Original Paper

Phonetica 2008;65:173–186
DOI: 10.1159/000144078

Received: September 5, 2006
Accepted: September 15, 2007

Effect of Speaking Rate on the Identification of Word Boundaries

Sandra Schwab[a] Joanne L. Miller[c] François Grosjean[a]
Michèle Mondini[b]

[a]Université de Neuchâtel, Neuchâtel, and [b]SVOX AG, Zürich, Switzerland;
[c]Northeastern University, Boston, Mass., USA

Abstract

Two experiments were conducted to determine whether listeners' ability to use allophonic variation to identify word boundaries is influenced by speaking rate. Listeners in both experiments were presented two-word sequences (such as *great eyes*) spoken by naturally fast and naturally slow talkers; in one experiment the sequences were presented in quiet and in the other they were presented in noise. The listeners' task was to identify the intended sequence from among four choices with alternative segmentations (e.g. *great eyes, gray ties, great ties, gray eyes*). In both experiments performance was worse for the sequences produced by the naturally fast talkers than for those produced by the naturally slow talkers. This finding suggests that the extent to which allophonic variation contributes to the identification of word boundaries may depend on the rate at which the speech was produced.

Introduction

In order for listeners to comprehend a spoken message, they must identify the individual lexical items of the language from the continuously varying stream of speech they hear. Speech is a complex spectral/temporal signal that does not contain clear and reliable acoustic markers to word boundaries, and the process by which listeners identify individual words from the speech stream is a topic of considerable interest. In this paper we examine how mapping the continuous speech signal onto discrete words, that is, how listeners identify the location of word boundaries, is influenced by the rate at which the speech is produced.

It is well established that speaking rate differs from one talker to another [e.g. Allen et al., 2003; Grosjean and Deschamps, 1975; Miller et al., 1984], and that even for a given talker rate may vary as a function of the nature of the verbal activities the talker is engaged in [e.g. Goldman-Eisler, 1968; Grosjean and Deschamps, 1972, 1973,

KARGER

Fax +41 61 306 12 34
E-Mail karger@karger.ch
www.karger.com

0031–8388/08/0653–0173
$24.50/0
Accessible online at:
www.karger.com/pho

Dr. Joanne L. Miller
Psychology Department, Northeastern University
360 Huntington Ave., Boston, MA 02115 (USA)
Tel. +1 617 373 3766, Fax +1 617 373 8714
E-Mail j.miller@neu.edu

1975; Lucci, 1983]. It is also widely known that variation in the rate at which speech is produced has perceptual consequences for the listener.

One consequence concerns the effect of rate on global comprehension. Most researchers studying this issue have varied speaking rate artificially using time-compressed speech. Across a variety of tasks, including answering questions [Fairbanks et al., 1957; Vaughan and Letowski, 1997], sentence identification and recall [Buttet et al., 1980; Wingfield et al., 1999], sentence repetition [Vaughan and Letowski, 1997], and intelligibility rating [Vaughan and Letowski, 1997], the findings have shown that a fast speaking rate can impair comprehension. Interestingly, it is also the case that listeners can at least partially accommodate for even high levels of compression after relevant experience with compressed speech [Dupoux and Green, 1997; Sebastián-Gallés et al., 2000]. Although studies using naturally produced variation in speaking rate are less numerous than those using time-compressed speech, it has been found that at least for certain tasks (e.g. answering questions involving inferences), naturally produced fast speech is more difficult to process than naturally produced slow speech [Cohen, 1979; Nicholas and Brookshire, 1986].

Another perceptual consequence of speaking rate concerns lexical processing. Although the evidence is more limited than that for global comprehension, there are findings (mostly using naturally produced speech) indicating that speaking rate can influence the processing of individual words. For example, Bradlow and Pisoni [1999] found that words produced at a fast speaking rate were more difficult to identify than words produced at a moderate or slow speaking rate. Moreover, even variation in speaking rate can affect performance: Sommers et al. [1994] found that word identification decreased when the words to be identified were produced with different speaking rates that varied from trial to trial, compared to when they were produced at a constant rate of speech.

Speaking rate has also been shown to influence perceptual processing at the level of individual phonetic segments. A change in speaking rate produces a complex, nonlinear expansion and compression of the speech signal, affecting many acoustic-phonetic properties that specify the identity of individual consonants and vowels, and numerous studies have shown that listeners are sensitive to such variation [for an early review, see Miller, 1981]. To take just one example, consider voice onset time (VOT), a major property distinguishing voiced and voiceless syllable-initial consonants in English: relatively short VOT values specify voiced consonants, whereas longer VOT values specify voiceless consonants. There is now substantial evidence that VOT varies systematically with speaking rate, such that VOT decreases as speaking rate becomes faster, especially for voiceless stop consonants [Kessinger and Blumstein, 1997; Miller et al., 1986; Volaitis and Miller, 1992]. Moreover, there is also evidence that listeners accommodate for such variation. As speaking rate becomes faster, both the voiced-voiceless boundary along a VOT series [Summerfield, 1981], and the location of the best voiceless exemplars along the series [Miller and Volaitis, 1989; Volaitis and Miller, 1992; Wayland et al., 1994] shift to shorter VOT values, in accord with the systematic changes that occur in VOT during production. Findings such as these, which have been reported for a variety of acoustic-phonetic properties, document the listener's sensitivity to speaking rate when processing the individual consonants and vowels of the language.

In the current paper, we focus our attention on the interface between lexical and phonetic processing, and examine the influence of speaking rate on word segmentation,

that is, on the identification of boundaries between words.[1] There has been considerable debate on the nature of the processes underlying word segmentation. A prominent view in the field is that word segmentation per se does not precede word recognition, but rather that word segmentation is itself a product of word recognition. That is to say, the boundaries between words are defined as words are recognized. Some early accounts within this general view proposed that word recognition – and hence word segmentation – occurs via a word by word recognition process, with words recognized one after the other [e.g. Cole and Jakimik, 1980; Marslen-Wilson and Welsh, 1978]. More recently, proponents of this general view have argued that word recognition does not proceed word by word, but rather is achieved through activation and competition among multiple lexical candidates that share potential word boundaries [e.g. McClelland and Elman, 1986; Norris, 1994; Norris et al., 1995]. Of particular relevance to the current paper, it is now well established that sublexical phonetically based information about lexical boundaries can influence word segmentation, and researchers find a role for phonetically based information in their accounts. For example, it has been proposed that such information could modulate the activation-competition process, so that lexical candidates consistent with the sublexical information are favored [Gow and Gordon, 1995; McQueen, 1998; Norris et al., 1997; Shatzman and McQueen, 2006].

One kind of phonetically based information that plays a role in word segmentation is allophonic variation. Although the speech signal does not contain pauses or other clear markers between each word, the signal does contain some acoustic-phonetic information that is correlated with word boundaries. More specifically, there is strong evidence that phonetic segments are articulated slightly differently according to their position within a word, and this leads to systematic allophonic differences in the speech signal. For example, vowels at the beginning of words typically show distinctive laryngealization and/or glottalization, and voiceless stop consonants at the beginning of words are especially strongly aspirated (leading to long VOT values) [Dilley et al., 1996; Lehiste, 1960; Nakatani and Dukes, 1977]. As a consequence of allophonic variation, two utterances (such as *gray ties* and *great eyes*), which have the same phonemic representation (/gretaɪz/), can differ systematically in their acoustic realization. Moreover, it is now well established that listeners can use allophonic variation to identify the location of word boundaries [Christie, 1974; Kirk, 2000; Lehiste, 1960; Nakatani and Dukes, 1977; Quené, 1993; Shatzman and McQueen, 2006; Yersin-Besson and Grosjean, 1996].

Thus one source of information for word segmentation is the detailed acoustic-phonetic information at word boundaries. However, as mentioned earlier, speaking rate affects acoustic-phonetic characteristics of the speech signal, and this alteration includes at least some of those characteristics known to play a role in word segmentation. The issue addressed in the current paper concerns the perceptual consequences of this alteration. More specifically, when the only source of information for segmentation is allophonic information (for example, when listeners are asked to determine whether a given stretch of speech was intended as *gray ties* or *great eyes*), will listeners perform equally well for fast and slow speech?

As noted earlier, it has been shown that listeners accommodate for changes in speaking rate during phonetic perception. Consider again the example of voicing, specified by VOT (reflecting degree of aspiration). As speech becomes faster the VOT values of voiceless stop consonants become shorter, and listeners accommodate for such variation

[1] In this paper we use the expressions 'word segmentation' and 'identification of word boundaries' interchangeably.

by shifting the location of the voiced-voiceless category boundary (and the location of the best voiceless exemplars) along a VOT series toward shorter VOTs [e.g. Volaitis and Miller, 1992]. It may be that the accommodation for changes in speaking rate extends beyond phonetic perception per se and operates as well when acoustic-phonetic properties are used for word segmentation. If so, then even though in fast speech voiceless stops have shorter VOTs – and are less strongly aspirated – listeners will nonetheless be able to use aspiration to segment words just as easily in fast speech as in slow speech. More generally, listeners might accommodate for the systematic variation in phonetic properties resulting from changes in speaking rate, such that the efficacy of allophonic information for word segmentation does not vary with modifications in rate.

However, there may be instances in which listeners accommodate only partially (or not at all) for rate-induced changes in acoustic-phonetic properties when using these properties for word segmentation. Again, consider the example of VOT for voiceless stop consonants. As noted above, listeners accommodate for the shorter VOTs in fast speech when using VOT to identify voiceless consonants; they shift the location of the voiced-voiceless category boundary toward shorter VOTs. But they might find it difficult to accommodate for the shorter VOT – the weakened aspiration – when using aspiration for the identification of word boundaries: weaker aspiration might provide a weaker cue to word-initial position. More generally, when allophonic properties are used to specify the location of word boundaries, segmentation may be more difficult for fast speech than for slow speech. We tested this possibility in experiment 1.

Experiment 1

The purpose of this experiment was to investigate whether speaking rate influences the ability of listeners to use allophonic information for the identification of word boundaries. We focused on naturally occurring rate variation between talkers. More specifically, we asked numerous talkers to produce two-word sequences (e.g. *gray ties*) at what they considered to be a comfortable rate, and chose the speech of 4 naturally fast talkers and 4 naturally slow talkers for use in a forced-choice perceptual identification experiment. Listeners were asked to indicate the intended segmentation for each two-word sequence (e.g. indicate whether the above sequence was intended as *gray ties, great eyes, great ties,* or *gray eyes*). We expected that listeners would be able to use the allophonic information at word boundaries to perform this task at a reasonably high level of accuracy. The main question was whether accuracy would be worse for the two-word sequences produced by the fast talkers than for those produced by the slow talkers.

Method

Subjects

Thirty Northeastern University students took part in the perceptual identification experiment and received course credit for their participation. Their mean age was 19 years.

Stimulus Materials

The stimulus materials used in the perceptual identification experiment consisted of naturally produced versions of 36 two-word sequences (henceforth termed word-pairs), recorded by 4 naturally

Table 1. Test word-pairs used in experiments 1 and 2

	C#C	C#V	V#C	V#V
/p/	wipe pink	wipe ink	why pink	why ink
	grape pail	grape ale	gray pail	gray ale
	keep part	keep art	key part	key art
/t/	might take	might ache	my take	my ache
	great ties	great eyes	gray ties	gray eyes
	neat tape	neat ape	knee tape	knee ape
/k/	bike coil	bike oil	buy coil	buy oil
	make coat	make oat	may coat	may oat
	weak cash	weak ash	we cash	we ash

fast talkers and 4 naturally slow talkers. The procedures for recording and selecting the word-pairs are described below.

Word-Pairs

The word-pairs to be recorded were originally designed by Mondini [2004] for an earlier study on word segmentation. Four kinds of word-pairs were recorded: 36 test word-pairs, 48 filler word-pairs, 12 practice-identification word-pairs, and 12 practice-identification filler word-pairs.

The 36 test word-pairs were made up of three quadruplets for each of the three voiceless stop consonants (/p/, /t/, and /k/) used as the juncture consonant. Each quadruplet contained two word-pairs with a contrastive juncture consonant (C#V, as in *great eyes*, and V#C, as in *gray ties*), one word-pair with a geminate juncture consonant (C#C, as in *great ties*), and one word-pair with no juncture consonant (V#V, as in *gray eyes*). The test word-pairs in the nine quadruplets are shown in table 1.

The 36 test word-pairs were developed by first compiling a list of 18 monosyllabic word-pairs with contrastive word juncture (9 C#V and 9 V#C). These 18 word-pairs were based on two separate lists of highly familiar words obtained from a lexical database for English (mean familiarity rating across all individual words was 6.96 on a scale from 1 to 7, as reported by Nusbaum et al. [1984]). One list comprised all monosyllabic words that ended with /p/, /t/, or /k/, with the constraint that when the final consonant was removed, a word remained (e.g. *great/gray*). The second list comprised all monosyllabic words that began with /p/, /t/, or /k/, with the constraint that when the onset consonant was removed, a word remained (e.g. *ties/eyes*). From these two lists, 18 word-pairs with contrastive word juncture (9 C#V and 9 V#C, as in *great eyes* and *gray ties*) were created, 3 C#V and 3 V#C word-pairs for each of the three juncture consonants /p/, /t/, and /k/.

The selection of these 18 test word-pairs was further constrained as follows. The first word of the word-pair was either CV(C) or CCV(C) (the contrastive consonant is in parentheses). In addition, given that according to English phonological constraints monosyllabic words can end in a long vowel but not a short vowel, the initial words in the word-pairs all contained a long vowel; these were /aɪ/, /i/, and /e/. The final words in the word-pairs were all (C)VC or (C)VCC (with no constraint on the vowel). Based on these 18 word-pairs with contrastive juncture (9 C#V and 9 V#C), two additional sets of word-pairs were constructed, a set of 9 word-pairs with geminate juncture consonants (C#C, as in *great ties*), and a set of 9 word-pairs with no juncture consonant (V#V, as in *gray eyes*). In the end, any given word-pair was part of a quadruplet that consisted of four juncture types, C#V, V#C, C#C, and V#V (as in *great eyes, gray ties, great ties, gray eyes*).

To help ensure that talkers produced the 36 test word-pairs as naturally as possible without exaggerating the juncture differences, 48 filler word-pairs were also generated. The 48 filler word-pairs consisted of 12 quadruplets of monosyllabic word-pairs made up of nouns that varied in terms of plural and possessive *s* (e.g. *pen's cap, pens' cap, pen caps, pen cap*). The filler word-pairs were not presented to listeners in the perceptual identification experiment.

Finally, 12 practice-identification word-pairs and 12 practice-identification filler word-pairs were also generated. The 12 practice-identification word-pairs, which had alternative segmentations

(e.g. *buy zinc, buys ink*), were to be used as practice stimuli for the perceptual identification experiment. The 12 practice-identification filler word-pairs were similar to the fillers described above.

The main production list to be presented to the talkers for recording was created as follows. A stimulus block consisting of the 36 test word-pairs and 48 filler word-pairs (for a total of 84 items) was repeated three times with a different within-block quasi-randomization. Care was taken to ensure that both within and across blocks, the alternate segmentations for a particular test word-pair were separated by at least 10 stimuli. After randomization, 4 practice-identification word-pairs and 4 practice-identification filler word-pairs were added to the beginning of each block in mixed order, for a total of 92 word-pairs per block. The main production list thus consisted of 276 word-pairs (92 per block x 3 blocks). A randomized practice production list for the talkers was also generated; this consisted of one instance of each of the 36 test word-pairs and 48 filler word-pairs.

Recordings

Fifty native English speakers (25 male and 25 female, mean age of 19 years) were recorded individually in a sound-treated booth. (These were different from the subjects who took part in the perceptual identification experiment.) All were Northeastern University students who received course credit for participation. Each talker's speech was recorded via a microphone (AKG C460B) onto digital audiotape (TASCAM DA-P1 DAT recorder). The word-pairs appeared on a computer screen in front of the talker, one word-pair at a time. Talkers were instructed to say each word-pair, as it appeared, in the neutral sentence frame 'He writes…' with emphasis on the first word of the sentence (*He*). Each word-pair remained on the screen for 2,750 ms, with an interstimulus interval of 1,500 ms. The talkers were asked to speak as naturally as possible, at whatever rate was most comfortable for them.

At the beginning of the session, the talker was familiarized with the recording procedure through the presentation of 8 word-pairs (different from those described above). This initial familiarization was followed by presentation of the practice production list (84 trials). Finally, the main production list (276 trials) was presented, with a break between blocks within the list. Recordings were transferred to a Pentium PC at a sampling rate of 20 kHz using hardware for the CSL system (Kay Elemetrics Corp.). Measurements were made from the sound files using Praat software [Boersma, 2001].

Talker and Token Selection

We selected one token of each of the 36 test word-pairs per talker as potential candidates for inclusion in the stimulus set to be used in the perceptual identification experiment. These tokens were selected such that there were no mispronunciations or hesitations, the two words in a given word-pair had approximately equal stress, and there was no pause between the two words of a word-pair. Two males were excluded because their productions did not yield at least one good token of each of the 36 test word-pairs. This left 48 talkers, 25 females and 23 males. For these talkers, we measured the overall duration of each selected word-pair, and then computed mean test word-pair duration over the 36 test word-pairs for each talker. A one-way analysis of variance (male vs. female) on these talker means showed a duration difference for male and female talkers [$F(1, 46) = 11.52$, $p < 0.01$], with males (mean = 757 ms) producing shorter word-pairs than females (mean = 844 ms).

So as not to confound differences in speaking rate with possible gender differences in the stimuli used for the identification experiment, we selected talkers of only one gender for the final stimulus set. Given that females showed a somewhat wider spread in test word-pair duration than males, we elected to use female talkers, and selected the 4 slowest and the 4 fastest talkers among the 25 females. The mean word-pair duration (averaged over the 36 test word-pairs) ranged from 938 to 1,050 ms for the 4 slowest talkers and ranged from 712 to 752 ms for the 4 fastest talkers. For each of these 8 female talkers, we also selected one of the practice-identification word-pairs, a different word-pair for each talker. There were thus 296 word-pairs selected in all, 8 practice-identification word-pairs (one per talker) and 288 test word-pairs (36 for each of 8 talkers). Each word-pair was extracted from its sentence frame, and the extracted word-pairs were equated for root mean square amplitude. The sound files for the word-pairs were then converted to a sampling rate of 22.05 kHz and digitally transferred to a Macintosh G3 for use in the perceptual identification experiment. For purposes of explication, we will refer to the word-pairs produced by the 4 naturally fast talkers as 'fast-normal' speech, and the word-pairs produced by the 4 naturally slow talkers as 'slow-normal' speech.

Procedure

The transferred word-pairs were used to create two stimulus sets, a practice set and a test set. The practice set consisted of the 8 practice-identification word-pairs, and the test set consisted of the 288 test word-pairs.

Listeners were run individually in a sound-treated booth. The stimuli were presented online from the Macintosh G3 using PsyScope software [Cohen et al., 1993], which also recorded the listeners' responses. The speech was presented through MDR-V6 earphones at 72 dB SPL. Listeners were instructed to listen to each word-pair (e.g. *great eyes*), to make a selection among the four possible choices that appeared in a row on the computer screen (e.g. *great eyes, great ties, gray ties, gray eyes*), and to click on their choice with the mouse. The left-to right order of the four choices was randomized across trials. The four-choice visual presentation appeared 500 ms after the end of the auditory stimulus, and the next auditory stimulus was presented 2,300 ms after the listener's mouse click.

Each session consisted of the presentation of the practice set (8 trials), followed by presentation of the test set (288 trials), with a break in the middle of the session. Each listener received a different randomization of the two sets.

Results and Discussion

For each listener, we computed percent correct identification for each of the four juncture types (C#C, C#V, V#C, and V#V) separately for the fast-normal speech and the slow-normal speech. As can be seen in figure 1, the fast-normal speech was identified less accurately than the slow-normal speech for each of the four juncture types. Overall, identification was 94.3% correct for the fast-normal speech and 96.8% correct for the slow-normal speech. It is also apparent that performance varied with juncture type, with identification best for the V#V stimuli and worst for the C#C stimuli. Two two-way analyses of variance with the factors speaking rate (fast, slow) and juncture type (C#C, C#V, V#C, V#V) performed on the arcsine-transformed data lend statistical support to these observations. For one analysis (F_1) subjects was the random factor and for the other analysis (F_2) items was the random factor. The effect of speaking rate was statistically significant in both analyses [$F_1(1, 29) = 33.61$, $p < 0.001$; $F_2(1, 32) = 7.60$, $p < 0.01$], as was the effect of juncture type [$F_1(3, 87) = 19.65$, $p < 0.001$; $F_2(3, 32) = 9.48$, $p < 0.001$]. The interaction of rate and juncture type was not significant in either analysis [$F_1(3, 87) = 1.35$, $p > 0.10$; $F_2(3, 32) < 1$], indicating that the rate effect was not modulated by juncture type.

The results of the experiment provide evidence that speaking rate can influence the identification of word boundaries, with speech produced at a faster rate identified less accurately than speech produced at a slower rate. Note, however, that the overall magnitude of the rate effect was small: the difference in accuracy for the fast-normal and slow-normal speech was only 2.5%. A question that arises is whether this difference accurately reflects the extent to which the difference between speech produced by naturally fast and naturally slow talkers can influence performance, or perhaps underestimates the influence of rate due to a ceiling effect. The overall identification accuracy (averaged across fast-normal and slow-normal speech) was very high (95.6% correct), suggesting that the task was quite easy for both the fast-normal and slow-normal speech, and that especially for the slow-normal speech, performance (96.8% correct) may have been at ceiling. To investigate whether a larger difference in identification of the fast-normal and slow-normal speech might emerge under more difficult listening conditions that precluded a ceiling effect, we conducted a second perceptual identification experiment that assessed word segmentation in the context of background noise.

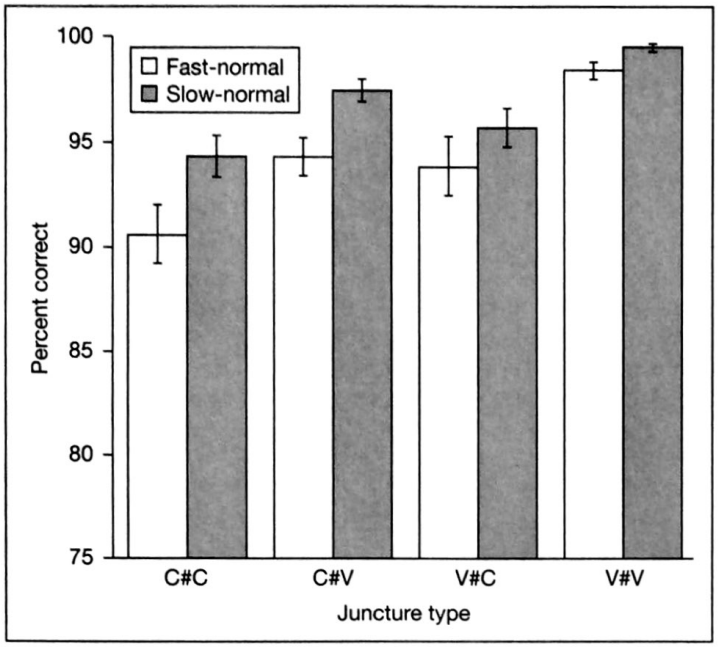

Fig. 1. Percent correct identification for each of the four juncture types, for the fast-normal and slow-normal speech, in quiet (experiment 1).

Experiment 2

Experiment 2 was identical to experiment 1, except that the word-pairs were presented in the context of background noise. Given the more difficult listening conditions, we expected to find that overall identification was more difficult for both the fast-normal and the slow-normal speech, rendering performance well below ceiling. We also expected that, as in experiment 1, the speech of the fast talkers would be identified less accurately than the speech of the slow talkers. The critical question was whether the magnitude of the difference between the fast-normal and slow-normal speech would be larger than that found in the first experiment.

Method

Subjects

Thirty new Northeastern University students took part in this experiment and received course credit for their participation. Their mean age was 19 years. One additional subject was tested but replaced because his performance was substantially lower than that of the other subjects (he scored more than two standard deviations below the mean). His replacement ensured that the error analyses performed on the data from this experiment (see below) were not disproportionately affected by the errors of one atypical subject.

Stimulus Materials and Procedure

The stimulus materials were the same as those used in experiment 1 (8 practice items and 288 test items). The procedure was also identical, except that the items were presented in a multitalker babble noise. This was a slightly modified version of the multitalker babble developed for the SPIN test [Kalikow et al., 1977]. The continuous babble noise was presented from a DAT player and mixed with the analog speech signal prior to earphone input. The babble noise was presented at 70 dB SPL and the speech at 67 dB SPL, yielding a signal-to-noise ratio of −3 dB.

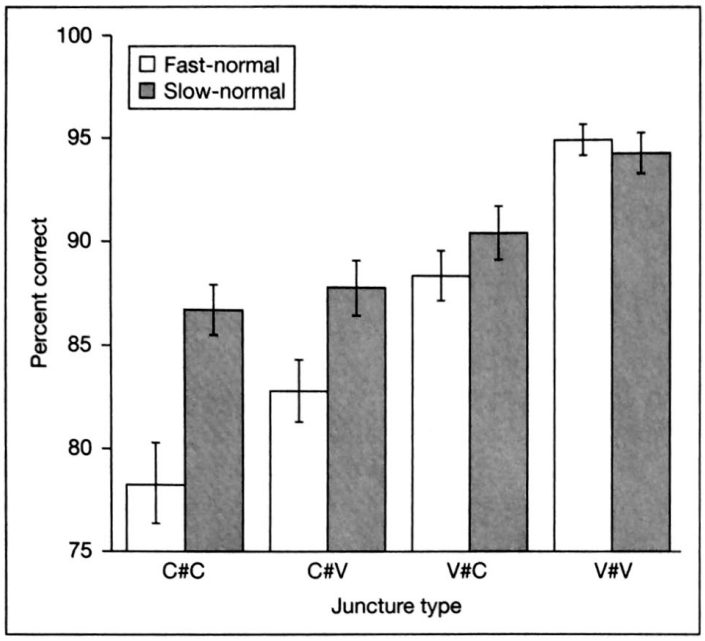

Fig. 2. Percent correct identification for each of the four juncture types, for the fast-normal and slow-normal speech, in noise (experiment 2).

Results and Discussion

As expected, overall identification performance was worse in this experiment (88.0% correct) than in the first experiment (95.6% correct). However, as shown in figure 2, the pattern of results was similar in the most relevant respects to that found in the first experiment. Performance again varied with juncture type, with identification best for the V#V stimuli and worst for the C#C stimuli. Most important, identification was again worse for the fast-normal speech (86.2%) than for the slow-normal speech (89.9%), and this effect was generally found across juncture type (with a slight reversal for the V#V stimuli).

As in experiment 1, two two-way analyses of variance with the factors speaking rate (fast, slow) and juncture type (C#C, C#V, V#C, V#V) were performed on the arcsine-transformed data, one with subjects as the random factor (F_1) and one with items as the random factor (F_2). The effect of speaking rate was again statistically significant in both analyses [$F_1(1, 29) = 21.86$, $p < 0.001$; $F_2(1, 32) = 4.18$, $p < 0.05$], as was the effect of juncture type [$F_1(3, 87) = 34.47$, $p < 0.001$; $F_2(3, 32) = 6.72$, $p < 0.005$]. The interaction of rate and juncture type was significant in the subjects analysis [$F_1(3, 87) = 5.44$, $p < 0.005$]. However, it failed to reach significance in the items analysis [$F_2(3, 32) = 1.53$, $p > 0.10$], indicating that any modulation of the rate effect by juncture type was not robust across stimulus items.

The most important finding of experiment 2 is that in noise, as in quiet, overall identification performance varied as a function of speaking rate. Moreover, the magnitude of the rate effect was similar in the two experiments, with a 3.7% difference in accuracy between the fast-normal and slow-normal speech in the current experiment, compared to a 2.5% difference in accuracy between these two types of speech in the first experiment. Two analyses of variance on the combined arcsine-transformed data from the two experiments (collapsing across juncture type), one by subjects (F_1) and

one by items (F_2), provide statistical support for this observation. The two factors in each analysis were speaking rate (fast, slow) and listening condition (quiet, noise). The analyses showed significant effects of both speaking rate [$F_1(1, 58) = 59.60$, $p < 0.001$; $F_2(1, 35) = 8.53$, $p < 0.01$) and listening condition [$F_1(1, 58) = 84.13$, $p < 0.001$; $F_2(1, 35) = 65.24$, $p < 0.001$], but no interaction between the two factors [$F_1(1, 58) < 1$; $F_2(1, 35) < 1$]. Thus, the change in rate had the same magnitude of effect on identification in quiet and in noise.

In a final analysis, which focused on the data from experiment 2, we examined the pattern of errors that occurred for the fast-normal and slow-normal speech. (We did not conduct a comparable analysis for the data in experiment 1 because there were so few errors overall.) There is evidence from the spoken word recognition literature suggesting that word onsets may play an especially important role in lexical access [for discussion, see Gow and Gordon, 1995]. More directly, previous research has indicated that postjuncture word-initial information is perceptually more salient than prejuncture word-final information in specifying the location of word boundaries [Nakatani and Dukes, 1977; Quené, 1993]. Given these findings, we might expect that for the word-pairs tested in the current experiment, errors will tend to preserve the word-initial phonetic segment of the second word in the word-pair. For example, consider the C#V juncture type. Three possible error responses are possible, C#C, V#C, and V#V. Only one of these, V#V, preserves the word-initial segment (V). Thus we would expect that V#V errors would predominate for a C#V stimulus.

To determine whether such a pattern held for our data, we examined the type of error response (C#C, C#V, V#C, V#V) that occurred for each stimulus type (C#C, C#V, V#C, V#V), and did so separately for the fast-normal and slow-normal speech. Table 2 presents the results of this analysis. For each stimulus type, the percentage of errors that occurred for each possible response type is shown. As is apparent, for all eight cases (four stimulus types at two speaking rates), the errors are not equally distributed across the possible response types. This observation was confirmed by separate χ^2 analyses for each of the eight cases on the number of errors across response type; in all cases, $\chi^2(2) > 17$, $p < 0.001$ (corrected for multiple tests, $\alpha = 0.006$). Moreover, the distribution of errors in all eight cases is consistent with the primacy of word-initial information for segmentation. For both the fast-normal and slow-normal speech, C#C stimuli yielded mostly V#C errors, C#V stimuli yielded mostly V#V errors, V#C stimuli yielded mostly C#C errors, and V#V stimuli yielded mostly C#V errors. Thus across speaking rates, errors tended to preserve the postjuncture phonetic segment.[2]

[2] The question arises as to whether the error pattern could be due to differences in the familiarity of the word-pairs. Although, as noted earlier, the 18 individual words used to construct the 9 word-pairs were all highly familiar, given the many constraints on word-pair construction, the word-pairs themselves clearly varied in familiarity. To obtain a rough estimate of familiarity, we used the Altavista search engine on the internet to determine the frequency of each of the 9 word-pairs in its database. The word-pairs received over 18 million hits and (collapsing across word-pairs) the ordering of stimulus types, from most to least frequent, was C#C, V#C, V#V, C#V. This ordering does not account for the relative accuracy in identifying these stimulus types: in both experiments 1 and 2, identification was least accurate for the C#C stimuli and most accurate for the V#V stimuli (for both the fast-normal and slow-normal stimuli). Moreover, this ordering does not account for the specific error pattern found in experiment 2: in two cases the most frequent response type was more frequent than the stimulus type (V#V for C#V and C#C for V#C), but in the other two cases it was less frequent (V#C for C#C and C#V for V#V). Thus although we cannot rule out the possibility that the familiarity of the word-pairs influenced the specific pattern of results in some way, it is unlikely that it was the major influence. Importantly, familiarity could not have accounted for the main finding, a decrease in accuracy for fast compared to slow speech, given that precisely the same word-pairs were tested at the two speaking rates.

Table 2. Distribution of error response types (in percentage) for each stimulus type, for fast-normal and slow-normal speech, in experiment 2

a Fast-normal speech

Stimulus type	Error response type			
	C#C	C#V	V#C	V#V
C#C	–	22	77	1
C#V	12	–	25	63
V#C	49	40	–	11
V#V	4	65	31	–

b Slow-normal speech

Stimulus type	Error response type			
	C#C	C#V	V#C	V#V
C#C	–	29	69	1
C#V	18	–	20	63
V#C	59	16	–	25
V#V	8	49	43	–

General Discussion

The purpose of this investigation was to examine whether speaking rate influences a listener's ability to identify the location of word boundaries. Earlier research has shown that one source of information listeners can use for word segmentation is allophonic variation: talkers pronounce some aspects of consonants and vowels differently depending on their position in a word, and listeners can use this information to help segment the stream of speech [e.g. Nakatani and Dukes, 1977]. It is also known that a change in speaking rate can alter the precise acoustic-phonetic characteristics of the individual consonants and vowels of the language [Miller, 1981], and does so in a way that causes at least some properties that can specify the location of word boundaries to become less distinctive [e.g. Miller et al., 1986]. Thus word segmentation based on allophonic information may become more difficult when speaking rate becomes faster. We tested this possibility in two experiments.

In both experiments, we presented listeners with two-word sequences that potentially had alternate segmentations involving voiceless stop consonants, and asked them to choose the intended segmentation in a forced-choice identification task. The naturally produced speech of 8 talkers was tested. The talkers had been instructed to produce the sequences at whatever rate was comfortable for them, and we chose the speech of 4 'fast talkers' and 4 'slow talkers' for the identification experiment. In the first experiment the two-word sequences were presented in quiet, and in the second experiment they were presented in the context of a multitalker babble noise. As expected, overall performance was poorer in noise than in quiet. Of critical importance, in both experiments there was a disadvantage for the speech of the fast talkers.

An examination of the error patterns for the second experiment indicated that for the speech of both the fast and the slow talkers, the errors made by listeners tended

to preserve word-initial information. This pattern of findings is in accord with previous research indicating that postjuncture word-initial information is more salient than prejuncture word-final information in specifying the location of word boundaries [Nakatani and Dukes, 1977; Quené, 1993] and, furthermore, indicates that the pattern holds across variation in rate. Thus although changes in rate can influence how well listeners use allophonic information to segment speech, they seem not to produce a qualitative change in listeners' relative use of postjuncture versus prejuncture information for locating word boundaries.

Taken together, the results of the two experiments provide clear evidence that speaking rate can affect the ability of listeners to use allophonic information for word segmentation, with poorer segmentation for faster speech. Thus listeners do not always accommodate fully for rate-induced changes in acoustic-phonetic properties, at least with respect to locating the boundaries between words.

The influence of speaking rate in the current experiments is particularly noteworthy given that the two-word sequences used as stimuli were all produced at what the talkers themselves considered to be a comfortable rate; the talkers were not induced to talk quickly or slowly. Thus natural variation in speaking rate across talkers is sufficient to produce changes in how well listeners can use acoustic-phonetic detail to segment speech. Note, however, that although the effect of speaking rate was statistically significant in both experiments, it was quite small. Averaging across the two experiments, there was a 3.1% advantage for the speech of the slow talkers over that of the fast talkers (the magnitude of the effect did not differ statistically across the experiments). Whether larger rate effects might be found with different stimulus or task conditions remains to be determined. For example, larger rate effects might emerge with larger differences in rate across talkers, or if rate were manipulated within talkers, by instructing them to vary speaking rate from very fast to very slow. Such manipulations would presumably magnify any difference in the distinctiveness of allophonic information across changes in rate, and this in turn could lead to a larger overall effect of rate on word segmentation. Task changes might also lead to larger rate effects. For example, an open-ended response format, rather than a forced-choice response format, might exacerbate the relative difficulty in segmenting fast speech, leading to a larger overall effect. Note that if larger effects do emerge under different conditions, it will be of particular interest to determine whether the pattern of errors found in the current study is maintained.

It is likely that many properties of the particular two-word sequences we used as stimuli were used by listeners to segment the speech into words. This presumably includes the degree of aspiration in the juncture consonant (/p, t, k/), as well as the degree of laryngealization/glottalization of the adjacent vowel. The precise nature of the critical allophonic variation listeners used and, most importantly, how this information changed across speaking rate are beyond the scope of the present paper and remain to be determined. Note that this will require not only detailed acoustic analyses, but also experiments that systematically manipulate potentially important allophonic information across changes in rate in order to assess the listener's use of such information for word segmentation. And, of course, it will also be important to examine how speaking rate influences other types of allophonic variation at word boundaries, not just that involving voiceless stop consonants. The present findings provide the basis for such future analyses and experiments by demonstrating that speaking rate can influence the efficacy of allophonic information for segmenting words.

Finally, the present findings have implications for models of word recognition. As noted in the introduction, a prominent view in the field is that word segmentation per se does not precede word recognition, but that it is a by-product of the word recognition process. Within this view, many argue that word recognition (and hence word segmentation) occurs through activation and competition among lexical candidates that share potential word boundaries, and some adherents of this view propose that sublexical information, including allophonic variation, can modulate this activation-competition process, with increased activation occurring in favor of lexical candidates that are consistent with the sublexical information [e.g. Shatzman and McQueen, 2006]. The current findings raise the possibility that the extent to which allophonic information modulates the process might itself depend on the rate at which the speech was produced, with a larger role in slower speech.

Acknowledgements

The authors thank Eliza Floyd and Timothy Hawes for their important contributions to this project. The research was supported by a Fellowship awarded to the first author by the Swiss National Science Foundation, and by a grant awarded to the second author by the National Institute on Deafness and Other Communication Disorders (NIH R01 DC000130).

References

Allen, J.S.; Miller, J.L.; DeSteno, D.: Individual talker differences in voice-onset-time. J. acoust. Soc. Am. *113*: 544–552 (2003).

Boersma, P.: Praat, a system for doing phonetics by computer. Glot int. *5*: 341–345 (2001).

Bradlow, A.R.; Pisoni, D.B.: Recognition of spoken words by native and non-native listeners: talker-, listener-, and item-related factors. J. acoust. Soc. Am. *106*: 2074–2085 (1999).

Buttet, J.; Wingfield, A.; Sandoval, A.W.: Effets de la prosodie sur la résolution syntaxique de la parole comprimée. Année psychol. *80*: 33–50 (1980).

Christie, W.M. Jr.: Some cues for syllable juncture perception in English. J. acoust. Soc. Am. *55*: 819–821 (1974).

Cohen, G.: Language comprehension in old age. Cognit. Psychol. *11*: 412–429 (1979).

Cohen, J.D.; MacWhinney, B.; Flatt, M.; Provost, J.: PsyScope: an interactive graphic system for designing and controlling experiments in the psychology laboratory using Macintosh computers. Behav. Res. Methods Instrum. Comput. *25*: 257–271 (1993).

Cole, R.A.; Jakimik, J.: A model of speech perception; in Cole, Perception and production of fluent speech, pp. 133–163 (Erlbaum Associates, Hillsdale 1980).

Dilley, L.; Shattuck-Hufnagel, S.; Ostendorf, M.: Glottalization of word-initial vowels as a function of prosodic structure. J. Phonet. *24*: 423–444 (1996).

Dupoux, E.; Green, K.: Perceptual adjustment to highly compressed speech: effects of talker and rate changes. J. exp. Psychol. hum. Perception Performance *23*: 914–927 (1997).

Fairbanks, G.; Guttman, N.; Miron, M.S.: Effects of time-compression upon the comprehension of connected speech. J. Speech Hear. Disorders *22*: 10–19 (1957).

Goldman-Eisler, F.: Psycholinguistics: Experiments in spontaneous speech (Academic Press, London 1968).

Gow, D.W.; Gordon, P.C.: Lexical and prelexical influences on word segmentation: evidence from priming. J. exp. Psychol. hum. Perception Performance *21*: 344–359 (1995).

Grosjean, F.; Deschamps, A.: Analyse des variables temporelles du français spontané. Phonetica *26*: 129–156 (1972).

Grosjean, F.; Deschamps, A.: Analyse des variables temporelles du français spontané. 2. Comparaison du français oral dans la description avec l'anglais (description) et avec le français (interview radiophonique). Phonetica *28*: 191–226 (1973).

Grosjean, F.; Deschamps, A.: Analyse contrastive des variables temporelles de l'anglais et du français: vitesse de parole et variables composantes, phénomènes d'hésitation. Phonetica *31*: 144–184 (1975).

Kalikow, D.N.; Stevens, K.N.; Elliott, L.L.: Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability. J. acoust. Soc. Am. *61*: 1337–1351 (1977).

Kessinger, R.H.; Blumstein, S.E.: Effects of speaking rate on voice-onset time in Thai, French, and English. J. Phonet. *25*: 143–168 (1997).

Kirk, C.: Syllabic cues to word segmentation; in Cutler, McQueen, Zondervan, Proceedings of the workshop on spoken word access processes, pp. 131–134 (Max Planck Institute for Psycholinguistics, Nijmegen 2000).
Lehiste, I.: An acoustic-phonetic study of internal open juncture. Phonetica *5* (Suppl.): 1–54 (1960).
Lucci, V.: Etude phonétique du français contemporain à travers la variation situationnelle (Université des langues et lettres de Grenoble, Grenoble 1983).
Marslen-Wilson, W.D.; Welsh, A.: Processing interactions and lexical access during word recognition in continuous speech. Cognit. Psychol. *10*: 29–63 (1978).
McClelland, J.L.; Elman, J.L.: The TRACE model of speech perception. Cognit. Psychol. *18*: 1–86 (1986).
McQueen, J.M.: Segmentation of continuous speech using phonotactics. J. Mem. Lang. *39*: 21–46 (1998).
Miller, J.L.: Effects of speaking rate on segmental distinctions; in Eimas, Miller, Perspectives on the study of speech, pp. 39–74 (Erlbaum Associates, Hillsdale 1981).
Miller, J.L.; Green, K.P.; Reeves, A.: Speaking rate and segments: a look at the relation between speech production and speech perception for the voicing contrast. Phonetica *43*: 106–115 (1986).
Miller, J.L.; Grosjean, F.; Lomanto, C.: Articulation rate and its variability in spontaneous speech: a reanalysis and some implications. Phonetica *41*: 215–225 (1984).
Miller, J.L.; Volaitis, L.E.: Effect of speaking rate on the perceptual structure of a phonetic category. Perception Psychophysics *46*: 505–512 (1989).
Mondini, M.: Perceiving non-native speech: word segmentation; PhD diss. Northeastern University (unpublished, 2004).
Nakatani, L.H.; Dukes, K.D.: Locus of segmental cues for word juncture. J. acoust. Soc. Am. *62*: 714–719 (1977).
Nicholas, L.E.; Brookshire, R.H.: Consistency of the effects of rate of speech on brain-damaged adults' comprehension of narrative discourse. J. Speech Hear. Res. *29*: 462–470 (1986).
Norris, D.: Shortlist: a connectionist model of continuous speech recognition. Cognition *52*: 189–234 (1994).
Norris, D.; McQueen, J.M.; Cutler, A.: Competition and segmentation in spoken-word recognition. J. exp. Psychol. Learn. Mem. Cogn. *21*: 1209–1228 (1995).
Norris, D.; McQueen, J.M.; Cutler, A.; Butterfield, S.: The possible-word constraint in the segmentation of continuous speech. Cognit. Psychol. *34*: 191–243 (1997).
Nusbaum, H.C.; Pisoni, D.B.; Davis, C.K.: Sizing up the hoosier mental lexicon: measuring the familiarity of 20,000 words. Indiana University. Res. Speech Percept. Progress Rep. *10*: 357–376 (1984).
Quené, H.: Segment durations and accent as cues to word segmentation in Dutch. J. acoust. Soc. Am. *94*: 2027–2035 (1993).
Sebastián-Gallés, N.; Dupoux, E.; Costa, A.; Mehler, J.: Adaptation to time-compressed speech: phonological determinants. Perception Psychophysics *62*: 834–842 (2000).
Shatzman, K.B.; McQueen, J.M.: Segment duration as a cue to word boundaries in spoken-word recognition. Perception Psychophysics *68*: 1–16 (2006).
Sommers, M.S.; Nygaard, L.C.; Pisoni, D.B.: Stimulus variability and spoken word recognition. 1. Effects of variability in speaking rate and overall amplitude. J. acoust. Soc. Am. *96*: 1314–1324 (1994).
Summerfield, Q.: Articulatory rate and perceptual constancy in phonetic perception. J. exp. Psychol. hum. Perception Performance *7*: 1074–1095 (1981).
Vaughan, N.E.; Letowski, T.: Effects of age, speech rate, and type of test on temporal auditory processing. J. Speech Lang. Hear. Res. *40*: 1192–1200 (1997).
Volaitis, L.E.; Miller, J.L.: Phonetic prototypes: influence of place of articulation and speaking rate on the internal structure of voicing categories. J. acoust. Soc. Am. *92*: 723–735 (1992).
Wayland, S.C.; Miller, J.L.; Volaitis, L.E.: The influence of sentential speaking rate on the internal structure of phonetic categories. J. acoust. Soc. Am. *95*: 2694–2701 (1994).
Wingfield, A.; Tun, P.A.; Koh, C.K.; Rosen, M.J.: Regaining lost time: adult aging and the effect of time restoration on recall of time-compressed speech. Psychol. Aging *14*: 380–389 (1999).
Yersin-Besson, C.; Grosjean, F.: L'effet de l'enchaînement sur la reconnaissance des mots dans la parole continue. Année psychol. *96*: 9–30 (1996).