Perception & Psychophysics
2007, 69 (6), 856-864

Statistical information and coarticulation as cues to word boundaries: A matter of signal quality

TÂNIA FERNANDES AND PAULO VENTURA
University of Lisbon, Lisbon, Portugal

AND

RÉGINE KOLINSKY
Fonds de la Recherche Scientifique and Free University of Brussels, Brussels, Belgium

We investigated how statistical information in the form of transitional probabilities (TPs) interacts with coarticulation, another sublexical cue to word boundaries, and examined the impact of signal quality on the weighting of these cues. In an artificial-language-learning setting, with phonetically intact speech, coarticulation overruled TPs, suggesting the predominance of subsegmental, low-level information. However, whereas the role of coarticulation in segmentation was highly modulated by signal quality, TPs were very resilient to noise. When coarticulation was rendered unreliable by strongly degrading the input with a 10-dB signal-to-noise ratio (SNR), only statistical information drove segmentation. In a more mildly degraded 22-dB SNR condition, in which some acoustic properties were still available, coarticulation was exploited, although with less reliability than in optimal conditions. These results can be interpreted according to a hierarchical approach (Mattys, White, & Melhorn, 2005) in which both the available segmentation cues and the listening conditions have an important role in speech segmentation.

Understanding how listeners locate word boundaries in fluent speech is a hard task, since the speech signal includes few reliable cues to word boundaries (see, e.g., Klatt, 1980; Liberman & Studdert-Kennedy, 1978). Several signal-derived sources of information have nevertheless been pointed out as potential cues to word boundaries, and combining these cues to lexical-driven mechanisms may be helpful to word segmentation (see, e.g., McQueen, Norris, & Cutler, 1994; Norris, McQueen, & Cutler, 1995). However, since each of these signal-derived cues is only probabilistically associated with word boundaries, it is essential to study multiple-segmentation cues not only in isolation but also in conjunction if we aim to obtain a realistic perspective on how listeners deal with them.

Furthermore, according to Mattys and colleagues (Mattys, 2004; Mattys, White, & Melhorn, 2005), the involvement of any segmentation cue, either lexically or signal-derived, is a graded rather than an all-or-none phenomenon. Mattys and colleagues defined a hierarchical organization[1] comprising three tiers. The first or top tier consists of lexical and postlexical knowledge, which is supposed to be the most reliable information in optimal listening conditions. Sublexical types of information—subsegmental, segmental, and suprasegmental—are confined to the lower tiers and are called upon when lexical information is unavailable or reduced. The second tier consists of the conjunction of segmental and subsegmental information. The lowest tier corresponds to metrical prosody, which acts as a last-resource segmentation heuristic, prevailing over the available information from the above tiers when lexical knowledge is not available (Mattys et al., 2005) and the signal is degraded, which impoverishes segmental and subsegmental information.

However, since Mattys and colleagues' (Mattys, 2004; Mattys et al., 2005) framework is quite recent, some of the possible interactions among the several information types available to listeners remain underspecified. For example, it is not clear how the various types of sublexical information interact with each other. Mattys et al. showed that when coarticulation (subsegmental) and phonotactic (segmental) information (both on Tier 2) were available, indicating the same segmentation points, their segmentation hypotheses were the ones used by adult listeners, in two scenarios: (1) when lexical information was not available in the signal; and (2) when signal quality was degraded enough that reliance on the semantic context was reduced, but the acoustic–phonetic aspects were still relatively available. However, because in Mattys et al.'s study phonotactic and coarticulatory information were never put in conflict, we do not know whether one of these two information types prevailed and, if so, which one.

In addition, Mattys and colleagues (Mattys, 2004; Mattys et al., 2005) have posited that segmental and subsegmental information were represented on the same tier;

P. Ventura, paulo.ventura@fpce.ul.pt

consequently, their model confers on these two information types the same importance. It is plausible, however, that even sublexical types of information represented on the same tier have independent impacts on segmentation and differ by their degree of dependence on or independence from signal quality. Indeed, as already posited by Mattys and colleagues, segmental and subsegmental types of information may be weighted differently according to their word boundaries' predictability, at least in some languages. This proposal was investigated in the present study with regard to coarticulation, a subsegmental, acoustic-dependent cue, and transitional probabilities, a statistical cue related to a higher structural level (here, syllabic).

Coarticulation is usually defined as a change in the acoustic–phonetic content of a speech segment due to anticipation or preservation of adjacent segments (see, e.g., Kühnert & Nolan, 1999). Although all fluent speech is coarticulated, the extent of coarticulation between adjacent segments is influenced by the presence of a prosodic boundary. Generally, there is more coarticulation within words than between them (see, e.g., Byrd, 1996; Byrd & Saltzman, 1998), and segment strength may convey information about local coherence versus disjunction in connected speech (Fougeron & Keating, 1997; Keating, 2006). Indeed, domain-initial strengthening is a general phenomenon, found in both stressed and unstressed syllables (Cho & McQueen, 2005).

The importance of coarticulation in speech segmentation has been demonstrated, in optimal listening conditions, from an early age (Johnson & Jusczyk, 2001) into adulthood (Mattys, 2004; Mattys et al., 2005). However, coarticulation seems to be strongly affected by the superimposition of noise on the speech input (Mattys, 2004). The main reason for this sensitivity to noise is possibly related to the nature of coarticulatory information (Mattys et al., 2005), whose use depends on the ability to process fine-grained, low-level acoustic properties. Thus, noise probably masks such information easily, reducing the effectiveness of coarticulation in speech segmentation.

Another important source of information that can help locate word boundaries is statistical information conveyed by the speech stream. The extraction of statistical information is a general mechanism (see, e.g., Conway & Christiansen, 2005; Saffran, Johnson, Aslin, & Newport, 1999) that seems to be universal in nature. As a matter of fact, within any language, the transitional probability (TP) from one unit of sound (e.g., a syllable) to the next is generally highest when the two units follow one another within a word, whereas the TPs that span a word boundary are relatively low. Human listeners in different stages of linguistic development can use this statistical information to locate "word" boundaries within an artificial language (AL) that consists of a continuous stream with no other segmentation cues (see, e.g., Saffran, Aslin, & Newport, 1996; Saffran, Newport, & Aslin, 1996).[2]

Until now, the question of how such statistical cues interact in speech segmentation with lower level cues like those linked to coarticulation has been left unsolved in the context of a completely mature speech system. Indeed, to the best of our knowledge, no study of adult listeners has ever evaluated the relative power of these cues within the same experiment. We already know, however, that with a high-quality signal, 8-month-old infants consider coarticulation to be a more reliable cue than TPs when coarticulation and TPs indicate different segmentation hypotheses (Johnson & Jusczyk, 2001). Whether this coarticulation preference holds true in the mature system and whether the observed weighting would be modulated by signal quality, which may be predicted by Mattys and colleagues' (Mattys, 2004; Mattys et al., 2005) hierarchical framework, was examined in the present study.

To this end, we adopted an AL paradigm, which has proved useful for understanding how adult listeners segment natural speech. Indeed, electrophysiological correlates (N100 amplitude enhancement; Sanders, Newport, & Neville, 2002) and neural signatures (left-lateralized fMRI signal increases in temporal cortices; McNealy, Mazziotta, & Dapretto, 2006) of online word segmentation in AL learning suggest that processes that may be central to speech segmentation are called upon in the segmentation of an AL. In addition, adult listeners attempt to integrate the output of statistical learning with knowledge of their native language acquired prior to the experimental task (see, e.g., Shukla, Nespor, & Mehler, 2007, and Vroomen, Tuomainen, & de Gelder, 1998, on prosodic properties; Onnis, Monaghan, Richmond, & Chater, 2005, on phonotactic probabilities). Since the AL paradigm provides precise control over the information available in the input, allowing reduction of the set of uncontrolled variables (see, e.g., Gómez & Gerken, 2000), it is ideally suited for studying the weighting of various sublexical cues in the absence (or near absence) of high-level information.

Nine groups of participants were presented with AL-learning situations (cf. Saffran, Newport, & Aslin, 1996). The groups were exposed to different levels of input intelligibility—namely, intact, mildly degraded, or strongly degraded speech. Within each input intelligibility condition, the same AL was presented in three conditions that differed according to the number and congruence of the segmentation cues available in the speech stream. To achieve realistic coarticulation, we used naturally produced utterances rather than synthesized speech and created concatenated versus coarticulated versions of the AL stimuli. In the single-cue condition, only TPs could help the listeners to locate word boundaries, since only the concatenated version of the AL was used. In the other two conditions, coarticulatory information was added. For these coarticulated versions, the TP-words in the congruent-cues condition (i.e., the stimuli that corresponded to the word boundaries defined by TPs) and the part-words in the incongruent-cues condition (i.e., the stimuli that crossed the TP-words' boundaries) were recorded in their coarticulated versions. Thus, in the congruent-cues condition, coarticulation and TPs pointed to the same word boundaries, whereas in the incongruent-cues condition, the two cues pointed to different word boundaries, since coarticulation corresponded to the part-words of the AL. This framework allowed us not only to explore the relative power of the different cue types when they were in conflict with each other in the incongruent case but also to estimate the performance gain afforded by redundant (and hence potentially cooperating) cues in the congruent case.

If, like infants (Johnson & Jusczyk, 2001), in good listening conditions, adult listeners also consider coarticulatory cues to be more reliable than statistical cues, then in the forced-choice test aimed at estimating how well the listeners learned the "words" of the AL, one would observe not only a performance gain in the congruent-cues condition (i.e., more TP-word choices) but also a strong decrease in TP-word choices in the incongruent cues condition in comparison with the single-cue condition (because listeners would consider the part-words to be the lexical units of the AL). If, on the contrary, statistical information was considered more reliable, then the TP-words would always tend to be considered to be the lexical units of the AL.

Even if the weighting of these cues in intact speech revealed itself to be similar to that found in infants (Johnson & Jusczyk, 2001), it is possible that the weighting is modulated by signal quality. Indeed, in adults, Mattys and colleagues (Mattys, 2004; Mattys et al., 2005) demonstrated that listeners undervalued coarticulatory cues in conditions of noise superimposition. In the present work, we used two levels of input degradation to examine whether statistical information was more resilient to noise than coarticulatory cues were.

The resilience of statistical information to noise may actually be linked to the special role of statistical information in development. Indeed, since the computation of TPs does not require any (not even minimal) lexical knowledge, TPs may provide infants with their first window into the acoustic regularities of their native language and thus may play a central role in speech acquisition (Thiessen & Saffran, 2003). This general segmentation mechanism may "occupy a pivotal position in the acquisition of not only words, but also other word boundary cues, such as stress, phonotactics, coarticulation, and allophony" (Mattys et al., 2005, p. 493). Hence, one would expect such a fundamental mechanism to be relatively immune to signal degradation.

However, at a high level of noise such as the 10-dB signal-to-noise ratio (SNR)[3] used in the present experiment, the fine-grained acoustical information linked to coarticulation may be almost unavailable. Thus, observing that TPs have more impact on the word segmentation process than coarticulation has would not be surprising. But in more mildly degraded conditions such as the 22-dB SNR condition also used in the present study,[4] the subsegmental information is still audible, and coarticulation could still facilitate segmentation (Mattys et al., 2005, Experiment 6B). We were thus interested in determining whether, in the latter condition, the mere presence of noise would alter the reliability that listeners attribute to low-level subsegmental cues like those linked to coarticulation. If this were the case, noise superimposition might allow statistical cues to override coarticulatory cues even at mild levels of noise.

METHOD

Participants

Eighty-seven undergraduate psychology students at the University of Lisbon participated in the experiment for course credit. All were monolingual European-Portuguese speakers, with no reported history of speech or hearing disorders. They were randomly assigned to nine groups according to the 3 (input intelligibility) × 3 (available segmentation cues) experimental design: 32 to the intact speech condition (9 in the single-cue, 11 in the congruent-cues, and 12 in the incongruent-cues condition); 28 to the mildly degraded (22-dB SNR) condition (6 in the single-cue, 10 in the congruent-cues, and 12 in the incongruent-cues condition), and 27 to the strongly degraded (10-dB SNR) condition (9 in each cue condition).

Materials

All natural speech stimuli were recorded in a soundproof room by a female native European-Portuguese speaker, sampled at a rate of 22.05 kHz and 16-bit conversion. The digitized versions of the stimuli (i.e., syllables, words, and part-words) were edited with Adobe Audition 1.5 and Praat 4.2.24 softwares (the latter available at www.fon.hum.uva.nl/praat/). These natural stimuli were closely matched to a synthesized stream (see the discussion of the pretest, below) in all relevant acoustic parameters, such as speaking rate and absence of lexical stress cues to word boundaries. This also allowed us to match the concatenated and coarticulated versions of the stimuli (see below) on mean duration (by applying a maximum compression rate of 30%) and on pitch contour (by flattening to a monotone 220 Hz).

The repertoire of the AL comprised four consonant sounds (/b/, /k/, /l/, and /f/) and three vowel sounds (/ɐ/, /i/, and /u/), which when combined yielded 12 possible CV syllables. The syllables of the AL were combined to create six trisyllabic TP-words: /bɐbuku/, /bukɐlɐ/, /lufɐbɐ/, /kɐfubi/, /fufibu/, and /kilɐbu/). The TPs of the trisyllables (words and trisyllabic part-words of the AL) were computed by averaging the two TPs associated with each stimulus. The TPs between adjacent syllables were always higher within words than between them, with the mean TP of adjacent syllables between words being .38. Since some syllables occurred more often than others, three words presented higher TPs (from .75 to 1.00) than did the other three (from .50 to .58). This distributional gradient probably mirrors what happens in natural languages (Saffran, Newport, & Aslin, 1996) and might allow a fine-grained evaluation of the effect of TPs in the learning process.

For the familiarization phase (which was syllable based), concatenated exemplars and coarticulated exemplars (which were based on TP-words or part-words) of the AL stimuli were constructed from natural speech. Three versions of the AL were created. Each version included the same sequence of syllables divided into three listening blocks of approximately 7-min duration each. Each block was created by concatenating 105 tokens of each of the six words (1,890 syllables, 630 tokens), with the criterion that two tokens of the same word never occurred adjacently in the stream.

The three natural versions of the AL differed with regard to the types of information available in the speech stream, as explained in the introduction and illustrated in Table 1, which presents an orthographic translation of a sample of the speech stream in each of the three versions of the AL.

A pretest checked whether using naturally produced utterances, chosen here to achieve realistic coarticulation, made the task easier (Thiessen & Saffran, 2003) than using synthesized speech, which is more common in AL learning experiments. Synthetic stimuli (created using text-to-speech MBROLA software; cf. Dutoit, Pagel, Pierret, Bataille, & van der Vrecken, 1996) were presented to 21 independent volunteer undergraduates in the same AL experiment as the main one, except that only the single-cue condition was used. Syllables were concatenated (with no other cues to word boundaries than TPs) using a European-Portuguese female diphone database (available at www.tcts.fpms.ac.be/synthesis/mbrola.html) at

Table 1
Orthographic Translation of a Sample of the Speech Stream Heard in the Familiarization Phase in the Three Cue Conditions

Cue Conditions		Part-Words	
(Available Cues)	Speech Stream	1 Syllable#2 Syllables	2 Syllables#1 Syllable
Single cue (TPs)	lu-fa-ba-#ki-la-bu-#ka-fu-bi-#ba-bu-ku-#bu-ka-la-#fu-fi-bu . . .	ba-#ki-la	ka-la-#fu
Congruent cues (coarticulated TP words)	LUFABA-#ki-la-bu-#ka-fu-bi-#ba-bu-ku-#bu-ka-la-#FUFIBU . . .	BA-#ki-la	ka-la#FU
Incongruent cues (coarticulated WPCs)	lu-fa-BA#KILA-bu-#ka-fu-bi-#ba-bu-ku-#bu-KALA#FU-fi-bu . . .	BA#KILA	KALA#FU

Note—"#" defines word boundaries according to transitional probabilities (TPs); "-" represents concatenation; uppercase letters represent coarticulated syllables.

22.05 kHz and with a speech rate of approximately 270 syllables/min. We evaluated two input intelligibility conditions (intact speech and 22-dB SNR), since in the more difficult (noisy) situation, a facilitation effect might appear for natural speech in comparison with synthetic speech. This was not the case. The TP-word preference was similar in the synthesized and in the natural single-cue conditions in the two signal intelligibility situations: With intact speech, $t(19) = .73$; with 22-dB SNR, $t(13) = .56$; $p > .10$ in both cases. Thus, in both intact and noisy conditions, the natural speech stimuli used in the main experiment induced statistical learning based on TPs in the same way that the synthetic material more commonly used in AL learning studies did.

In order to create the two degraded-signal conditions, white noise was superimposed on each block of all three natural versions of the AL at 22-dB SNR and 10-dB SNR. These SNRs were selected on the basis of a pretest run on an independent group of 23 volunteer undergraduate students, with five between-subjects conditions: intact speech, 22-dB SNR, 10-dB SNR, 5-dB SNR, and 0-dB SNR. All the words and part-words of the AL were presented in randomized order, one at a time. They were played through headphones at 76-dB sound pressure level (SPL), which is approximately the level of conversational speech. Participants were informed that on each trial they would hear a pronounceable trisyllabic nonsense sequence and that they were required to write it down. This allowed us to choose the SNR that would reduce the intelligibility of the stimuli by approximately 50%, a reduction similar to the one used by Mattys (2004), with intelligibility operationalized as the total number of stimuli correctly identified. As expected for unfamiliarized listeners, the TP-words and the part-words of the AL did not yield significantly different correct responses [$t(22) = -1.534$, $p > .10$]. The best performance was observed for intact speech (an average of 91.7% correct identification). The 22-dB SNR reduced performance by nearly 50%, leading to 47.9% correct identification on average; the other SNR conditions induced much poorer performance (16.6% correct identification for 10-dB SNR, 5% for 5-dB SNR, 0% for 0-dB SNR).

In addition, the mean number of phonemes correctly identified (in their correct order) with 22-dB SNR was 5.37 in sequences of six phonemes each. This indicates that the incorrect responses in the 22-dB SNR condition were not due to an inability to identify any of the phonemes of the stimuli, as was the case in the 0-dB SNR condition, for example. Thus, although the 22-dB SNR impoverished the signal, it did not make the phonetic information inaudible. This makes a 22-dB SNR condition a perfect option, since the signal is degraded but not to the degree that many phonetic aspects of the speech are unavailable. Nevertheless, there are important differences between this pretest task, in which naive participants were required to identify the TP-words and part-words of the AL presented one at a time, and the AL learning situation, in which participants were repeatedly exposed to these stimuli embedded in longer streams of speech. Thus, the intelligibility level obtained with a specific SNR in the identification-in-noise task may not correspond to the degree of intelligibility obtained with the same SNR in an AL-learning task. One should note that for words, a 0-dB SNR usually reduces intelligibility by about 50%. Although our AL material is rather different from real words, it is plausible that repeated exposure to this material would lead to a higher level of intelligibility in the AL situation than in the one used by Mattys (2004). Thus, another condition with superimposition of a higher level of white noise (10-dB SNR) was also chosen. In the pretest, the 10-dB SNR had a strong impact on the identification of the AL stimuli, reducing dramatically the correct identification not only of the nonwords but also of the phonemes that constituted them (3.9 out of 6).

The forced-choice test included the six TP-words and six part-words. These part-words consisted of syllables of two different words that had appeared adjacently in the speech stream. Three part-words (1 syllable#2 syllables in Table 1) comprised the last syllable of one word and the first two syllables of the next word; for example, /bɐ/, which is the last syllable of /lufɐbɐ/, was combined with /kilɐ/, the first two syllables of the next word, /kilɐbu/, to form the part-word /bɐkilɐ/. The other three part-words (2 syllables#1 syllable in Table 1) comprised the last two syllables of one word and the first syllable of the next word. For all conditions and groups, the stimuli used in the test phase were produced by concatenating the three syllables that constituted each word or part-word without white noise superimposed, thus avoiding responses based on acoustic matching between the stimuli of the familiarization and test phases.

Procedure

The familiarization phase was presented with Windows Media Player, with the auditory stimuli presented at a comfortable level through Sennheiser HD 280 headphones. For the test phase, stimuli were also presented through headphones, with presentation, timing, and data collection controlled by E-Prime 1.1 (Schneider, Eschman, & Zuccolotto, 2002a, 2002b).

All participants were instructed to listen to a new language that contained words, but no meaning or grammar. Their task was to find out which words constituted the new language. No information about the structure, phonology, or length of the words was given. Participants were informed that the experiment consisted of three short listening blocks, followed by a test of their knowledge of the words that constituted the language. Participants in the degraded-signal conditions were warned of the poor signal quality. A 5-min break followed each of the 7-min blocks. After the listening phase, participants were presented with a two-alternative forced-choice test. Each trial started with a warning tone, followed by two trisyllabic strings that were separated by 500 msec of silence. One of these strings was a word from the AL, the other was a part-word. Each word was paired exhaustively with each part-word, rendering 36 trials. A new trial began immediately after participants gave their answer or if no answer was registered after 10 sec. Participants were told to always provide an answer even if they were not totally sure about their decision. Nevertheless, accuracy was also emphasized. The test began with four practice trials, in which an animal and an

Table 2
TP Word Responses (in Percentages) According to Cue Condition (Single, Congruent, or Incongruent) and Signal Quality (Intact Speech, 22-dB SNR, 10-dB SNR), Considering the TP Gradient (High or Low) of the TP Words of the Artificial Language

	Signal Quality					
	Intact		22-dB SNR		10-dB SNR	
AL Condition	High TP	Low TP	High TP	Low TP	High TP	Low TP
Single cue	66.1	66.6	62.2	62.2	55.5	60.5
Congruent cues	82.2	85.0	66.6	77.7	57.7	63.3
Incongruent cues	39.4	41.6	50.0	48.1	55.5	59.4

Note—Chance corresponds to 50%.

environmental sound were presented and participants had to decide which of the two was an animal sound. Feedback was provided only for practice trials. Order of presentation of test trials was randomized for each participant, and order of presentation of the stimuli within trials was counterbalanced within each group.

RESULTS AND DISCUSSION

The percentage of TP-word choices was computed for each participant. Table 2 presents the average AL learning performances broken down by TP level (high vs. low) of the TP-words.

In all input intelligibility conditions, participants presented with the single cue chose the TP-word significantly more often than the part-word: with intact signal, 67% [$SD = 6.4$, $t(8) = 7.917$, $p < .0001$]; with mildly degraded signal (22-dB SNR), 62% [$SD = 8.2$, $t(5) = 3.605$, $p < .025$]; with strongly degraded signal (10-dB SNR), 58% [$SD = 7.1$, $t(8) = 3.53$, $p < .01$]. A significant learning effect was also found in all intelligibility conditions for the participants presented with congruent cues: with intact signal, 84% [$SD = 9.4$, $t(10) = 11.985$, $p < .001$]; with 22-dB SNR, 72% [$SD = 12.2$, $t(9) = 5.69$, $p < .001$]; with 10-dB SNR, 61% [$SD = 4.7$, $t(8) = 6.897$, $p < .001$].

In sharp contrast with this pattern, with intact signal, participants presented with incongruent cues discarded the TP-words, choosing TP-words only 41% of the time on average ($SD = 10.5$). Thus, in this condition, participants judged the part-words to be the lexical units of the new language significantly more often than they did the TP-words [$t(11) = -3.040$, $p = .01$]. With the mildly degraded signal, performance with incongruent cues did not differ from chance [$t(11) = -.389$, $p = .704$], with TP-words chosen 50% of the time ($SD = 12.5$). Thus, no learning effect was observed. It was only in the strongly degraded condition that participants presented with incongruent cues significantly preferred the TP-words over the part-words, choosing the TP-words 57% of the time on average [$SD = 7.6$, $t(8) = 2.794$, $p < .025$].

In order to evaluate directly both the weighting of the two studied cues and the impact of signal quality on this weighting, we ran an ANOVA including cue condition (single cue vs. congruent cues vs. incongruent cues) and signal quality (intact speech vs. 22-dB SNR degraded signal vs. 10-dB SNR degraded signal) as between-subjects factors. We also included words with high-level TPs versus words with low-level TPs as a within-subjects factor, to determine whether there was a TP gradient.

The cue condition effect [$F(2,78) = 49.1$, $MS_e = 5.66$, $p < .001$] was significantly modulated by signal quality, as revealed by the interaction between these two factors [$F(4,78) = 11.551$, $p < .001$]. No other main effect or interaction was significant (all Fs < 2, ps $> .1$). We further investigated the nature of the significant interaction through pairwise comparisons, using the Bonferroni-corrected alpha rate of .017. As Figure 1 clearly illustrates, with intact signal, a main effect of cue condition was found [$F(2,29) = 64.77$, $MS_e = 10.87$, $p < .001$], with the congruent-cues condition leading to better performance than both the single-cue [$F(1,29) = 16.55$, $p < .0005$] and the incongruent-cues [$F(1,29) = 128.74$, $p < .0001$] conditions. Not surprisingly, performance in the incongruent-cues condition, which led participants to prefer part-words over TP-words, also differed from performance in the single-cue condition [$F(1,29) = 43.47$, $p < .001$].

Figure 1 also shows that with the mildly degraded (22-dB SNR) signal—main effect of cue condition [$F(2,25) = 11.24$, $MS_e = 17.33$, $p < .001$]—incongruent cues led to a lower performance than did congruent cues

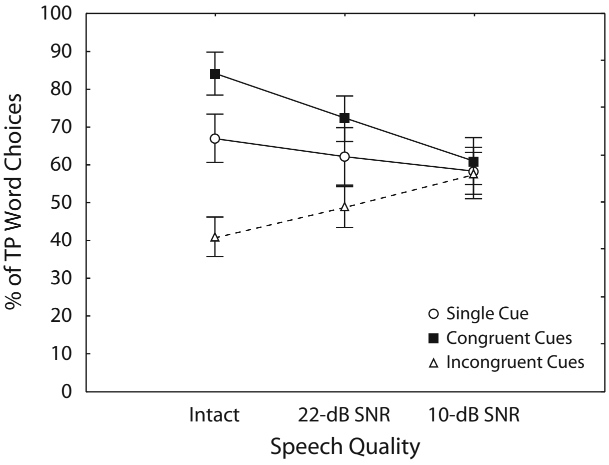

Figure 1. Performance patterns (TP-word responses, in percentages) according to cue condition (single, congruent, incongruent) and signal quality (intact speech, 22-dB SNR, 10-dB SNR). Vertical bars denote 0.95 confidence intervals. Chance corresponds to 50%.

[$F(1,25) = 20.29$, $p < .001$]. With a strongly degraded (10-dB SNR) signal, all three cue conditions resulted in a similar performance level [$F(2,24) = 0.73$, $MS_e = 2.815$, $p = .48$], contrary to what was observed in the former two input intelligibility conditions.

The data show that signal quality had no major impact on results in the single-cue condition [$F(2,21) = 3.33$, $MS_e = 6.58$, $p > .05$]. Thus, the statistical learning mechanism operating on TPs between adjacent syllables seems to be very resilient to noise, allowing speech segmentation to be almost as efficient when the signal is quite distorted (i.e., with a 10-dB SNR) as when it is highly intelligible [$F(1,78) = 3.84$, $p > .05$].

In contrast, signal quality significantly affected performance in the congruent-cues condition [$F(2,27) = 14.94$, $MS_e = 11.46$, $p < .001$]. A significant linear trend [$F(1,78) = 30.05$, $p < .001$] suggests that in good listening conditions, congruent cues to word boundaries were integrated, allowing the optimization of the speech segmentation process. However, as signal degradation increased, integration was affected and hence the redundancy gain[5] was largely reduced. Indeed, the number of TP-word choices was significantly higher with intact speech than with noise superimposition: 22-dB SNR and 10-dB SNR [$F(1,78) = 8.47$ and 30.05, respectively, $p < .005$]. In addition, performance was also better with mildly (22-dB SNR) than with strongly (10-dB SNR) degraded speech [$F(1,78) = 6.72$, $p = .011$].

This pattern of progressive reduction of the redundancy gain seems related to the strong sensitivity of coarticulatory cues to noise. Indeed, as reported above, the redundancy gain observed in the congruent-cues condition compared with that of the single-cue condition was significant only with intact speech. It was numerically still present but not statistically significant anymore with mildly degraded speech (22-dB SNR) and was no longer observed at all with strongly degraded speech (10-dB SNR).

The sensitivity to noise of coarticulatory cues is even more clearly revealed by the performance pattern observed with incongruent cues. With incongruent cues, the effect of noise was significant [$F(2,30) = 6.16$, $MS_e = 14.53$, $p < .01$], and modulation of performance by signal quality is reflected by a significant linear trend [$F(1,78) = 15.74$, $p < .001$]. As already reported, participants relied on coarticulation rather than on statistical information only when exposed to intact speech, a listening condition that differed significantly from strongly degraded (10-dB SNR) speech [$F(1,78) = 15.74$, $p = .0001$]. Indeed, with low noise intensity (22-dB SNR), the inconsistency between the two types of information disrupted performance, but the influence of coarticulation totally vanished only with the most degraded input (10-dB SNR). In this strongly degraded speech condition, listeners considered only statistical information, performing at the same level as listeners exposed to either a single cue or congruent cues.

GENERAL DISCUSSION

In an AL learning experiment, we used three signal quality conditions to investigate the relative impact of two types of sublexical information in speech segmentation: statistical information (TPs between adjacent syllables) and subsegmental information (coarticulation). The present results suggest, in accordance with Mattys and colleagues' (Mattys, 2004; Mattys et al., 2005) proposal, that the segmentation process adopted by listeners varies as a function of both signal quality and the types of cues available in the speech flow. With phonetically intact speech, coarticulation was a powerful segmentation cue, able to drive the segmentation process. Most importantly, coarticulation overruled statistical information when both cues were put in conflict in the speech stream, which is in line with Johnson and Jusczyk's (2001) infant data. Mattys and colleagues (Mattys, 2004; Mattys et al., 2005) had already demonstrated coarticulation reliability in adults when coarticulatory cues were in conflict with lexical stress. Our results add to this evidence. Thus, in good listening conditions, coarticulation is given priority over either lexical stress (Mattys, 2004; Mattys et al., 2005) or a general segmental statistical cue (in the present study, TPs). It is not clear, however, whether the local coherence given by coarticulation or the perception of edges between concatenated and coarticulated parts of the speech stream (or even both factors) forms the basis for a coarticulation-driven segmentation procedure.

In any case, the perceptual salience of coarticulatory information is extremely affected by signal degradation. Both the present study and Mattys and colleagues' (Mattys, 2004; Mattys et al., 2005) data have shown that coarticulation is much more affected by the presence of noise than are other cue types. Indeed, when only sublexical information is available in the speech stream, segmentation driven by coarticulation is observed only with phonetically intact speech. When information from each of the three tiers, including both lexical and sublexical information, is available, coarticulation has an effect only in mildly noisy conditions (Mattys et al., 2005). Remarkably, Mattys et al. (2005), in a condition of moderate noise superimposition, observed no preponderance of any of the incongruent sublexical cues, as was also the case in the mildly degraded condition of the present study. Nevertheless, it seems that in conditions of signal degradation in which some acoustic properties are still available, the cue that is less reliable in noisy conditions (i.e., coarticulation) is still sufficiently available to disrupt the segmentation driven by an inconsistent cue (e.g., lexical stress in Mattys et al., 2005, and TPs in the present study), although, as noise increases, this cue becomes unable to drive the segmentation process by itself and override statistical information. This pattern is reflected in the present 22-dB SNR condition by two pieces of evidence: (1) the nonsignificant trend toward a redundancy gain observed when coarticulation and statistical cues suggested the same word boundaries and (2) the significant interference effect (leading to no AL learning) when these two cues pointed toward conflicting segmentation points.

In short, the incongruent-cues condition showed that coarticulation overrides TPs only in intact speech, whereas TPs override coarticulation in strongly degraded speech. The pattern observed with congruent cues suggests an additive effect of converging cues (at least in good listening conditions), allowing the optimization of the speech segmentation process. However, whether this redundancy gain is synergistic (greater than the sum of the isolated cues effect) or conjunctive (Christiansen & Curtin, 2005) cannot be determined from the present results, since no coarticulation-only condition was assessed (the single-cue condition was a TPs-only condition). In any case, there was a performance gain when statistics and coarticulation were consistent in comparison with the condition in which only statistical information was available. However, integration was affected and hence the redundancy gain was largely reduced as signal degradation increased.

The pattern of results reported here was not modulated by the TP level of the TP-words. In fact, we did not find any hint of a TP gradient effect: TP-words with higher TPs, ranging from 0.75 to 1.00, were not better learned than those with TPs ranging from 0.50 to 0.58. Since such a gradient had been reported by Saffran, Newport, & Aslin (1996), at least in a condition in which only statistical information was available, we suspect that this null result is partly due to the fact that the TP range within the TP-words used in our study was narrower (from 0.50 to 1.00) than that of Saffran, Newport, & Aslin, which varied from 0.31 to 1.00.

Much more important, the pattern of the present results cannot be attributed to either the use of natural speech or to the frequency difference of TP-words and part-words. Indeed, it has been suggested that the use of natural rather than synthesized speech in AL-learning studies could make the tasks easier (Thiessen & Saffran, 2003). However, the similarity between the performance levels of the TP groups exposed to synthesized speech and those exposed to natural speech (see the description of the pretest in the Method section) clearly demonstrates that this dimension cannot account for the present results.

As regards the raw frequency differences between TP-words and part-words, we know that although this factor is potentially important to word discovery by infants (Brent & Cartwright, 1996) and can explain participants' performance in some studies (see, e.g., Dahan & Brent, 1999), it cannot always explain the impact of TPs (Aslin, Saffran, & Newport, 1998). In the present study, we tried to minimize the potential confound between raw frequency of occurrence of the stimuli and TPs. In AL conditions in which coarticulatory information was available, the speech stream was constituted by concatenated syllables, and the number of coarticulated tokens of TP-words in the congruent-cues condition was equivalent to the number of coarticulated tokens of part-words in the incongruent-cues condition. Thus, the raw frequency of coarticulated exemplars was matched in these two conditions. The learning pattern (or the significant absence of learning, in the incongruent-cues condition with intact signal) shows that listeners did not rely (at least not only) on the absolute frequency of the items. Had listeners used only the frequency of the trisyllabic items to analyze the speech stream, performance would have been similar in the three cues conditions—all groups would have learned the AL equally well, with performance depending exclusively on signal quality. Instead, the AL learning patterns differed according to both the segmentation cues available in the speech stream and noise contingency.

The present study also shows that both the influence of noise and the involvement of the various sources of information available vary in a graded manner. In addition, the weighting modulation by noise seems to be mainly related to the listeners' inability to exploit coarticulatory information when the signal is distorted. It is only in that case that the segmentation process is driven solely by statistical information. Indeed, in the 10-dB SNR condition, a similar AL learning performance was observed independently of the number and congruence of the segmentation cues available in the stream. This suggests that coarticulation was no longer available to assist segmentation. Thus, the weighting change does not seem to support the notion that the less one cue type is used, the more other cue types will gain importance. Instead, it largely depends on both the unavailability of coarticulatory cues and the high resilience of statistical information to noise superimposition.

In summary, noise superimposition does not affect all segmentation cues to a similar degree—not when these cues pertain to the distinct tiers defined by Mattys et al. (Mattys, 2004; Mattys et al., 2005) nor when they are sublexical, as was the case here. Does this imply that the segmental and subsegmental information should be assigned to qualitatively different tiers, contrary to what is posited by Mattys et al.'s (2005) hierarchical model? Providing a definite answer to this question is difficult given the available evidence. In conditions in which segmental (phonotactic; cf. McQueen, 1998) information and subsegmental (coarticulation; cf. Mattys, 2004) information competed individually with metrical prosody, both overruled the last cue. This may support the need for a broad distinction between sublexical (either segmental or subsegmental) cues and prosodic information. However, as Mattys et al. have already suggested and as the present results clearly demonstrate, although segmental information and subsegmental information in any natural language tend to be intrinsically correlated, they might be differentially weighted.

Part of this differential weighting, and hence part of the resilience of cues to noise, might be related to the structural grain of the cues, as proposed by Mattys et al. (2005) and already commented on. Indeed, the use of cues defined at a lower structural level depends on the ability to process fine-grained, low-level acoustic properties, which are more easily masked by noise than higher level units like syllables (involved here in the TPs). Note, however, that cue reliability cannot be reduced to perceptual salience per se. Indeed, if it were always easier to extract information from highly salient syllables than from perceptually less salient subsyllabic (segmental or subsegmental) structures, coarticulation would never overrule statistical information when both cues are put in conflict. Yet this was the case with intact speech, both in the present experiment and in Johnson and Jusczyk's (2001) study of infants. Nevertheless, to better understand the role of the structural grain of the cues, it would be interesting to contrast, under various noise conditions, the power of (subsegmental) coarticulation and that of distributional regularities defined at the segmental level, such as phonotactic cues, which have been shown to intervene in AL learning, at least when nonadjacent TPs are considered

(Onnis et al., 2005). It would also be interesting to compare, as we are presently doing, the resilience to cognitive noise of the two types of cues used here (as well as of other cues, such as prosodic ones) through the use of (potentially) interfering tasks of various levels of difficulty (cf. Toro, Sinnett, & Soto-Faraco, 2005). If TPs were found to be the most resilient segmentation procedure, not only to physical noise, as in the present study, but also to cognitive noise, any interpretation based only on physical masking of acoustic properties would be dismissed.

In fact, the differential weighting of cues may depend not only on the structural grain of the cues but also on two basic factors, the first one being domain generality. The ability to track TPs involves a domain-general learning mechanism, as demonstrated by the fact that TPs are also extracted in tone (Saffran et al., 1999) and visual (Fiser & Aslin, 2001) sequences. On the contrary, coarticulatory information is speech specific. Within the speech domain, it might be useful to consider further the distinction between universal cues, such as intonational phrases that partly correspond to physiological mechanisms like breath groups (Shukla et al., 2007), and language-specific cues, such as properties that are unique to a particular language. Indeed, the latter properties obviously need to be learned, whereas both domain-general statistical mechanisms and speech-specific universal cues are available from a very early phase in language acquisition and are used by adults even with unknown or foreign languages (Shukla et al., 2007).

Both domain-general statistical mechanisms and speech-specific universal cues might thus play a central role in development as guides to other segmentation cues. To be reliable guides, such cues should be relatively immune to signal degradation. In other words, as suggested by Mattys et al. (2005), the more resilient (but, in normal listening conditions, lower weighted) cues in adult speech segmentation could correspond to the earliest acquired and thus the most critical cues at the onset of language development. Further work should be aimed at testing these propositions. For example, since various languages have different patterns of coarticulation that often reflect the phonetic contrasts that are emphasized (see Manuel, 1999, for review), the importance of the language specificity of cues versus their universality may be assessed by contrasting the role of TPs with the role of language-specific as opposed to language-general patterns of coarticulation in speech segmentation. Contrasting the universal prosodic cues used by Shukla et al. (2007) with language-specific prosodic cues like the word stress patterns used by Mattys (2004) and Mattys et al. (2005) may also shed light on the relevance of this distinction. It would also be worth contrasting domain-general statistical mechanisms with speech-specific universal cues like the universal prosodic properties examined by Shukla et al. (2007) under various physical and cognitive noise conditions. Indeed, with intact speech, Shukla et al. observed that phrasal prosodic cues act as a filter, suppressing possible word-like sequences (trisyllabic sequences with high TPs) that straddle two prosodic constituents. Whether this would hold true in noisy situations remains to be tested.

In summary, the AL-learning patterns observed in this study have yielded three important findings: (1) the modulation of coarticulation reliability by signal quality; (2) the high resilience of statistical information based on TPs to noise superimposition; (3) the strong signal contingency of the weighting of the cues used to segment speech into words. This pattern of results can be well accommodated by Mattys and colleagues' (Mattys, 2004; Mattys et al., 2005) hierarchical proposal, and it highlights the importance of studying speech segmentation in the context of multiple cues (see, e.g., Christiansen & Curtin, 2005) and in different listening conditions (Mattys, 2004; Mattys et al., 2005). Importantly, an integrated approach must seek to comprehend the role of sublexical information in speech segmentation, as well as how sublexical cues interact with lexical and supralexical information and how the weighting of several cue types is affected by listening conditions.

AUTHOR NOTE

Preparation of this article was supported by Grant SFRH/BD/12290/2003 from the Fundação para a Ciência e a Tecnologia, Ministério da Ciência, Tecnologia e Ensino Superior to support the doctoral work of T.F. This research was also supported by a joint grant from Fundação para a Ciência e a Tecnologia and European Community FEDER funding (Project POC/PSI/56901/2004, "Visual phonology and auditory orthography"). The Centro de Psicologia Clínica e Experimental: Desenvolvimento, Cognição e Personalidade of the University of Lisbon also provided support for the preparation of this article. R.K. is Senior Research Associate of the Belgian Fonds National de la Recherche Scientifique. We are grateful to Mark Pitt and three anonymous reviewers for their helpful comments on an earlier version of the manuscript. Correspondence concerning this article should be addressed to P. Ventura, Faculdade de Psicologia e de Ciências da Educação da Universidade de Lisboa, Alameda da Universidade, 1649-013 Lisbon, Portugal (e-mail: paulo.ventura@fpce.ul.pt or paulo.ventura@gmail.com).

REFERENCES

Aslin, R. N., Saffran, J. R., & Newport, E. L. (1998). Computation of conditional probability statistics by 8-month-old infants. *Psychological Science*, **9**, 321-324.

Brent, M. R., & Cartwright, T. A. (1996). Distributional regularity and phonotactic constraints are useful for segmentation. *Cognition*, **61**, 93-125.

Byrd, D. (1996). Influences on articulatory timing in consonant sequences. *Journal of Phonetics*, **24**, 209-244.

Byrd, D., & Saltzman, E. (1998). Intragestural dynamics of multiple prosodic boundaries. *Journal of Phonetics*, **26**, 173-199.

Cho, T., & McQueen, J. M. (2005). Prosodic influences on consonant production in Dutch: Effects of prosodic boundaries, phrasal accent and lexical stress. *Journal of Phonetics*, **33**, 121-157.

Christiansen, M. H., & Curtin, S. (2005). Integrating multiple cues in language acquisition: A computational study of early infant speech segmentation. In G. Houghton (Ed.), *Connectionist models in cognitive psychology* (pp. 347-372). Hove, U.K.: Psychology Press.

Conway, C. M., & Christiansen, M. H. (2005). Modality-constrained statistical learning of tactile, visual, and auditory sequences. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **31**, 24-39.

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.

Dutoit, T., Pagel, V., Pierret, N., Bataille, F., & van der Vrecken, O. (1996). The MBROLA project: Towards a set of high quality speech synthesizers free of use for non commercial purposes. *Proceedings of the International Conference on Spoken Language Processing*, **3**, 1393-1396.

Fiser, J., & Aslin, R. N. (2001). Unsupervised statistical learning of higher-order spatial structures from visual scenes. *Psychological Science*, **12**, 499-504.

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

Gómez, R. L., & Gerken, L. (2000). Infant artificial language learning and language acquisition. *Trends in Cognitive Sciences*, **4**, 178-186.

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

Keating, P. A. (2006). Phonetic encoding of prosodic structure. In J. Harrington & M. Tabain (Eds.), *Speech production: Models, phonetic processes, and techniques* (pp. 167-185). New York: Psychology 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.

Kuhn, G., & Dienes, Z. (2005). Implicit learning of nonlocal musical rules: Implicitly learning more than chunks. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **31**, 1417-1432.

Kühnert, B., & Nolan, F. (1999). The origin of coarticulation. In W. J. Hardcastle & N. Hewlett (Eds.), *Coarticulation: Theory, data, and techniques* (pp. 61-75). Cambridge: Cambridge University Press.

Liberman, A. M., & Studdert-Kennedy, M. (1978). Phonetic perception. In R. Held, H. Leibowitz, & H.-L. Teuber (Eds.), *Handbook of sensory physiology: Vol 8. Perception* (pp. 143-178). Berlin: Springer.

Manuel, S. (1999). Cross-language studies: Relating language-particular coarticulation patterns to other language-particular facts. In W. J. Hardcastle & N. Hewlett (Eds.), *Coarticulation: Theoretical and empirical perspectives* (pp. 179-198). Cambridge: Cambridge University Press.

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

Mattys, S. L., White, L., & Melhorn, J. F. (2005). Integration of multiple speech segmentation cues: A hierarchical framework. *Journal of Experimental Psychology: General*, **134**, 477-500.

McNealy, K., Mazziotta, J. C., & Dapretto, M. (2006). Cracking the language code: Neural mechanisms underlying speech parsing. *Journal of Neuroscience*, **26**, 7629-7639.

McQueen, J. M. (1998). Segmentation of continuous speech using phonotactics. *Journal of Memory & 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, & Cognition*, **20**, 621-638.

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

Onnis, L., Monaghan, P., Richmond, K., & Chater, N. (2005). Phonology impacts segmentation in online speech processing. *Journal of Memory & Language*, **53**, 225-237.

Perruchet, P., & Pacton, S. (2006). Implicit learning and statistical learning: One phenomenon, two approaches. *Trends in Cognitive Sciences*, **10**, 223-238.

Perruchet, P., & Vinter, A. (1998). PARSER: A model for word segmentation. *Journal of Memory & Language*, **39**, 246-263.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, **274**, 1926-1928.

Saffran, J. R., Johnson, E. K., Aslin, R. N., & Newport, E. L. (1999). Statistical learning of tone sequences by human infants and adults. *Cognition*, **70**, 27-52.

Saffran, J. R., Newport, E. L., & Aslin, R. N. (1996). Word segmentation: The role of distributional cues. *Journal of Memory & Language*, **35**, 606-621.

Sanders, L. D., Newport, E. L., & Neville, H. J. (2002). Segmenting nonsense: An event-related potential index of perceived onsets in continuous speech. *Nature Neuroscience*, **5**, 700-703.

Schneider, W., Eschman, A., & Zuccolotto, A. (2002a). *E-Prime references guide*. Pittsburgh: Psychology Software Tools.

Schneider, W., Eschman, A., & Zuccolotto, A. (2002b). *E-Prime user's guide*. Pittsburgh: Psychology Software Tools.

Shukla, M., Nespor, M., & Mehler, J. (2007). An interaction between prosody and statistics in the segmentation of fluent speech. *Cognitive Psychology*, **54**, 1-32.

Thiessen, E. D., & Saffran, J. R. (2003). When cues collide: Use of stress and statistical cues to word boundaries by 7- to 9-month-old infants. *Developmental Psychology*, **39**, 706-716.

Toro, J. M., Sinnett, S., & Soto-Faraco, S. (2005). Speech segmentation by statistical learning depends on attention. *Cognition*, **97**, B25-B34.

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

NOTES

1. One may argue that the term *hierarchical organization* is misleading, since there is no structural relationship between the sources of information of the different tiers: They are simply weighted differently, with a bottom-up increase in reliability of the segmentation cues in optimal listening conditions. For clarity, however, we have retained this terminology when referring to Mattys et al.'s model (Mattys, 2004; Mattys et al., 2005).

2. Whereas Saffran and colleagues (Saffran, Aslin, & Newport, 1996; Saffran, Newport, & Aslin, 1996) considered statistical learning to be based on statistical computations, another mechanism involving the formation of chunks has been proposed (Perruchet & Vinter, 1998). In terms of their explanatory powers, it is difficult to decide between these two interpretations (Perruchet & Pacton, 2006); however, learning effects based on nonadjacent TPs (see, e.g., Kuhn & Dienes, 2005; Onnis et al., 2005) seem to challenge the chunking explanation.

3. The SNR ratio is measured as the noise intensity in relation to the average intensity of the speech signal (here, the AL signal). Thus, a 10-dB SNR means that if the signal intensity is 76 dB, as was the case here, then the noise intensity was set at 66 dB.

4. The 22-dB SNR was chosen because this ratio reduced intelligibility by approximately 50%, a value similar to the one used by Mattys (2004); for details, see the discussion of the pretest in the Method section.

5. The term *redundancy gain* is used to emphasize the improvement in the segmentation process (revealed by the superior AL-learning performance in the congruent cues condition with intact speech) as the result of the availability of different consistent segmentation cues in the stream.

(Manuscript received May 31, 2006;
revision accepted for publication November 29, 2006.)