that the sensitivity to statistical regularities may in fact precede stress-based segmentation. In their experiments, they noticed that 7.5-month-olds solved segmentation conflicts using statistical regularities rather than stress, whereas 9-month-olds did the opposite. As the authors noted, it is possible that sensitivity to statistically recurring patterns draws the infant's attention to the predominance of stress-initial words in the signal. Thus, sensitivity to statistical regularities, perhaps as a population- and domain-general tool (Fiser & Aslin, 2002; Hauser, Newport, & Aslin, 2001), could occupy a pivotal position in the acquisition of not only words but also word-boundary cues, such as stress, phonotactics, coarticulation, and allophony. Its exact role in the development of the hierarchy is currently the object of a great deal of research.

References

Acton, W. I. (1970). Speech intelligibility in a background of noise and noise-induced hearing loss. *Ergonomics, 13,* 546–554.

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995). *The CELEX lexical database* (Release 2) [CD-ROM]. Philadelphia: Linguistic Data Consortium, University of Pennsylvania.

Blank, M. A., & Foss, D. J. (1978). Semantic facilitation and lexical access during sentence processing. *Memory & Cognition, 6,* 644–652.

Bortfeld, H., Morgan, J., Golinkoff, R. M., & Rathbun, K. (2005). Mommy and me: Familiar names help launch babies into speech-stream segmentation. *Psychological Science, 16,* 298–304.

Bregman, A. S. (1999). *Auditory scene analysis: The perceptual organization of sound.* Cambridge, MA: MIT Press.

Brent, M. R. (1999). An efficient, probabilistically sound algorithm for segmentation and word discovery. *Machine Learning, 34,* 71–106.

Cairns, P., Shillcock, R., Chater, N., & Levy, J. P. (1997). Bootstrapping word boundaries: A bottom-up corpus-based approach to speech segmentation. *Cognitive Psychology, 33,* 111–153.

Chambers, K. E., Onishi, K. H., & Fisher, C. (2003). Infants learn phonotactic regularities from brief auditory experience. *Cognition, 87,* B69–B77.

Christiansen, M. H., Allen, J., & Seidenberg, M. S. (1998). Learning to segment speech using multiple cues: A connectionist model. *Language and Cognitive Processes, 13,* 221–268.

Cole, R. A., & Jakimik, J. (1980). Segmenting speech into words. *Journal of the Acoustical Society of America, 64,* 1323–1332.

Cutler, A., & Butterfield, S. (1992). Rhythmic cues to speech segmentation: Evidence from juncture misperception. *Journal of Memory and Language, 31,* 218–236.

Cutler, A., & Carter, D. M. (1987). The predominance of strong initial syllables in the English vocabulary. *Computer Speech and Language, 2,* 133–142.

Cutler, A., Dahan, D., & van Donselaar, W. A. (1997). Prosody in the comprehension of spoken language: A literature review. *Language and Speech, 40,* 141–202.

Cutler, A., Mehler, J., Norris, D., & Segui, J. (1992). The monolingual nature of speech segmentation by bilinguals. *Cognitive Psychology, 24,* 381–410.

Cutler, A., & Norris, D. G. (1988). The role of stressed syllables in segmentation for lexical access. *Journal of Experimental Psychology: Human Perception and Performance, 14,* 113–121.

Dahan, D., & Brent, M. R. (1999). On the discovery of novel wordlike units from utterances: An artificial-language study with implications for native-language acquisition. *Journal of Experimental Psychology: General, 128,* 165–185.

Davis, M. H., Marslen-Wilson, W. D., & Gaskell, M. G. (2002). Leading up the lexical garden-path: Segmentation and ambiguity in spoken-word recognition. *Journal of Experimental Psychology: Human Perception and Performance, 28,* 218–244.

Dilley, L., Shattuck-Hufnagel, S., & Ostendorf, M. (1996). Glottalization of vowel-initial syllables as a function of prosodic structure. *Journal of Phonetics, 24,* 423–444.

Erber, N. P. (1971). Auditory detection of spondaic words in wideband noise by adults with normal hearing and by children with profound hearing loss. *Journal of Speech and Hearing Research, 14,* 373–381.

Fiser, J., & Aslin, R. N. (2002). Statistical learning of new visual feature combinations by infants. *Proceedings of the National Academy of Sciences, USA, 99,* 15822–15826.

Foss, D. J., & Blank, M. A. (1980). Identifying the speech codes. *Cognitive Psychology, 22,* 609–632.

Fougeron, C., & Keating, P. A. (1997). Articulatory strengthening at edges of prosodic domains. *Journal of the Acoustical Society of America, 101,* 3728–3740.

Frauenfelder, U. H., & Peeters, G. (1990). On lexical segmentation in TRACE: An exercise in simulation. In G. T. M. Altmann (Ed.), *Cognitive models of speech processing: Psycholinguistic and computational perspectives* (pp. 50–86). Cambridge, MA: MIT Press.

Friederici, A., & Wessels, J. (1993). Phonotactic knowledge of word boundaries and its use in infant speech perception. *Perception & Psychophysics, 54,* 287–295.

Gaskell, M. G., & Marslen-Wilson, W. D. (2001). Lexical ambiguity and spoken word recognition: Bridging the gap. *Journal of Memory and Language, 44,* 325–349.

Gómez, R. L. (2002). Variability and detection of invariant structure. *Psychological Science, 13,* 431–436.

Gordon, M., & Ladefoged, P. (2001). Phonation types: A cross-linguistic overview. *Journal of Phonetics, 29,* 383–406.

Gow, D. W., & Gordon, P. C. (1995). Lexical and prelexical influences on word segmentation: Evidence from priming. *Journal of Experimental Psychology: Human Perception and Performance, 21,* 344–359.

Grosjean, F., & Gee, J. (1987). Prosodic structure and spoken word recognition. *Cognition, 25,* 135–155.

Grossberg, S., & Myers, C. W. (2000). The resonant dynamics of speech perception: Interword integration and duration-dependent backward effects. *Psychological Review, 4,* 735–767.

Hallé, M., & Vergnaud, J.-R. (1987). *An essay on stress.* Cambridge, MA: MIT Press.

Harrington, J., Watson, G., & Cooper, M. (1989). Word boundary detection in broad class and phoneme strings. *Computer Speech and Language, 3,* 367–382.

Hauser, M., Newport, E. L., & Aslin, R. N. (2001). Segmentation of the speech stream in a non-human primate: Statistical learning in cotton-top tamarins. *Cognition, 78,* B41–B52.

Hayes, B. (1995). *Metrical stress theory: Principles and case studies.* Chicago: University of Chicago Press.

Isel, F., & Bacri, N. (1999). Spoken-word recognition: The access to embedded words. *Brain and Language, 68,* 61–67.

Johnson, E. K., & Jusczyk, P. W. (2001). Word segmentation by 8-month-olds: When speech cues count more than statistics. *Journal of Memory and Language, 44,* 548–567.

Jusczyk, P. W., Houston, D. M., & Newsome, M. (1999). The beginnings of word segmentation in English-learning infants. *Cognitive Psychology, 39,* 159–207.

Kalikow, D. N., Stevens, K. N., & Elliott, L. L. (1977). Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability. *Journal of the Acoustical Society of America, 61,* 1337–1351.

Keating, P., Cho, T., Fougeron, C., & Hsu, C. (2003). Domain-initial articulatory strengthening in four languages. In J. Local, R. Ogden, & R. Temple (Eds.), *Papers in laboratory phonology: Vol. VI. Phonetic interpretation* (pp. 143–161). Cambridge, England: Cambridge University Press.

Klatt, D. H. (1980). Speech perception: A model of acoustic-phonetic analysis and lexical access. In R. A. Cole (Ed.), *Perception and production of fluent speech* (pp. 243–288). Hillsdale, NJ: Erlbaum.

Levelt, C., & van de Vijver, R. (2004). The acquisition of syllable types in cross-linguistic and developmental grammars. In R. Kager, J. Pater, & W. Zonneveld (Eds.), *Constraints in phonological acquisition* (pp. 204–218). Cambridge, England: Cambridge University Press.

Liss, J. M., Spitzer, S., Caviness, J. N., Adler, C., & Edwards, B. (1998). Syllabic strength and lexical boundary decisions in the perception of hypokinetic dysarthric speech. *Journal of the Acoustical Society of America, 104,* 2457–2466.

Liss, J. M., Spitzer, S., Caviness, J. N., Adler, C., & Edwards, B. (2000). Lexical boundary error analysis in hypokinetic and ataxic dysarthria. *Journal of the Acoustical Society of America, 107,* 3415–3424.

Luce, P. A., & Cluff, M. S. (1998). Delayed commitment in spoken word recognition: Evidence from cross-modal priming. *Perception & Psychophysics, 60,* 484–490.

Luce, P. A., & Large, N. (2001). Phonotactics, neighborhood density, and entropy in spoken word recognition. *Language and Cognitive Processes, 16,* 565–581.

Luce, P. A., & Lyons, E. A. (1999). Processing lexically embedded words. *Journal of Experimental Psychology: Human Perception and Performance, 1,* 174–183.

Luce, P. A., & Pisoni, D. B. (1998). Recognizing spoken words: The neighborhood activation model. *Ear & Hearing, 19,* 1–36.

Marslen-Wilson, W. D. (1984). Function and process in spoken word-recognition. In H. Bouma & D. G. Bouwhuis (Eds.), *Attention and performance: X. Control of language processes* (pp. 125–150). Hillsdale, NJ: Erlbaum.

Marslen-Wilson, W. D., & Warren, P. (1994). Levels of perceptual representation and process in lexical access: Words, phonemes, and features. *Psychological Review, 101,* 653–675.

Mattys, S. L. (2004). Stress versus coarticulation: Toward an integrated approach to explicit speech segmentation. *Journal of Experimental Psychology: Human Perception and Performance, 30,* 397–408.

Mattys, S. L., Jusczyk, P. W., Luce, P. A., & Morgan, J. L. (1999). Phonotactic and prosodic effects on word segmentation in infants. *Cognitive Psychology, 38,* 465–494.

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

McQueen, J. M. (1998). Segmentation of continuous speech using phonotactics. *Journal of Memory and Language, 39,* 21–46.

McQueen, J. M., Norris, D., & Cutler, A. (1994). Competition in spoken word recognition: Spotting words in other words. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 621–638.

Morgan, J. L., & Saffran, J. R. (1995). Emerging integration of sequential and suprasegmental information in preverbal speech segmentation. *Child Development, 66,* 911–936.

Nazzi, T., Bertoncini, J., & Mehler, J. (1998). Language discrimination by newborns: Toward an understanding of the role of rhythm. *Journal of Experimental Psychology: Human Perception and Performance, 24,* 756–766.

Norris, D. (1994). Shortlist: A connectionist model of continuous speech recognition. *Cognition, 52,* 189–234.

Norris, D., Cutler, A., McQueen, J. M., & Butterfield, S. (2005). *Phonological and conceptual activation in speech comprehension.* Manuscript submitted for publication.

Norris, D., McQueen, J. M., & Cutler, A. (1995). Competition and segmentation in spoken-word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21,* 1209–1228.

Norris, D., McQueen, J. M., Cutler, A., & Butterfield, S. (1997). The possible-word constraint in the segmentation of continuous speech. *Cognitive Psychology, 34,* 191–243.

Oller, D. K. (1973). The effect of position in utterance on speech segment duration in English. *Journal of the Acoustical Society of America, 54,* 1235–1247.

Palmer, S. E. (1999). *Vision science: Photons to phenomenology.* Cambridge, MA: MIT Press.

Perrin, F., & García-Larrea, L. (2003). Modulation of the N400 potential during auditory phonological/semantic interaction. *Cognitive Brain Research, 17,* 36–47.

Radeau, M., Besson, M., Fonteneau, E., & Castro, S. L. (1998). Semantic, repetition and rime priming between spoken words: Behavioral and electrophysiological evidence. *Biological Psychology, 48,* 183–204.

Sanders, L. D., & Neville, H. J. (2000). Lexical, syntactic, and stress-pattern cues for speech segmentation. *Journal of Speech, Language, and Hearing Research, 43,* 1301–1321.

Sanders, L. D., Neville, H. J., & Woldorff, M. G. (2002). Speech segmentation by native and non-native speakers: The use of lexical, syntactic, and stress-pattern cues. *Journal of Speech, Language, and Hearing Research, 45,* 519–530.

Seligman, M. E. P. (1970). On the generality of the laws of learning. *Psychological Review, 77,* 406–418.

Shillcock, R. (1990). Speech segmentation and the generation of lexical hypotheses. In G. T. M. Altmann (Ed.), *Cognitive models of speech processing: Psychological and computational perspectives* (pp. 24–49). Cambridge, MA: MIT Press.

Smith, M. R., Cutler, A., Butterfield, S., & Nimmo-Smith, I. (1989). The perception of rhythm and word boundaries in noise-masked speech. *Journal of Speech and Hearing Research, 32,* 912–920.

Suomi, K., McQueen, J. M., & Cutler, A. (1997). Vowel harmony and speech segmentation in Finnish. *Journal of Memory and Language, 36,* 422–444.

Swinney, D. A. (1981). Lexical processing during sentence comprehension: Effects of higher order constraints and implications for representation. In T. Meyers, J. Laver, & J. Anderson (Eds.), *The cognitive representation of speech* (pp. 201–209). Amsterdam: North-Holland.

Tabossi, P., Burani, C., & Scott, D. R. (1995). Word recognition in connected speech in Italian. *Journal of Memory and Language, 34,* 440–467.

Thiessen, E. D., & Saffran, J. R. (2003). When cues collide: Statistical and stress cues in infant word segmentation. *Developmental Psychology, 39,* 706–716.

Turk, A. E., & Shattuck-Hufnagel, S. (2000). Word-boundary-related duration patterns in English. *Journal of Phonetics, 28,* 397–440.

Tyler, L. K., & Wessels, J. (1983). Quantifying contextual contributions to word-recognition processes. *Perception & Psychophysics, 34,* 409–420.

van Donselaar, W., Koster, M., & Cutler, A. (2005). Exploring the role of lexical stress in recognition. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology, 58*(A), 251–273.

Vecera, S. P., Vogel, E. K., & Woodman, G. F. (2002). Lower region: A new cue for figure-ground assignment. *Journal of Experimental Psychology: General, 131,* 194–205.

Vihman, M. M., Nakai, S., DePaolis, R. A., & Hallé, P. (2004). The role of accentual pattern in early lexical representation. *Journal of Memory and Language, 50,* 336–353.

Vitevitch, M. S., & Luce, P. A. (1999). Probabilistic phonotactics and neighborhood activation in spoken word recognition. *Journal of Memory and Language, 40,* 374–408.

Vitevitch, M. S., Luce, P. A., Charles-Luce, J., & Kemmerer, D. (1997). Phonotactic and syllable stress: Implications for the processing of spoken nonsense words. *Language and Speech, 40,* 47–62.

Vivalda, E. (1987). Italian text-to-speech synthesis: The linguistic processor. *Olivetti Research & Technical Review, 4,* 47–60.

Vroomen, J., & de Gelder, B. (1995). Metrical segmentation and lexical inhibition in spoken word recognition. *Journal of Experimental Psychology: Human Perception and Performance, 21,* 98–108.

Vroomen, J., & de Gelder, B. (1997). Activation of embedded words in spoken word recognition. *Journal of Experimental Psychology: Human Perception and Performance, 23,* 710–720.

Vroomen, J., Tuomainen, J., & de Gelder, B. (1998). The roles of word stress and vowel harmony in speech segmentation. *Journal of Memory and Language, 38,* 133–149.

Weber, A. (2000, May). *The role of phonotactics in the segmentation of native and non-native continuous speech.* Paper presented at the Workshop on Spoken Word Access Processes, Max Plank Institute for Psycholinguistics, Nijmegen, the Netherlands.

Weber, C., Hahne, A., Friedrich, M., & Friederici, A. D. (2004). Discrimination of word stress in early infant perception: Electrophysiological evidence. *Cognitive Brain Research, 18,* 149–161.

White, L. S. (2002). *English speech timing: A domain and locus approach.* Unpublished doctoral dissertation, Edinburgh University, Edinburgh, Scotland.

(*Appendixes follow*)

Appendix A

Primes and Context Stimuli in Experiment 1

Word used as basis for primes		Context (either SW or WS)	
SW	WS	SW	WS
marathon	material	/'revə/	/vɑl'keɪ/
modular	melodic	/'nəʊtə/	/vɪk'tɔ/
laminate	lamenting	/'regjʊ/	/ɲjʊə'rəʊ/
utilize	utopia	/'mædʒɪ/	/rɪ'se/
vaccinate	victorian	/'nævɪ/	/rɪ'sɪ/
lavender	linguistic	/'mæksɪ/	/mə'məʊ/
longitude	logistic	/'rædɪ/	/ju'təʊ/
magical	malicious	/'vɜsə/	/nɑk'tɜ/
maximum	mechanic	/'lævɪn/	/rɪ'be/
regular	recipient	/'mɑdjʊ/	/lɪŋ'gwɪ/
messenger	medallion	/'revə/	/vɑl'keɪ/
motivate	memorial	/'ridʒə/	/rɪ'fre/
navigate	neurosis	/'mesən/	/mə'lɑ/
notable	nocturnal	/'retɪ/	/mə'tɪ/
radical	rebellion	/'væksɪ/	/mɪ'dæ/
regional	recession	/'məʊtɪ/	/mə'lɪ/
reticent	refreshing	/'junɪ/	/lə'dʒɪ/
revenue	religion	/'jutɪ/	/ju'næɪ/
universe	united	/'lɑŋgɪ/	/mɪ'kæ/
versatile	volcano	/'mærə/	/lə'men/

Note. SW = strong–weak; WS = weak–strong.

Appendix B

Table B1
Mean Durations (ms) of Onset Consonants in Prime-Initial Syllables in the Concatenation and Coarticulation Conditions for Experiment 1

	Concatenated (favorable)		Coarticulated (unfavorable)	
Coarticulation	ms	*n*	ms	*n*
Phone class of onset				
/l/	48	12	70	12
/m/	37	24	87	24
/n/	40	8	82	8
/v/	62	6	98	6
Mean duration (total occurrences)	43	50	83	50
SD	14.84		13.63	

The waveform and spectrogram of each utterance were examined to determine the acoustic characteristics of the juncture between context and prime created by the concatenation (decoarticulation) procedure. As shown in Table B1, the most noticeable effect was the relative shortening of the onset consonant following the decoarticulation point. Mean duration of the onset consonant of the prime-initial syllable was thus significantly greater in the coarticulated condition than in the concatenated (decoarticulated) condition, $t(98) = 14.00$, $p < .001$, $MSE = 202.69$, $\eta^2 = .66$. This durational difference did not extend to the nucleus of the prime-initial syllable, $t(94) = 1.09$, $p = .28$. Note that the analyses of onset and nucleus duration were carried out on slightly different subsets of the data, because the duration of the syllable-initial consonants could not be reliably measured from the waveform and spectrogram for all tokens. For this reason, all tokens in which the prime began with /ɹ/ ($n = 40$) or /j/ ($n = 16$), plus four tokens beginning with /v/, were excluded from statistical analysis. The analysis was therefore based on 63% of the total recorded tokens. However, even in those cases, the mean duration of onset plus nucleus could be analyzed. In the concatenated condition, onset plus nucleus had a mean duration of 114.00 ms ($SD = 29.47$), compared with a mean of 153.00 ms ($SD = 42.55$) in the coarticulated condition, $t(38) = 3.38$, $p < .005$, $MSE = 1{,}339.67$, $\eta^2 = .23$. It may be reasonably inferred that lengthening of the onset /ɹ/ underlies this effect, given that the coarticulation versus concatenation condition consistently affects the onset but not the nucleus for /l/, /m/, /n/, and /v/. Onset consonants in the initial syllables of utterance-medial words are generally lengthened relative to onset consonants in word-medial position (Oller, 1973; Turk & Shattuck-Hufnagel, 2000), and word-initial onsets may be lengthened further at higher-level prosodic boundaries (Fougeron & Keating, 1997). However, at the initial edge of an utterance, the pattern is reversed: Utterance-initial onset consonants are generally shorter than in word-initial position away from the utterance edge (Fougeron & Keating, 1997; White, 2002). In our stimuli, the relative shortening of the onset consonant following the decoarticulation point was the primary indicator of a preceding utterance boundary. Other acoustic–phonetic cues to an utterance boundary, such as the presence of murmured voice in prejuncture vowels (Gordon & Ladefoged, 2001), were also observed, which strongly suggests that decoarticulation was realized successfully in the concatenation condition.

Appendix C

Primes and Context Stimuli in Experiment 2

Word used as basis for primes		Context (either SW or WS)	
SW	WS	SW	WS
customer	cathedral	/'gɑstem-ŋ/	/trə'zem-ŋ/
compliment	canary	/'gɪspem-ŋ/	/pə'ræm-ŋ/
continent	composer	/'træpem-ŋ/	/fɪ'dem-ŋ/
criminal	conclusion	/'wæsɪm-ŋ/	/te'plɪm-ŋ/
terminal	tremendous	/'blɪʃəm-n/	/kə'rum-n/
tactical	torrential	/'plækəm-n/	/sə'pum-n/
tropical	terrific	/'mektɪm-n/	/ʃən'deɪm-n/
tyranny	tomato	/'dreskəm-n/	/mor'teɪm-n/
punitive	parental	/'tɑkɪʒ-s/	/rə'neɪʒ-s/
pyramid	perception	/'spærəʒ-s/	/kə'ruʒ-s/
passenger	potato	/'glirəʒ-s/	/fə'rɪʒ-s/
primary	perspective	/'snærɪʒ-s/	/pə'teʒ-s/
galaxy	galactic	/'tæfəm-ŋ/	/də'ʃɪm-ŋ/
gravity	guerrilla	/'grɪdəm-ŋ/	/fə'pæm-ŋ/
gratitude	gregarious	/'petrɪm-ŋ	/dɪ'trɪm-ŋ/

Note. For each context stimulus, a dash separates the segment contributing to the low-frequency diphone from that contributing to the high-frequency diphone. SW = strong–weak; WS = weak–strong.

Appendix D

Primes and Context Stimuli in Experiment 3

Word used as basis for primes		Context	
SW	WS	Words	Nonwords
compromise	courageous	criminal	lectinal
dictionary	devotion	captivate	duptilate
diplomat	detector	doctorate	nermorate
lavender	linguistic	telescope	plesticope
longitude	logistic	tropical	broginal
marathon	mechanic	coconut	patidut
messenger	medallion	labyrinth	ramasinth
modular	melodic	terrible	pamible
motivate	memorial	scavenger	skopinger
notable	nocturnal	faculty	drapulty
radical	rebellion	government	tebament
revenue	religion	dormitory	kerbitary
universe	united	deficit	thomenit
vaccinate	victorian	burglary	simblary
versatile	volcano	architect	fermilect
		cognition	umbation
		sarcastic	thimpastic
		fanatic	jemealic
		neurosis	rolasis
		gymnastic	fortestic
		recession	rasellion
		refreshing	besifying
		enormous	geremous
		casino	lafarno
		banana	pelina
		pragmatic	dustotic
		offensive	enentive
		departure	rinterdure
		abnormal	perdonal
		citation	rembation

Note. SW = strong–weak; WS = weak–strong.

(*Appendixes continue*)

Appendix E

Test Stimuli in Experiment 4

Context		
Word	Nonword	Target
calculus	baltuluf	male
consequence	pentiluf	nest
handkerchief	pendefeej	list
photograph	segopraj	land
paragraph	tapitram	rust
autograph	ipokram	role
ambulance	anterinj	list
acquaintance	edwilltanj	line
business	keernef	mark
oblivious	etliviuf	milk
hilarious	miraliuf	moon
circumstance	leerdemstang	mind
psychosis	liedoming	mate
ambiguous	elpituof	nine
crucifix	prulymif	nose
mysterious	nilteriouf	near
mischievous	willkenuz	face
courageous	torakuz	fact
masculine	walculiz	file
suspension	derelshiouz	farm
cinnamon	lennerez	fine
physician	teridez	field
discipline	midipliz	feel
forbidden	lorpiddez	fresh
gentleman	nentromuz	fund
discussion	gispushiuz	firm
telephone	meletoz	film
conclusion	benclueruz	force

Appendix F

Test Stimuli in Experiment 5

Context		
Congruent	Incongruent	Target
deepening	pseudonym	gap
dressing	mayhem	gown
joking	maxim	clown
purring	podium	cat
mooing	denim	cow
loving	victim	kiss
mountain	kingdom	top
garden	system	tool
afternoon	misinform	tea
construction	momentum	team
African	Birmingham	tribe
bonus	mirage	point
purchase	garage	price
citrus	species	fruit
chivalrous	legalize	knight
practice	bailiff	match
devious	dandruff	mind
campus	barrage	life
photograph	sabotage	lab
cactus	ménage	plant
noxious	series	fumes
lawyers	axes	fee
religious	neuroses	faith
admirals	organize	fleet
restless	mischief	night
mining	pilgrim	gold
trousers	montage	press
obvious	sometimes	fact
lawless	handcuff	mob
lettuce	massage	leaf

Appendix G

Test Sentences in Experiments 6A and 6B

An alternative to traditional burial is to creMATE the dead.
Judas was the disciple who would beTRAY Jesus Christ.
The politician fought a long camPAIGN on this mandate.
He worked hard for many companies to further his caREER in business.
Before deciding which of them to buy, he wanted to comPARE the prices.
In the immediate aftermath of the earthquake, the government has yet to conFIRM numbers of dead.
The two players left in the tournament will conTEST the final.
As an agnostic, he couldn't decide whether to beLIEVE in God.
Whenever Queen Victoria passed through villages, she liked to conVERSE with local people.
The 3-1 loss was the first time the team had suffered deFEAT that season.
The lawyer strongly rebutted the claims in deFENSE of his client.
She longed to be an actress so she could perFORM on the stage.
To solve the problem, a reward was offered to inventors who could create the deVICE needed.
The religious miners put their miraculous escape down to diVINE intervention.
The sailors said their goodbyes before going to emBARK on the voyage.
The storyteller could enCHANT little children.
Before a claim was lodged, insurers demanded an independent assessment of the exTENT of the damage.
The bass player in the band learned guiTAR when he was young.
After she cleared the bar so easily she wanted to inCREASE its height.
At the time of the explosion, people a mile away felt the imPACT of the blast.
In a medieval joust, knights were literally attempting to imPALE opponents with a lance.
Jimmy spent all his money on a flash car to imPRESS the girls.
Years of hard practice had helped him to imPROVE immeasurably.
We like watching the geese fly overhead as they miGRATE for the winter.
After such a terrible holiday I admitted it was a misTAKE to go there.
Safety measures were introduced to preVENT any further disasters.
When I buy something new, I keep the reCEIPT in case I need to change it.
He was angry and made a rash reMARK to the press.
The band released their second reCORD in 1978.
After it is collected in its raw form, it is necessary to reFINE it.
He is held in high reGARD by his enemies.
The best four runners were asked to compete in the reLAY team.
After the system crashed due to excessive demand, administrators decided to reSTRICT access.
Because they were outnumbered, the soldiers decided to reTREAT to base.
After a long time away, she wanted to reTURN home.
The new bra was designed to provide extra supPORT where needed.
After a spate of break-ins, most families made sure to seCURE their homes.
In the heat of the day, the pace of life is quite seDATE and laid back.
During the food riots, many protesters went to surROUND the presidential palace.
His incredible success was put down to his ability to tranSCEND his poor childhood.

(*Appendix continues*)

Appendix H

Acoustic Measurements for Experiment 6A–6B

We examined the waveform and spectrogram of each utterance from Experiments 6A and 6B to determine the acoustic–phonetic characteristics of the neutral (Experiment 6A), W#S (Experiment 6B) and #WS (Experiment 6B) junctures created by the decoarticulation procedure, where # indicates a decoarticulation point. As shown in Table H1, the primary difference between the two decoarticulation conditions was the substantially greater duration of the strong syllable onset in the #WS condition compared with the W#S condition. An analysis of variance showed a significant effect of onset class (fricative, nasal, etc.; see Table H1), $F(6, 93) = 18.39$, $p < .001$, $MSE = 1{,}466.97$, $\eta^2 = .54$, and a significant effect of decoarticulation (neutral, W#S, #WS), $F(2, 93) = 7.30$, $p = .001$, $MSE = 1{,}466.97$, $\eta^2 = .14$. The two variables did not significantly interact. Additional comparisons showed that the duration of the strong syllable onset was shorter in the W#S condition than in the #WS or neutral conditions ($p < .001$ in both cases) and that the durations in the latter two were not significantly different. This difference was comparable to that found in Experiment 1A: The decoarticulation condition appears to have been phonetically realized as an utterance boundary, with the shortening of the onset that is characteristic of such a boundary.

Evidence of shortening of the weak syllable onset in the decoarticulation condition arose mainly in qualitative terms, probably because onset duration was difficult to measure in weak syllables starting with an approximant (nine instances). Additionally, some weak syllables lacked an onset (eight instances). Thus, an analysis of variance showed a significant effect of onset class, $F(3, 54) = 47.50$, $p < .001$, $MSE = 543.11$, $\eta^2 = .71$, but no significant effect of decoarticulation and no significant interaction. However, weak syllables in the #WS condition were frequently associated with allophony, such as glottalization. In contrast, the boundaries in the coarticulated conditions tended to be manifested as an unbroken transition from vowel to vowel or from vowel to approximant. Glottalization is known to be a feature of vowel-initial words, particularly at major prosodic junctures, such as phonological phrase boundaries (Dilley, Shattuck-Hufnagel, & Ostendorf, 1996), which suggests that the speaker might have used different methods of decoarticulation before the weak syllable and before the strong syllable. Given that the weak syllables followed a word boundary in the carrier sentences, it might have been more natural to realize the juncture as a higher level prosodic boundary, at least in some cases. In contrast, strong syllables were word medial. Therefore, given the relative unnaturalness of a prosodic break at this point, it seems that the speaker rather opted for a complete restart of speech, with the phonetic characteristics of an utterance boundary. Despite the apparent differences in the phonetic realization of the decoarticulation in the W#S and #WS conditions, the results reported in Table H1 suggest that in both cases the juncture was salient and used in segmentation in at least some listening conditions.

Table H1
Mean Durations (ms) of Onset Consonants of the Strong (Word-Medial) and Weak (Word-Initial) Syllables in the W#S and #WS Conditions for Experiment 6B

Decoarticulation condition	W#S		#WS	
	ms	*n*	ms	*n*
Strong syllables				
Phone class of onset				
Affricate	203	1	183	1
Approximant	89	2	100	2
Fricative	105	11	150	11
Nasal	53	2	123	2
Voiceless stop	138	16	193	16
Voiced stop	48	4	110	4
Cluster	220	2	234	2
M (total occurrences)	118	38	165	38
SD	53.03		49.42	
Weak syllables				
Phone class of onset				
Fricative	147	4	131	4
Nasal	104	2	62	2
Voiceless stop	162	10	174	10
Voiced stop	89	7	92	7
M (total occurrences)	132	23	132	23
SD	38.89		48.87	

Note. The duration of the onset consonant could not be reliably measured from the waveform and spectrogram for all tokens, particularly when the onset was an approximant. In addition, the weak syllables that started with a vowel were excluded from statistical analyses (eight instances). Voiced stop duration was measured from the onset of closure to stop release. Voiceless stop duration includes closure duration and aspiration duration. W#S and #WS = weak–strong test words, with # indicating the decoarticulation point.

Received September 8, 2004
Revision received May 24, 2005
Accepted May 30, 2005 ■