Available online at www.sciencedirect.com
ScienceDirect
Cognitive Psychology 54 (2007) 1–32
Cognitive Psychology
www.elsevier.com/locate/cogpsych

An interaction between prosody and statistics in the segmentation of fluent speech ☆

Mohinish Shukla [a,*], Marina Nespor [b], Jacques Mehler [a,c]

[a] *International School for Advanced Studies, Trieste, Italy*
[b] *University of Ferrara, Ferrara, Italy*
[c] *Laboratoire de Sciences Cognitives et Psycholinguistique, EHESS-ENS-CNRS, Paris, France*

Accepted 9 April 2006
Available online 19 June 2006

Abstract

Sensitivity to prosodic cues might be used to constrain lexical search. Indeed, the prosodic organization of speech is such that words are invariably aligned with phrasal prosodic edges, providing a cue to segmentation. In this paper we devise an experimental paradigm that allows us to investigate the interaction between statistical and prosodic cues to extract words from a speech stream. We provide evidence that statistics over the syllables are computed independently of prosody. However, we also show that trisyllabic sequences with high transition probabilities that straddle two prosodic constituents appear not to be recognized. Taken together, our findings suggest that prosody acts as a filter, suppressing possible word-like sequences that span prosodic constituents.

Keywords: Prosody; Transitional probability; Speech segmentation; Artificial speech; Language acquisition

☆ Various parts of the study were supported by the HFSP Grant RGP 68/2002, the Regione Friuli-Venezia-Giulia (L.R. 3/98), the McDonnell Foundation, CEE contract 12778 (NEST) specific targeted project CALACEI and the Progetto FIRB. The research is also supported in the framework of the European Science Foundation EUROCORES program "The Origin of Man, Language and Languages" as well as C.O.F.I.N. 2003–2005. We thank the editor and anonymous reviewers for useful suggestions.
* Corresponding author. Fax: +39 040 3787 615.
E-mail address: shukla@sissa.it (M. Shukla).

0010-0285/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.cogpsych.2006.04.002

1. Introduction

How do infants learn to segment fluent speech into a series of words? Statistical strategies were proposed for the segmentation of words, based on distributional properties over sub-lexical units like phonemes or syllables (e.g., Brent & Cartwright, 1996; Batchelder, 2002; Dahan & Brent, 1999; Harris, 1955; Swingley, 2005). Saffran and colleagues (e.g., Aslin, Saffran, & Newport, 1998; Saffran, Newport, & Aslin, 1996; Saffran, Aslin, & Newport, 1996) showed that troughs (''dips'') in transition probabilities (TPs)[1] between syllables are used by infants and adults to segment synthetic streams of continuous speech that lack prosodic cues (see also Peña, Bonatti, Nespor, & Mehler, 2002; Saffran, 2001; Thiessen & Saffran, 2003).

The prosody of language might also be a cue to word (or phrase) boundaries. Speech is typically organized into prosodically cohesive units ranging from the syllable to the utterance (e.g., Hayes, 1989; Nespor & Vogel, 1986; Selkirk, 1984). The boundaries of prosodic units are associated with acoustic cues like final lengthening and pitch decline (Beckman & Pierrehumbert, 1986; Cooper & Paccia-Cooper, 1980; Klatt, 1976; Wightman, Shattuck-Hufnagel, Ostendorf, & Price, 1992 amongst others). Thus, such acoustic/prosodic cues can mark domains in speech, even though there may not be a one-to-one relationship between the prosodic constituents and underlying syntactic constituents (see, for example, Fisher & Tokura, 1996; Gerken, Jusczyk, & Mandel, 1994; Shattuck-Hufnagel & Turk, 1996). Nevertheless, phonologists have proposed that phrasal prosodic constituents can be exhaustively parsed into a sequence of non-overlapping words (e.g., Nespor & Vogel, 1986; Selkirk, 1984, 1996; Shattuck-Hufnagel & Turk, 1996). Thus, boundaries of phrasal prosodic constituents are also word boundaries, and they may serve as segmentation cues to discovering words, perhaps even in very early stages of language acquisition.

In this paper, we explore experimentally whether prosodic cues are used to segment a speech stream into smaller-sized groups of syllables. We examine the interaction between the aforementioned segmentation cues: phrasal prosodic cues and TPs between syllables. That is, since troughs in TP are used to segment speech, and in addition words are aligned with phrasal prosodic edges, we investigate empirically how these two cues to word boundaries might interact. The larger goal of this research is to understand how multiple cues are utilized for acquiring language by infants and by adults learning a second language.

Do infants perceive and utilize the prosodic aspects of speech? Several studies indicate that newborns use prosody to discriminate two languages as long as these differ in their rhythm (see Christophe & Morton, 1998; Mehler et al., 1988; Mehler, Dupoux, Nazzi, & Dehaene-Lambertz, 1996; Nazzi, Bertoncini, & Mehler, 1998; Ramus, Hauser, Miller, Morris, & Mehler, 2000 amongst others). Several authors have found that prosody influences the way infants organize speech stimuli. Hirsh-Pasek et al. (1987) showed that 4.5-month-old (and 9-month-old) infants prefer utterances with artificially inserted pauses at clause boundaries as opposed to utterances with pauses inserted in the middle of clauses (see also Jusczyk, Pisoni, & Mullennix, 1992; Kemler, Nelson, Hirsh-Pasek, Jusczyk, & Wright-Cassidy, 1989; Morgan, 1994). Such a preference was present even for low-pass filtered speech (Jusczyk, 1989; Jusczyk et al., 1992). Other studies have shown that infants are also sensitive to even smaller prosodic units. Soderstrom et al. (2003) showed that

[1] The TP from any syllable x to another syllable y is given by: $TP(x \rightarrow y) = \frac{\text{frequency}(xy)}{\text{frequency}(x)}$.

6- and 9-month-old infants are sensitive to the phonological phrase (which typically consists of a content word and its associated function words, see Nespor & Vogel, 1986, for a technical definition). Gout, Christophe, and Morgan (2004) showed that slightly older (10- and 12.5-month) infants use phonological phrase boundaries to constrain on-line lexical access, suggesting that infants do not attempt lexical access on syllable sequences that span such boundaries (Christophe, Gout, Peperkamp, & Morgan, 2003).

Infants rapidly discover specific prosodic properties of the maternal language. Several authors have shown that infants are able to perceive lexical stress very early on (e.g., Fowler, Smith, & Tassinary, 1986; Sansavini, Bertoncini, & Giovanelli, 1997), and around nine months, they show a preference for the predominant stress pattern of their maternal language (see Houston, Jusczyk, Kuijpers, Coolen, & Cutler, 2000; Jusczyk, Cutler, & Redanz, 1993; Jusczyk, Houston, & Newsome, 1999 amongst others). Christophe and colleagues (Christophe, Dupoux, Bertoncini, & Mehler, 1994; Christophe, Mehler, & Sebastián-Gallés, 2001) demonstrated that newborns discriminate lists of bisyllabic items, one list consisting of items spliced from inside words (e.g., French 'mati' from 'mathématicien' or Spanish 'lati' from 'relativamente') while the other containing items spliced from the last syllable of one word and the first syllable of the next (e.g., French 'mati' from 'panorama typique' or Spanish 'lati' from 'manuela timida'). Morgan and Saffran (1995) showed that both 6- and 9-month-old infants represented sequences of syllables with an appropriate rhythmic pattern (for their language of exposure) as coherent units (see also Johnson & Jusczyk, 2001). In addition, Mattys, Jusczyk, Luce, and Morgan (1999) demonstrated that English 9-month-old infants preferred prosodically appropriate sequences over phonotactically appropriate ones, suggesting that prosodic cues might take precedence over phonotactic cues in the development of word segmentation strategies.

Infants can thus use prosody to organize speech into units including a few words (Christophe & Dupoux, 1996; Christophe, Nespor, Guasti, & van Ooyen, 1997; Guasti, 2002). This organization reduces the problem of segmenting words, since, as mentioned above, boundaries of prosodic constituents are also word boundaries. Words internal to prosodic constituents can then be extracted using various distributional strategies. For example, several authors have suggested that the distributional properties of segments at the edges of utterances can aid in word segmentation (e.g., Brent & Cartwright, 1996). The prosodic organization of speech would, at the very least, provide additional 'edges' in otherwise continuous speech, enhancing such distributional strategies.

The view that prosody can be used for segmentation is strengthened by the finding that in adults, prosody also influences the segmentation of artificial speech. Saffran et al. (1996), testing American adults, found that segmentation was facilitated when the final syllable of trisyllabic nonce words was lengthened, as compared to when the first syllable was lengthened. Bagou, Fougeron, and Frauenfelder (2002) found a similar facilitation with Swiss French adults when either the last syllable was lengthened or when it had a higher pitch. Christophe, Peperkamp, Pallier, Block, and Mehler (2004) found that prosody also constrained lexical access, in a word recognition paradigm. In this study, French adults had to respond to the presence of a target word (for example, *chat* /ʃa/[2] "cat") that could occur in a locally ambiguous context (e.g., *chat grincheux* /ʃagʀɛ̃ʃø/ where *chagrin*/ʃagʀɛ̃/ is a French word), or in a locally unambiguous context (like *chat drogué*/ʃadʀoge/; there is

[2] Pronunciations are marked in IPA throughout.

no French word starting with /ʃad/). The authors found that the word *chat* was responded to faster in the unambiguous than in the ambiguous context. However, this delay in detecting *chat* in an ambiguous context disappeared when a phonological phrase boundary occurred immediately after the target word (for example, *[le gros chat] [grimpait ...]*, /ləgʀoʃa#gʀɛ̃pɛ/, wherein the possible word *cha#grin* is now interrupted by a phonological phrase boundary as indicated). In other words, phonological phrase boundaries appear to act as natural boundaries, after which the cohort of activated candidates stops (see Marslen-Wilson & Tyler, 1980).

In this study, we use prosodic cues associated with Intonational Phrases (IPs). IPs account for natural break points in speech, being one of the highest levels of the prosodic hierarchy (Kager, 1999; Nespor & Vogel, 1986; Selkirk, 1984). An IP contains at least one syllable that bears (intonational phrasal) stress, and it ends with a boundary tone that is a cue associated with its right edge (Pierrehumbert & Hirschberg, 1990). IPs can be preceded and followed by pauses, and in fluent speech, pauses can be inserted at IP boundaries without perturbing the pitch contour. In 1, square brackets mark IPs (from Nespor & Vogel, 1986):

(1) [Lions,]$_{IP}$ [as you know,]$_{IP}$ [are dangerous.]$_{IP}$

Some of the prosodic correlates of IPs might be universal, being based ultimately on physiological mechanisms like breath groups (Bolinger, 1964; Lieberman, 1967; Ohala, Dunn, & Sprouse, 2004; Vaissière, 1995, but see Ladd, 1996). Several behavioral results have shown an effect of IP boundaries in processing fluent speech (for example, Watson & Gibson, 2004; see Cutler, Dahan, van, & Donselaar, 1997 for a review of prosodic effects in speech comprehension). Finally, ERP studies have shown that IPs elicit a characteristic component (Closure Positive Shift), even when speech is low-pass filtered (e.g., Friederici, Steinhauer, & Pfeifer, 2002; Steinhauer, Alter, & Friederici, 1999; Steinhauer & Friederici, 2001; see Steinhauer, 2003 for an overview).

Thus, IPs contain cues that are available to very young infants, and could be used to begin segmenting speech. Such cues thus might be important in very early stages of language acquisition, as opposed to other phonological cues like lexical stress or allophonic regularities, which are language specific, and might operate in a task-specific manner.

In this paper, we examine how Italian adult participants react to different segmentation cues in artificial speech streams. In Experiment 1, we investigate whether pseudowords internal to (Italian) IPs are better recalled than those that straddle IPs. In Experiment 2 we strip the speech stream used in Experiment 1 of prosody, to explore the role of statistical computations. In Experiment 3 we explore if different positions inside prosodic contours are equally well recognized. Based both on the results from these experiments and on theoretical considerations, we consider possible processing models for an interaction between prosodic and statistical cues. In Experiment 4, we furnish empirical evidence favorable to one of these models. Finally, in Experiments 5 and 6, we present some evidence to suggest that the effects of prosody might be in part due to universal properties of IPs, by replicating the effects of Italian IPs with Japanese IPs in Italian adult participants.

2. Experiment 1

Experiment 1 examines the effect of IP prosody on the extraction of statistically well-defined 'words' in a fluent speech stream. The role of TPs has been explored using artificial

speech streams composed by concatenating a few multisyllabic pseudowords. For example, in Saffran et al. (1996), the familiarization stream consisted of four trisyllabic 'words' concatenated at random (with no immediate repetitions of the 'words'). To place artificial 'words' in different positions in relation to prosodic domains without changing their statistical properties, we modified the nature of the familiarization stream. In our design, instead of concatenating trisyllabic pseudowords at random, we placed these inside a carrier sequence of 'noise' syllables.

2.1. Methods

2.1.1. Participants

Twenty adults (university students and researchers) participated in this experiment (9 males and 11 females, mean age 23.9 years, range 20–36 years). In this and subsequent experiments, participants were paid 3 euro each, reported no auditory or language-related problems and were naïve with respect to the aims of the experiment. All participants were monolingual, native speakers of Italian.

2.1.2. Materials

The artificial speech stream was conceived as a series of *frames*. We defined a frame as a sequence of 10 CV (Consonant–Vowel) syllables (σ). A single frame of ten syllables can be represented as follows:

$$[\sigma 1–\sigma 2–\sigma 3–\sigma 4–\sigma 5–\sigma 6–\sigma 7–\sigma 8–\sigma 9–\sigma 10]$$

In each frame there was one trisyllabic 'word' (for example at positions 4–5–6 or 5–6–7), and straddling two consecutive frames there was a trisyllabic 'word' (at position 9–10–1′ or 10–1′–2′, where 1′ and 2′ represent syllable slots from the successive frame, see Fig. 1).

We defined four trisyllabic nonsense words: /pu-le-a/, /ni-da-fo/, /te-ki-me/, and /vo-ge-tʃu/. In this experiment, the first two words were placed at contour-internal positions 4–5–6 or 5–6–7, and the other two were placed at contour-straddling positions 9–10–1′ or 10–1′–2′. This ensures that no two artificial 'words' can be adjacent; there is at least one noise syllable intervening between any two consecutive 'words.' There were 100 tokens of each word in the familiarization stream. Each frame contained one contour-internal

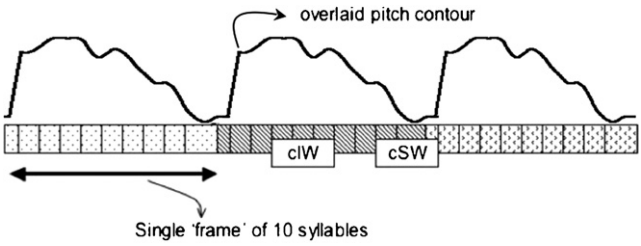

Fig. 1. Schematic outline of the structure of the familiarization stream for Experiment 1. A series of three frames, each containing 10 syllable slots is shown. Duration and pitch characteristics are the suprasegmentals that define the overlaid prosodic contour. Pitch is represented schematically by the average 'shape' of the eight contours. The pitch undergoes a 80-Hz excursion, from 130 to 210 Hz. Possible positions of one contour-internal 'word' (cIW) and one contour-straddling 'word (cSW) are shown.

'word,' and straddling each pair of successive frames there occurred one contour-straddling 'word.' There were thus 201 frames in all (corresponding to 200 straddling positions for the two straddling 'words'). The remaining syllabic slots in all frames were occupied by one of ten different 'noise' syllables. The noise syllables were /du/, /ko/, /mi/, /pa/, /po/, /ro/, /sa/, /ti/, /va/, and /ve/. These syllables were thus interspersed randomly between the 'words.' Care was taken to ensure that no bisyllabic sequence resembled an Italian word. Although the noise syllables were chosen at random from the limited pool, their average frequency over the entire stream was 100. An algorithm, implemented in MATLAB (Mathworks, Inc.) generated the sequence of frames. TPs between the syllables that formed the four 'words' were 1.0, while all other TPs were between 0.05 and 0.2, with a mean value of 0.1.

To add prosody to the frames, we proceeded as follows. A single Italian female speaker recorded nine short Italian declarative clauses, each one corresponding to a single IP. These were embedded in carrier sentences (listed in the Appendix A), and were between one and five words in length. The material was recorded with a Sony ECM microphone connected to a SoundBlaster sound card on a PC under Windows 2000™. CoolEdit (Syntrillium Corp.) was used to record and digitally manipulate the speech waveforms. The speech segments corresponding to the IPs were digitally excised. For each IP, we measured the pitch contour, smoothly interpolating across unvoiced segments using PRAAT (Boersma, 2001). A single pitch contour was converted into a vector of 400 pitch points (which thus results in time-normalized IPs). Thus, 20 pitch points per phoneme could be used to shape each of the 20 phonemes constituting the 10 CV syllables in each frame. From the nine recorded IPs, we thus obtained nine different pitch contour vectors.

We next measured the durations of the first and last syllables of each IP. These durations were divided by the number of segments in the syllables, to get a normalized value. We found that the average normalized duration of the phonemes of the last syllable (99.6 ms) was significantly greater from the average normalized duration of the phonemes of the first syllable (79.9 ms), paired t-test, $t(8) = 2.8$, $p = .02$. Since we routinely use phoneme durations of between 116 and 120 ms (e.g., Peña et al., 2002), we chose a mean value of 120 ms for all the phonemes except for those of the first and the last syllables. Phonemes of the first were shortened by 20 ms each, for a final value of 100 ms, while those of the last were lengthened to 140 ms each. Thus, on average, all the phonemes in a frame had a duration of 120 ms.

We used eight of the nine pitch contours. Each of the 201, 10-syllabic frames were associated with one of the eight contours. The ten syllables in each frame went from an initial syllable of 200 ms followed by 8 syllables of 240 ms to a final syllable of 280 ms. The total duration of the familiarization stream was 8 min, 3 s. Fig. 1 shows a schematic outline of the model of prosody we implement. Pitch is depicted schematically by the average of the eight Italian IP pitch contours we used.

The sequence of phonemes with the added prosodic characteristics was used to generate an artificial speech stream using MBROLA (Dutoit, 1997) and the Spanish male diphone database (es1). We chose the Spanish diphone database since pilot studies showed that artificial speech synthesized using this database resulted in speech that was perceived by Italian adults better than with other similar databases, including the Italian diphone database. Notice that the diphone database does not encode sentential prosody. The 22.05 kHz, 16-bit, mono wave file was converted to stereo and the initial and final 5 s of the file were ramped up and down in amplitude to remove onset and offset cues.

Trisyllabic sequences corresponding to the four words and four non-words were separately created for the test phase, using MBROLA and the es1 diphone database. The non-words were trisyllabic sequences constructed by concatenating the last two syllables of one 'word' and the first syllable of another 'word.' Non-words had not occurred during familiarization.[3] All phonemes of the test items were 120 ms in duration, and had a constant pitch of 100 Hz. All trisyllabic items were separately generated as 22.05 kHz, 16-bit, mono wave files. These were converted to stereo files for use in the test phase. We used such acoustically 'flat' test items, to approximate a neutral prosody that minimizes biasing the choice of the participants. Thus, the 'words' heard during test are acoustically different from those heard during familiarization.

The 'words' and non-words were pre-tested on 14 naïve participants for possible initial biases in the material. These participants heard a fully randomized sequence of all the chosen syllables for 2 minutes, followed by a test phase identical to that of Experiment 1 (see below). Overall, participants did not have a preference for 'words' over non-words, the mean score was 49.1%, $t(13) = -0.22$, $p = .83$. Looked at separately, neither the 'words' that were to be placed in contour-internal positions, nor those to be placed in contour-straddling positions were preferred over non-words, internal 'words,' 50.9%, $t(13) = 0.15$, $p = .89$, straddling 'words,' 47.3%, $t(13) = -0.51$, $p = .62$. Further, there were no differences between the 'words' to be placed in contour-internal and contour-straddling positions, $t(26) = 0.44$, $p = .66$. Thus, the material presents no initial biases overall for 'words' over non-words, and neither are the two groups of internal and straddling 'words' different from each other.

2.1.3. Apparatus

The experiments were conducted in a sound attenuated room. The experimental design was prepared and delivered using E-Prime V1.1 (Psychological Software Tools, 2002) under the Windows 98™ operating system. Sound was delivered through Sennheiser™ headphones attached to Harman-Kardon™ HK 19.5 speakers that themselves received input from SoundBlaster™ audio cards on the PCs. In the test phase, participants responded by pressing pre-marked keys on the E-Prime button box or keyboard.

2.1.4. Procedure

Each participant was seated in front of a computer screen where instructions were displayed. In the first phase, participants were instructed to listen to a speech stream in an 'invented' language and to try to pick up 'words' from this language. At the end of the familiarization phase (which lasted about 8 min), participants were instructed to listen to 16 pairs of auditory test items. Each pair consisted of a 'word' (contour-internal or contour-straddling 'word') and a non-word. After listening to each pair, participants had to press the left key on the button box or keyboard if the first item of the pair was judged as belonging to the artificial language heard in the familiarization phase and the right key if the second item was so judged (a typical two-alternative forced choice task, 2AFC). A response was coded as being correct if the key-press selected a contour-internal 'word' or a contour-straddling 'word' rather than a non-word. The order of 'words' and

[3] The non-words in this experiment are similar to the part-words in previous studies (e.g., Saffran et al., 1996). However, since they never actually occurred during familiarization, we refer to them as non-words, in keeping with the previous literature.

non-word was counterbalanced across trials. That is, in each trial, a 'word' was paired with a non-word. Each of the four 'words' in the familiarization phase was paired with each of the four non-words for a total of 16 trials. The 'words' occurred an equal number of times as the first or the second choice of the 2AFC. The two trisyllables in each trial were separated by a pause of 500 ms.

2.2. Results

The overall score, indicating correct segmentation of the speech stream, was 56.56% (*SD* 13.37), and was significantly different from chance, $t(19) = 2.2$, $p = .04$, $d = 0.49$. In this paper, all *t*-tests are two-tailed. Effect sizes are reported as Cohen's *d* for all *t*-tests (using pooled *SD* estimates for comparison of two groups), and η^2 for ANOVAs. Note that in this and the following experiments, chance is 50%, which implies no preference for 'words' over non-words in the 2AFC.

From Fig. 2, it can be seen that there appears to be a difference in the segmentation of contour-internal and contour-straddling 'words.'

The mean score of 68.13% (*SD* 22.39) for the contour-internal 'words' was significantly different from chance, $t(19) = 3.62$, $p < .002$, $d = 0.81$, while the mean score of 45% (*SD* 16.42) for the contour-straddling 'words' was not, $t(19) = -1.36$, $p = .19$, $d = 0.3$. The contour-internal 'words' differed significantly from the contour-straddling 'words,' $t(38) = 3.73$, $p < .001$, $d = 1.18$.

2.3. Discussion

The results indicate that, in the presence of prosody, only the contour-internal 'words' are recognized. That is, although all the 'words' are statistically indistinguishable, both in terms of their frequencies and their TPs, only those 'words' that lie inside IPs are correctly recognized. To establish that this is due to prosody, in the next experiment we stripped the

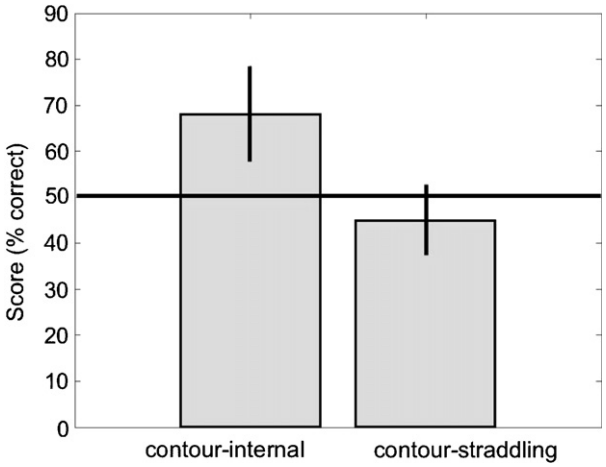

Fig. 2. The mean scores (% correct) from Experiment 1. Only contour-internal 'words' appear to be correctly segmented. Error bars represent 95% confidence limits of the means.

speech stream of its suprasegmental, prosodic overlay. This renders all the 'words' equivalent for both statistical and prosodic properties. It is thus to be expected that all the 'words' are correctly recognized. In addition, this further controls for a priori preferences of the two groups of 'words.' In fact, the participants in Experiment 2 are in exactly the same context as participants in Experiment 1, apart from prosody.

3. Experiment 2

In this experiment, we examine whether 'words' in a monotonous stream with interspersed syllabic noise can be segmented. The aim of this experiment is to remove prosody from the familiarization stream described in Experiment 1, and examine the effect of such a manipulation on the segmentation of statistically well-formed 'words' (that is, 'words' with high average TPs). We maintained the exact sequence of syllables as in the familiarization stream of Experiment 1, so as not to alter the statistics over the segments. The test phase was identical to the one in Experiment 1.

3.1. Methods

3.1.1. Participants

Twenty Italian adults (university students and researchers) participated in this experiment (5 males and 15 females, mean age 24.8 years, range 19–40 years).

3.1.2. Materials

Starting from the MBROLA file from Experiment 1, each phoneme was assigned a constant duration of 120 ms and a constant pitch of 100 Hz. This sequence of phonemes was used to generate artificial speech using the diphone-based speech synthesizer MBROLA, and the es1 (Spanish male) diphone database. The resultant was a 22.05 kHz, 16-bit, mono wave file with a duration of 8 min, 2 s. This file was converted into a stereo file, and the initial and final 5 s were ramped up and down in amplitude. This speech stream thus lacks all prosodic cues.

3.1.3. Apparatus and procedure

These were identical to Experiment 1.

3.2. Results

In this experiment we found that 'words' were significantly preferred over non-words, (mean 65.6%, *SD* 13.37), $t(19) = 5.23$, $p < .001$, $d = 1.17$.

As can be seen from Fig. 3, 'words' corresponding to contour-internal positions in Experiment 1 were preferred over the non-words (mean 68.75%, *SD* 15.97), $t(19) = 5.25$, $p < 0.001$, $d = 1.17$, as were the erstwhile contour-straddling 'words' (mean 62.5%, *SD* 20.28), $t(19) = 2.76$, $p < .015$, $d = 0.62$. In addition, the mean scores for the two 'word' types were not different from each other, $t(38) = 1.08$, $p = .29$, $d = 0.34$.

To compare the results from Experiment 2 with Experiment 1, we ran an ANOVA with Word Type (contour-internal or contour-straddling) as a within-subject factor and experimental Condition ('flat' familiarization or prosodic familiarization) as a between-subject factor. We found a significant main effect of Word Type, $F(1,38) = 11.95$, $p = .001$,

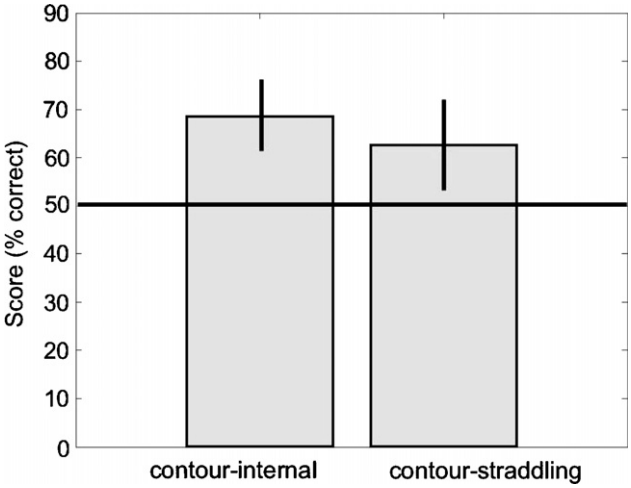

Fig. 3. The mean scores (% correct) from Experiment 2. All 'words' appear to be correctly segmented. Error bars represent 95% confidence limits of the means.

$\eta^2 = 0.12$, a significant main effect of Condition, $F(1,38) = 4.6$, $p = .04$, $\eta^2 = 0.047$, and a significant Word Type × Condition interaction, $F(1,38) = 3.95$, $p = .05$, $\eta^2 = 0.041$. These results indicate that the effect of prosody was primarily limited to the straddling 'words.' The internal 'words' in the two conditions were recognized at similar levels (68.75% in this experiment, versus 68.13% in Experiment 1), while the straddling 'words' were recognized above chance in this experiment (62.5%), but were at chance in Experiment 1 (45%).

3.3. Discussion

The results from Experiment 2 alone demonstrate that the presence of syllabic noise does not prevent the extraction of statistically defined, trisyllabic 'words.' Several studies have shown that both infants and adults can use troughs in TPs to extract 'words' from a fluent speech stream (amongst others, Saffran et al., 1996). In those experiments, the continuous streams were composed only of nonce words without any intervening noise syllables. Here, we extend those results, showing that words behave like figures that emerge from a background of 'noise syllables.'

One reason why segmentation is not affected by the presence of the interspersed noise syllables might be related to TPs at the edges of 'words.' In previous experiments (e.g., Peña et al., 2002; Saffran et al., 1996), where up to 6 'words' were concatenated at random, the TP from the last syllable of one 'word' to the first syllable of another was 0.2[4] (since immediate repetitions are not allowed, each 'word' can be followed by one of the other 5 'words'). In the present experiment, in contrast, the ten interspersed noise syllables occur at random. Thus, the TP from the last syllable of a 'word' to any of the noise syllables is 0.1. However, at the left (leading) edge of 'words,' any noise syllable can be followed by any of the other nine noise syllables, but also by the first syllable of the four 'words.' Thus,

[4] These are approximate, average values, since the speech streams are finite and randomly created.

the TP at the leading edge of 'words' is 0.077. Thus, possibly, the presence of interspersed syllabic noise actually enhances the detection of words because of a large ratio between the word-internal TPs and the TPs at the edges of words (see Gomez, 2002; Onnis, Monaghan, Christiansen, & Chater, 2004 for other studies demonstrating a beneficial effect of increased variability). Preliminary experiments in our lab show evidence for such a 'pop-out' effect in the presence of a large number of interspersed noise syllables. We suggest that this experimental paradigm is potentially of great use to study statistical models of segmentation of fluent speech.

Taken together, Experiments 1 and 2 demonstrate an effect of prosody on the segmentation of statistically defined 'words' in fluent speech. Experiment 2 established that in a monotonous speech stream, all statistically defined 'words' are correctly segmented, as expected. Experiment 1 demonstrated that when prosody is superimposed on the prosodically 'flat' speech stream, only those 'words' that lie internal to prosodic constituents are recognized. That is, when the two cues are in conflict, prosodic cues take precedence, resulting in prosodically ''bad'' syllabic sequences being rejected.

These results are compatible with proposals suggesting that prosodic constituents help segment speech (Christophe et al., 1997; Guasti, 2002). Our results thus suggest that prosody organizes the speech stream into constituents (hereafter referred to as 'phrases'), and this somehow results in contour-spanning 'words' being harder to detect.

How might we test the notion that prosody indeed carves up fluent speech into a series of 'phrases'? One possibility is suggested by studies on human memory, where it has been shown that learning an arbitrary sequence of verbal items is facilitated when the sequence can be perceptually chunked into shorter sequences (e.g., Burgess & Hitch, 1999; Hitch, Burgess, Towse, & Culpin, 1996). It has been known that edges of sequences are better recalled than their middles (e.g., Ebbinghaus, 1964; Miller, 1956), resulting in ∪-shaped recall curves (Baddeley, 1990; Brown, Preece, & Hulme, 2000). In perceptually chunked verbal lists, such ∪-shaped recall is observed *even for each of the subsequences* (Burgess & Hitch, 1999; Hitch et al., 1996; Henson, 1998; see Ng & Maybery, 2002 for a review).

Such evidence suggests that if prosody can divide a familiarization stream into 'phrases,' we might expect that the edges of such 'phrases' are more salient than their middles. In other words, finding an advantage for the recall of trisyllabic 'words' at edges over trisyllabic 'words' in the middles would constitute further empirical evidence in favor of a model wherein prosody serves to segment the input. Such a result is also warranted from other results from our lab which demonstrate an advantage for the processing of edges of multisyllabic sequences in adults. For example, Endress, Scholl, and Mehler (2005), showed that participants were better able to generalize repetition-based structures when repetitions were located at the edges rather than in the middles of seven-syllable sequences.

Thus, this experiment links language acquisition with general properties of the cognitive system, in this case, the salience of edges.

4. Experiment 3

This experiment is aimed at establishing if there is an advantage for 'words' at the edges of prosodic contours over 'words' in the middle. We modified the preparation of the speech stream in several ways. While in Experiment 1 'words' occurred either contour-internally or straddling two contours, in the current experiment all the 'words' occurred at contour-internal positions; either aligned with the edges or placed in the middles. We used

two different streams such as to place two 'words' at the right edge of IPs in one stream and at the left edge in the other. In both streams two other 'words' were placed in the middles.[5] One group of participants was exposed to a stream with 'words' at the left edges and in the middles of IPs. A separate group was exposed to a stream with 'words' at the right edges and in the middles of IPs.

We know from Experiment 1 that 'words' in the middle are correctly segmented. Also, the scores for the contour-internal 'words' were similar in the presence of prosody (68.13%) and in its absence (68.75%). This suggests that contour-internal 'words' are extracted using statistical mechanisms that are similar with or without prosody. We surmised that if the task of segmentation using statistical cues was made more difficult, it would worsen performance. This decline in performance should be greater for the middle-'words,' if indeed there is an added advantage for the edge-'words.' Thus, we decided to reduce the amount of statistical information by halving the amount of familiarization, providing 50 tokens of each 'word' instead of 100.

4.1. Methods

4.1.1. Participants

Fourteen Italian adults (university students and researchers) were exposed to the stream with edge-'words' at the left edge (2 males and 12 females, mean age 23.9 years, range 18–30 years). A separate group of twelve Italian adults were exposed to the stream with edge-'words' at the right edge (6 males and 6 females, mean age 23.3 years, range 19–32 years).

4.1.2. Materials

We created two new sequences of frames, resulting in two new speech streams. For the two streams, in each frame, we placed the contour-internal 'words' from Experiment 1 at position 1–2–3 (left-edge stream) or position 8–9–10 (right-edge stream), and designated them edge-'words'. The two contour-straddling 'words' from Experiment 1 were placed at position 6–7–8 (left-edge stream) or position 4–5–6 (right-edge stream) inside each frame and were designated the middle-'words'. The edge-'words' and the middle-'words' occurred 50 times each during the entire stream. The remaining slots in all frames were filled at random with interspersed noise syllables such that these had an average frequency of 50 across the entire stream. Each frame was randomly assigned one of eight prosodic contours from Experiment 1.

The entire sequences of phonemes were fed to MBROLA, using the Spanish male diphone database (es1). The final output files were 22.05 kHz, 16-bit, mono wave files, 4 min in duration. These files were converted into stereo files; the initial and final 5 s were ramped up and down to eliminate onset or offset cues to edge-'words' and middle-'words'. The test phase was identical for both group of participants and was identical to Experiments 1 and 2. Notice that in this experiment too, the non-words have zero frequency during familiarization.

[5] Having 'words' at both edges in the same stream would have necessitated an unnecessarily longer familiarization, since having both edges occupied simultaneously would have meant that two edge-'words' were often adjacent.

4.1.3. Apparatus and procedure

These were identical to Experiment 1.

4.2. Results

The overall score for the left-edge stream (mean 67.86%, *SD* 9.45) was significantly different from chance, $t(13) = 7.07$, $p < 0.0001$, $d = 1.8$. Similarly, for the right edge stream, the overall score (mean 66.25%, *SD* 19) was significantly different from chance, $t(11) = 3.23$, $p < .01$, $d = 0.93$. In Fig. 4, the scores for edge-'words' and middle-'words' for the left- and right-edge groups are shown separately.

The edge-'words' at the left edge (mean 75.9%, *SD* 13.4) were recognized significantly above chance, $t(13) = 7.23$, $p < .0001$, $d = 1.93$. Similarly, edge-'words' at the right edge (mean 81.3%, *SD* 15.54) were recognized significantly above chance, $t(11) = 6.97$, $p < .0001$, 2.01. The middle- 'words' in the two conditions were less well recognized, left edge: 59.8%, *SD* 17.11, $t(13) = 2.15$, $p = .05$, .57, right edge: 54.17%, *SD* 27.87, $t(11) = 0.52$, $p = .62$, $d = 0.15$.

Pooling the data in an ANOVA with factors Edge (left or right) and Position (edge-'word' or middle-'word') revealed a significant effect of Position, $F(1,24) = 20.42$, $p < .001$, $\eta^2 = 0.26$. The Edge condition was not significant ($p > .9$, $\eta^2 = 0.00$), and neither was the interaction ($p > .2$, $\eta^2 = 0.02$). Post-hoc (Scheffe) tests revealed that the edge-'words' were recognized better than the middle-'words' in both groups (left edge, $p = .02$, $d = 1.05$, right edge, $p < .001$, $d = 1.2$).

Since the ANOVA revealed no main effect of Edge or interaction between Edge and Position, we collapsed the data from the left- and right-edge groups. The combined data revealed that edge-'words' were recognized better than chance, mean 78.4% (*SD* 14.4), $t(25) = 10.1$, $p < .0001$, $d = 1.97$, while the middle-'words' were not, mean 57.2% (*SD* 22.4), $t(25) = 1.64$, $p = .11$, $d = 0.32$. In addition, the score for the combined data for the edge-'words' was significantly different from the score for the combined data for the middle-'words,' $t(50) = 4.1$, $p < 0.001$, $d = 1.12$.

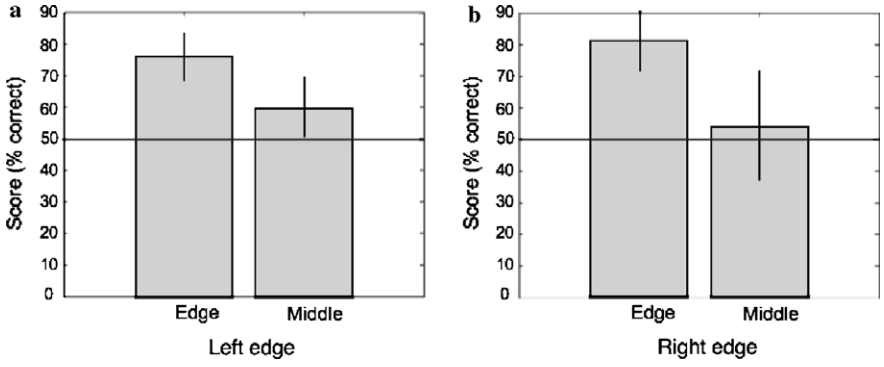

Fig. 4. Mean scores (% correct) for edge-'words' (Edge) and middle-'words' (Middle) from Experiment 3. In (a), edge-'words' occurred at the left edge of IPs, while in (b), edge-'words' occurred at the right edge of IPs. Edge-'words' are efficiently segmented, while middle-'words' are segmented with much less efficiency. Error bars represent 95% confidence limits of the means.

4.3. Discussion

The finding that edge-'words' are better recognized than middle-'words' suggests that prosody serves to chunk the speech stream. Recall that no preferences are observed for the 'words' or non-words when subjects are exposed to fully randomized streams (last paragraph of the Materials section, Experiment 1). In addition, in a statistically similar stream (apart from the number of tokens of the 'words'), all 'words' are significantly preferred over non-words (Experiment 2). Thus, the superiority of edge-'words' over middle-'words' is due to the fact that the former are placed at the left or right edges of prosodic breaks. Moreover, it is the edge that appears to cause the superiority of the edge-'words', since both the left and the right edges yield similar results. Last but not least, if we accept that prosody segments the streams into units, our results become compatible with the aforementioned experiments in human memory, wherein chunking an arbitrary list of verbal items results in a 'multiply-bowed recall curve' (Ng & Maybery, 2002), that is, in ∪-shaped curves within each of the chunks. By analogy, 'words' placed at the edges of prosodic contours should be better recalled than 'words' placed in their middles.

It is conceivable that the acoustic properties of prosodic breaks may recruit the attention of participants. If only increased attention to the vicinity of the breaks was responsible for the better recognition of edge-'words' in this experiment, we should have seen straddling 'words' being recognized better than middle-'words' in Experiment 1. Instead, the results suggest that attention is *aligned* with the edges of prosodic contours. That is, the focus of attention appears to be bounded by an edge, such that it toggles between being aligned with the right or the left edge of prosodic contours.

The results thus far are compatible with the model suggested in the introduction: prosody serves to organize a syllabic representation into 'phrases,' and TP computations yield the possible words inside each prosodic 'phrase.' There are, however, at least two possibilities as to how (higher-level) prosody might intervene: (a) prosody might directly segment the syllabic representation, such that TPs are *computed within* prosodically defined syllabic 'phrases,' or (b) prosody might act to *filter the output* of the TP system. The two possibilities are shown schematically in Fig. 5.

In the model in Fig. 5(a), prosodic information blocks the computation of TPs across two phrasal constituents. In 5(b) in contrast, the TP computations proceed automatically, and are unaffected by the prosodic properties of the speech stream. Instead, the local prosodic contours are co-indexed with all the 'words' that are the outputs of TP computations. This co-indexing results in straddling 'words' being disallowed, while no such filtering is obtained for the 'words' co-indexed with a non-straddling prosody.

Both possibilities outlined in Fig. 5 make the same predictions for the experiments described so far. But, there is an underlying difference between the two with regards to Experiment 1. In Experiment 1, we had contour-internal and contour-straddling 'words' during familiarization, and we found that only the contour-internal 'words' were correctly recognized in the test phase. The two proposals suggest different reasons for the failure of the contour-straddling 'words' to be recognized. According to proposal 5(a), the contour-straddling 'words' are not segmented at all. That is, prosody carves the input into 'phrases' and TP computations are bounded and are set to start again with every new 'phrases.' Therefore, straddling words are not ''visible'' to TP computations. In contrast, proposal 5(b) proposes that TP computations extract all 'words,' but straddling 'words' are suppressed during test.

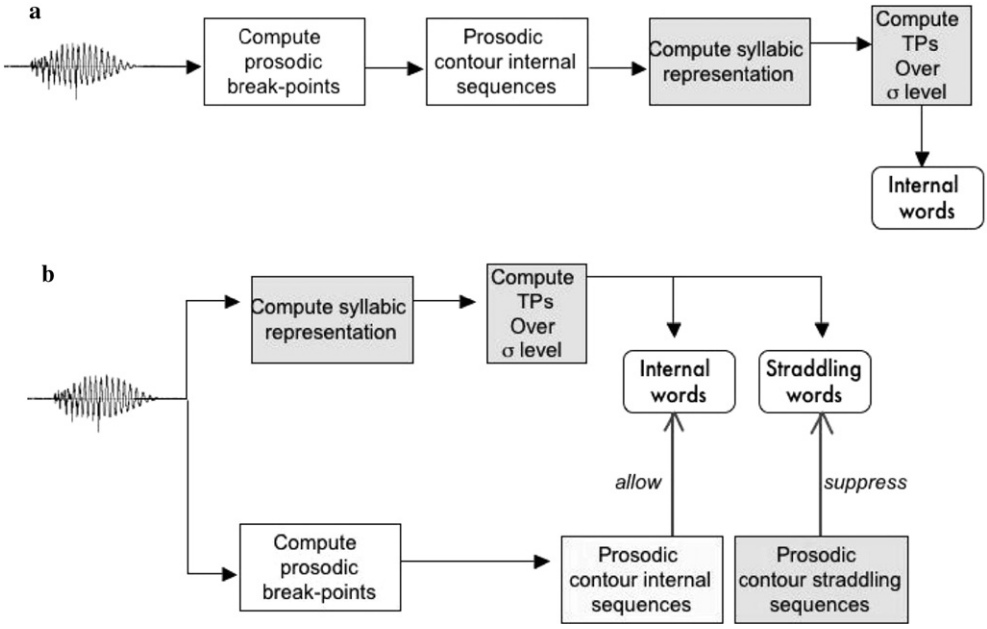

Fig. 5. Two possibilities for an interaction between prosody and statistics. In (a), the speech stream is broken up by prosody into 'phrases' (domains), and TPs are calculated only inside such domains. In (b), statistical analyses extract all possible 'words' from a syllabic representation. Prosody acts to suppress statistical 'words' that span prosodic boundaries.

If both internal and straddling 'words' are successfully segmented as suggested in 5(b), why are the straddling test items not recognized? One might argue that this is simply due to the acoustic differences in the way in which middle and straddling 'words' are realized during familiarization and test. Indeed, while familiarization in Experiment 1 used prosodic contours, test items were synthesized with a flat prosody; that is, all the phonemes were implemented with the same pitch and duration characteristics. It is known that the edges of prosodic domains are acoustically marked. Thus, the edges of contours might be acoustically more 'distant' from a flat contour than the middles. Hence, a straddling 'word' during familiarization, which contains two edges, would be more different from its flat counterpart during test. Conversely, a middle-'word' heard during familiarization, containing no edges, would be more similar to its flat counterpart during test.

However, such a proposal would predict that edge-'words,' which contain one edge, should be recognized worse than middle-'words,' which contain no edges. Instead, we find that edge-'words' are recognized significantly *better* than middle-'words'. This suggests that, rather than merely acoustic dissimilarity, it is the very fact that straddlers contain a perceptual break that causes them not to be recognized during test.

In addition, technically, to test the acoustic-distance hypothesis, one would need to match the items between familiarization and test in terms of their acoustic properties. But, given that the experimental material included a variety of contours, and that 'words' occurred at different positions in the different contours, it is difficult to determine what a prototypical middle or straddling contour should be. Second, even if we were to assume some model of what such prototypical contours should be like, participants might simply

react differently to the acoustic realization of items in the test phase containing two edges (straddlers) or no edges (middles).

Finally, there is a linguistic reason why such an acoustic control is not adequate: portions of IPs are themselves not well-formed IPs. As mentioned in the introduction, natural speech is organized in a hierarchical fashion, such that higher constituents dominate lower ones. Any utterance (which represents the top-most level of the prosodic hierarchy) must also contain at least one of each of the lower levels. Thus, the acoustic properties of words in isolation are different from the acoustic properties of the same word spoken as part of an IP. In other words, a portion of an IP (or of two adjacent IPs) is not well-formed in isolation, and it is not clear how participants might react to such tokens.

Nevertheless, we require empirical evidence to choose between the alternatives shown in Fig. 5. Thus, we conjectured that if, following the prosodic familiarization in Experiment 1, we could find a way to tap *only* the abstract, syllabic level of representation over which statistics are presumed to be computed, we might be in a position to find recall of not only the internal 'words,' but also the straddling 'words.' We tested this hypothesis in the next experiment.

5. Experiment 4

How can prosodic effects be bypassed? The model proposed in Fig. 5(b) suggests that perceptual inputs are analyzed by different subsystems. Thus, the incoming auditory information during familiarization might be encoded both as a sequence of prototypical representation of syllables and as a sequence of pitch movements or feature transitions.

To tap into the syllabic level, we decided to follow prosodic familiarization with a visual test phase. We hypothesized that, upon reading the test items, participants might generate internal representations that are equivalent to the stored representations of the segmented 'words' of the familiarization stream. This is especially the case for Italian, which has a very transparent orthography; we can thus make sure that the pronunciation of the visually presented letter strings correspond precisely to the artificial 'words.' Since the visual input does not contain acoustic information, a representation as a string of syllables will be dominant. If indeed all the 'words' had been extracted during familiarization as suggested by 5(b), then we would expect no difference in the recognition of contour-internal or contour-straddling 'words.' In contrast, if TP computations are restricted to the prosodic domains, then the presentation of a visual test phase will yield the same results as we find in Experiment 1.

5.1. Methods

5.1.1. Participants

Fourteen Italian adults (university students and researchers) participated in this experiment (4 males and 10 females, mean age 24.6 years, range 20–32 years).

5.1.2. Materials

The familiarization phase used the same artificial speech file as that of Experiment 1. In the test phase, instead of the two trisyllabic sequences presented aurally in each trial, the same items were presented visually on the screen. The first word was displayed to the left and the second to the right of the screen centre. The same instructions as for the previous experiments were used.

5.1.3. Apparatus and procedure

These were identical to Experiments 1 and 2 (apart from the visual modality of the test phase).

5.2. Results

The overall score, indicating correct segmentation of the speech stream was 66.96% (*SD* 14.80), and was significantly different from chance, $t(13) = 4.29$, $p < .001$, $d = 1.15$.

Fig. 6 shows that both the contour-internal and the contour-straddling 'words' were recognized significantly better than chance; contour-internal 'words' had a mean score of 66.96% (*SD* 21.77), $t(13) = 2.92$, $p = .01$, $d = 0.78$, and contour-straddling 'words' had a mean score of 66.96% (*SD* 20.57), $t(13) = 3.09$, $p < .01$, $d = 0.82$. The two groups were not different from each other, $t(26) = 0.0$, $p = 1$, $d = 0.0$.

The pattern of results from this experiment was compared to that from Experiment 1 in an ANOVA with factors Word Type (contour-internal or contour-straddling) and Test Type (auditory or visual). The results showed a main effect of Word Type, $F(1,32) = 5.1$, $p = 0.03$, $\eta^2 = 0.066$, a main effect of Test Type, $F(1,32) = 4.57$, $p = .04$, $\eta^2 = 0.053$ and a significant interaction between Word Type and Test Type, $F(1,32) = 5.1$, $p = .03$, $\eta^2 = 0.066$.

5.3. Discussion

The results from this experiment suggest that *all* statistically well-formed 'words' are extracted during prosodic familiarization. To explore whether these results were due to an insensitivity of the visual test to prosodic properties of the familiarization stream, we ran a control, replicating Experiment 3b with a visual test phase. In Experiment 3b we had found that 'words' that appear at the right edges of IPs were better recognized than 'words' in their middles. We ascribed this to the salience of edge positions. Since we know

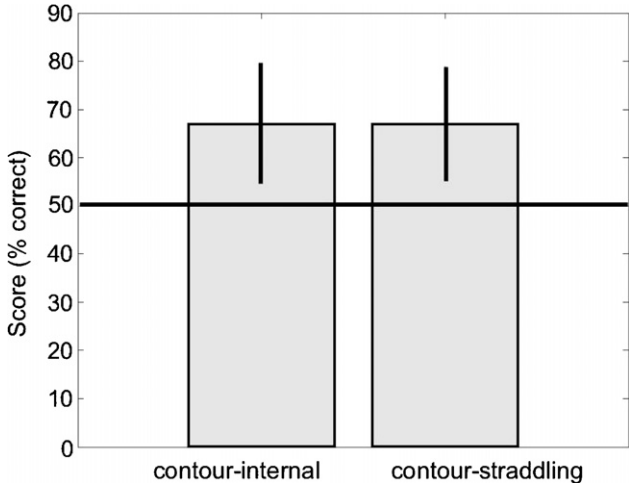

Fig. 6. Mean scores (% correct) and 95% limits for contour-internal 'words' and contour-straddling 'words' in a visual presentation of the test phase.

that edges are better encoded than middles, we expect the same results in the control (visual test words) as in Experiment 3b (auditory test words). Indeed, with the same familiarization as in Experiment 3b and a visual test phase, we found that edge-'words' (77.7%, *SD* 17.8) were better recognized than middle-'words' (61.6%, *SD* 22.7), $t(26) = 2.08$, $p = .047$, $d = 0.79$. Considering Experiment 3b and the visual-test control together, we found only a main effect of Position (edge-'words' were better recognized than middle-'words', $p < .001$, $d = 0.99$), while the modality of the test phase (auditory or visual) and the interaction between Position and Modality were not significant (both $p > .35$). Thus, we can conclude that visual presentation of the test phase is not insensitive to prosodic familiarization.

The finding that even the straddling 'words' are recognized better than chance is difficult to reconcile with the possibility, suggested in 5(a), that TPs are computed only inside syllabic 'phrases' aligned with prosodic domains. Instead, the results suggests that, in contrast to the model in 5(a), straddling 'words' are extracted by computing TPs, since they are segmented and recognized. The results thus favor the model depicted schematically in 5(b): TPs are computed over a syllabic representation of the speech stream, and prosody influences the output of the TP computational system. We propose that prosody acts as a *filter*; suppressing syllable sequences that straddle IP boundaries. We discuss some possible mechanisms for such a filtering effect in the general discussion.

This view is also coherent with the observation that TP computations appear to be implicit and automatic (Saffran, Newport, Aslin, Tunick, & Barrueco, 1997). Thus, TP computations over syllabic representations of speech might be an encapsulated, automatic process that is itself unaffected by other properties of the speech stream.

6. A universal mechanism?

It has been proposed that syllables are the natural units in which speech is perceived even by very young infants (Bertoncini & Mehler, 1980; Bertoncini & Mehler, 1981; Bertoncini, Bijeljac-Babic, Jusczyk, Kennedy, & Mehler, 1988; Mehler, 1981; Mehler, Dupoux, & Segui, 1990, but see Cutler, McQueen, Norris, & Somejuan, 2001) and statistical computations over syllables have been demonstrated in infants as young as 7 months of age (Thiessen & Saffran, 2003). However, it is not clear if the filtering effect of prosody could be obtained even in infants who have not had sufficient experience with the prosody of their language.

Indeed, it might be argued that Italian adults, with several years of experience with their language, might have learned to associate the edges of IPs with the edges of words. Then, such a strategy might not be available to very young infants, who might rely exclusively on TPs (as suggested in Thiessen & Saffran, 2003). If, instead, prosody *can* be used for bootstrapping word segmentation by breaking up fluent speech even in infants, it would greatly aid the process of speech segmentation. If we can show that adults rely on the proposed universal cues to IP boundaries that are constant across different languages, it would raise the possibility that infants, being sensitive to such cues, might use them in a like manner as well. We furnish such evidence in the next experiment.

7. Experiment 5

The present experiment explores whether Italian adults, like those that participated in the previous experiments, react in a like manner when faced with foreign IP contours. By

confronting adult participants with unfamiliar prosodic contours, we put them in a situation that might be analogous to that of a newborn listening to language. It has been proposed that IPs might have universal characteristics, like an initial pitch rise followed by a pitch decline, and final lengthening (e.g., Bolinger, 1964). If so, then, independently of other prosodic properties of a specific language, IP contours from any language might be equally effective in segmenting fluent speech into prosodic units. Thus, the 'universality' would consist in pitch and duration characteristics that define phrases without considering detailed pitch pattern variations within IPs, e.g., those due to language-specific tonal melodies, lexical stress and both the location and the physical realization of relative prominence in smaller phonological units (Nespor, Avesani, Donati, & Shukla, submitted for publication).

We thus chose to use Japanese IP characteristics instead of Italian ones, with comparable groups of Italian adults. There are two straightforward possibilities from such a manipulation. The first is that, in the face of an unheard prosody, Italian adults might not use prosody to filter the statistical computations, and so (a) both internal and straddling 'words' will be correctly recognized and (b) there will be no advantage for 'words' at edges with respect to 'words' in the middles of Japanese IPs.

A second possibility is that the aspects of prosody that are relevant for the filtering effect are the cues that are proposed to be universal, like the decline in pitch, such that Japanese IP prosody will evince the same results as those obtained with Italian IPs.

7.1. Methods

7.1.1. Participants

Fourteen Italian adults (university students and researchers) participated in this experiment (5 males and 9 females, mean age 25.3 years, range 19-36). None reported any contact with spoken Japanese.

7.1.2. Materials

To get Japanese IPs, a set of sentences were constructed.[6] Each set of sentences was constructed such that there was one clear IP corresponding to a single simple declarative clause, flanked by two IPs. The list of sentence sets is given in the Appendix A.

A single Japanese female speaker recorded the entire material. The material was recorded with an Audio-Technica ATR20 microphone connected to a SoundBlaster™ sound card on a PC under Window 2000™. CoolEdit (Syntrillium Corp.) was used to record and digitally manipulate the speech waveforms. The speech segments corresponding to the IPs were digitally excised. As in the Italian case, for each IP, we measured the pitch contour, smoothly interpolating across unvoiced segments using PRAAT (Boersma, 2001). A single pitch contour was converted into a vector of 400 pitch points. Thus, 20 pitch points per phoneme could be used to shape each of the 20 phonemes (from 10 CV syllables) in a single frame. From the nine recorded IPs, we thus obtained nine different pitch contour vectors; eight were used as in the experiments with Italian IPs.

[6] We are grateful to Yuki Hirose at the department of Human Communication, The University of Electro-Communications, Tokyo, Japan, and Hifumi Tsubokura at the Tokyo Women's Medical University, Tokyo, Japan, for help in the preparation of the material.

Fig. 7 shows a comparison of the average Japanese and Italian IP contours. As is clear from the figure, in both languages there is a final decline in pitch.

We next measured the durations of the first and last syllables of each Japanese IP. These durations were divided by the number of segments in the syllables to get a normalized value of the duration of each phoneme. The average normalized duration of the phonemes of the last syllable (99.8 ms) was significantly greater than the average normalized duration of the phonemes of the first syllable (73.19 ms), paired t-test, $t(8) = 3.72$, $p < .005$. These values correspond rather closely to those obtained for Italian, specially for the final syllable (Italian values: 79.9 and 99.6 ms for the initial and final phonemes respectively), and we decided to keep the duration of the final phonemes unchanged, while the duration of the initial phoneme was reduced by 5 ms. Thus, in these experiments, the duration of the phonemes comprising the first syllable was 95 ms (as opposed to 100 ms for Italian). The phonemes in the last syllable were 140 ms, and the remaining phonemes were 120 ms in duration, as for Italian.

Thus, both in terms of pitch decline and of final lengthening, the Japanese and Italian IPs we recorded show similar characteristics, although Japanese has a higher IP-initial pitch, and thereby a larger pitch-reset at IP boundaries.

To prepare the familiarization stream, the same MBROLA input file as used for Experiment 1 was used, and the eight Italian contours were each replaced by one of eight Japanese contours. This ensures that the statistical properties of the two sound streams (with Italian and with Japanese prosody) were largely matched. That is, they are identical for all the distributional properties at the level of the syllable, and the order of appearance of the IP contours (though not their identity, naturally) are identical. Thus, any difference in results could be attributed solely to the prosodic characteristics of the Japanese IPs.

The resulting file was converted to a wave file using MBROLA and the es1, Spanish male diphone database. The test phase was identical to Experiment 1.

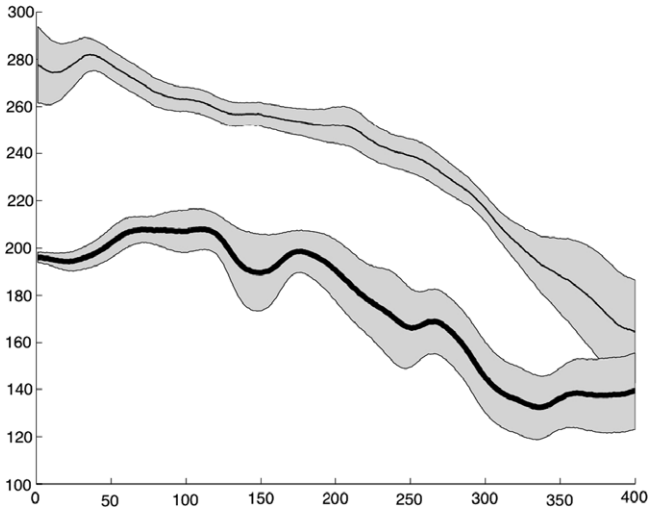

Fig. 7. Comparison of Italian and Japanese pitch contours. The shaded regions represent ±1 *SE*s around the means; thin line: Japanese, thick line: Italian. The x-axis represents (normalized) time, the y-axis is frequency.

7.1.3. Apparatus and procedure

These were identical to the Experiments 1–3.

7.2. Results

In Fig. 8, the results from this experiment are displayed. It is clear that Japanese prosody appears to have the same effect on Italian adults as Italian prosody: contour-internal 'words' are significantly preferred over the non-words, while the straddling 'words' are not.

Overall segmentation was 57.59% (*SD* 12.1), $t(13) = 2.36$, $p = .035$, $d = 0.63$. The mean score for the contour-internal 'words' was 73.21% (*SD* 17.58), $t(13) = 4.94$, $p < .0005$, $d = 1.32$, while the mean score for the straddling 'words' was 41.96% (*SD* 14.38), $t(13) = -2.09$, $p = .06$, $d = 0.56$. The score for the straddling 'words' thus reflects a tendency for the participants to prefer non-words. The two groups of 'words' were themselves significantly different, $t(13) = 5.15$, $p < .0001$, $d = 1.95$.

An ANOVA was used to compare Experiment 5 with Experiment 1 (wherein Italian prosody was used) to look for differences in using Italian or Japanese. Language (Italian or Japanese) was one factor, while Position (internal or straddling) was the other. The ANOVA revealed a main effect of Position, $F(1,64) = 36.48$, $p \leqslant .0001$, $\eta^2 = 0.36$, while there was no main effect of Language, $F(1,64) = 0.05$, $p = .82$, $\eta^2 = 0.00$. Also, there was no significant interaction between the two, $F(1,64) = 0.815$, $p = .37$, $\eta^2 = 0.01$.

7.3. Discussion

The result that prosody contributes to filtering straddlers suggests that prosody might contribute universal cues that organizes fluent speech. If Japanese and Italian IPs share

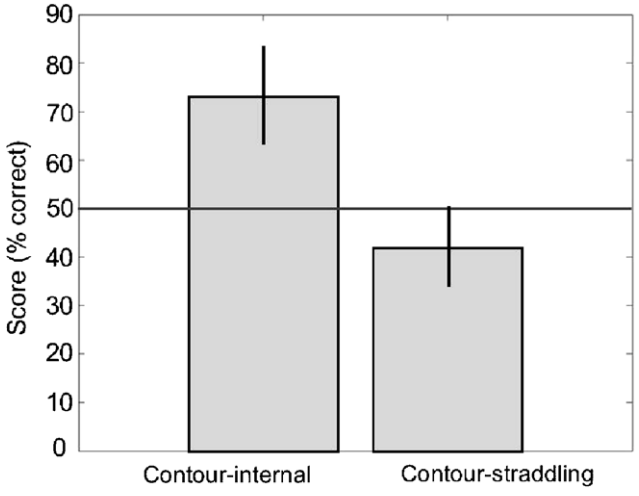

Fig. 8. Results of Experiment 6. Japanese IP characteristics replicate the findings with Italian IPs: IP-internal 'words' are accurately chosen, while straddling 'words' are not. Error bars represent 95% confidence limits of the mean.

universal cues that structure fluent speech, we expect to find that the edges of Japanese IPs are also more salient than their middles. Experiment 6 investigates if this is indeed the case.

8. Experiment 6

This experiment is aimed at establishing if there is an advantage for 'words' at the edges of prosodic contours over 'words' in the middle with Japanese IPs, as was the case with Italian IPs.

This experiment largely parallels Experiment 3. One group of participants was exposed to the stream with edge-'words' at the left edges of IPs and the other to the stream with edge-'words' at the right edges of IPs.

8.1. Methods

8.1.1. Participants

Fourteen Italian adults (university students and researchers) were exposed to the stream with edge-'words' at the left edge (4 males and 10 females, mean age 23.5 years, range 20–28 years). A separate group of twelve Italian adults were exposed to the stream with edge-'words' at the right edge (1 male and 11 females, mean age 22.9 years, range 19–27 years).

8.1.2. Materials

The preparation of the familiarization stream for this experiment paralleled that of the previous one. The MBROLA files from Experiment 3 (in which the edges of Italian IPs were examined) were used as a starting point, and each Italian IP contour in those files was replaced by one Japanese contour. Again, as in the previous experiment, this ensures that the distributional properties with respect to the syllables is matched in the present experiment and in Experiment 3.

The entire sequences of phonemes were fed to MBROLA, using the Spanish male diphone (es1) database. The final output files were 22.05 kHz, 16-bit, mono wave files of duration 4 min each. These files were converted into stereo files and the initial and final 5 s were ramped up and down to eliminate onset or offset cues to edge-'words' and middle-'words'. The test phase was identical for both groups of participants and was identical to Experiment 3.

8.1.3. Apparatus and procedure

These were identical to Experiment 3.

8.2. Results

The overall score for the left-edge stream group (mean 59.38%, *SD* 11.43) was significantly different from chance, $t(13) = 3.07$, $p < .01$, $d = 0.82$. However, for the right edge stream group, the overall score (mean 56.77%, *SD* 13.18) was not significantly different from chance, $t(11) = 1.78$, $p = .1$, $d = 0.51$. In Fig. 9, the scores for edge-'words' and middle-'words' for the left- and right-edge groups are shown separately.

The edge-'words' at the left edge (mean 74.11%, *SD* 18.65) were recognized significantly above chance, $t(13) = 4.84$, $p < .001$, $d = 1.3$. Similarly, edge-'words' at the right edge (mean 65.63%, *SD* 22.06) were recognized significantly above chance, $t(11) = 2.45$,

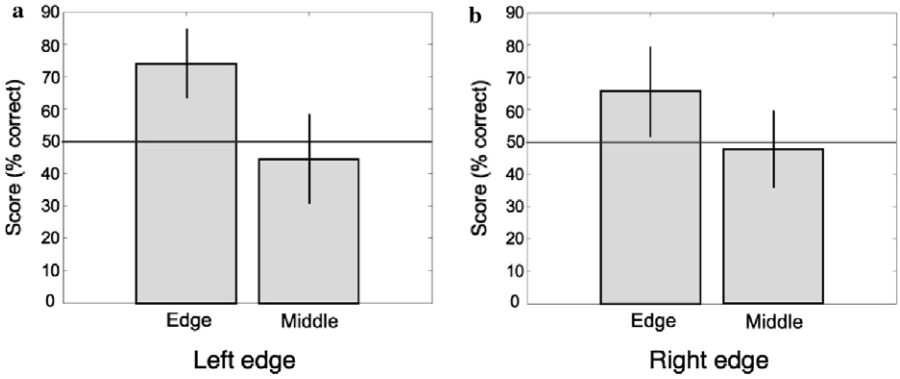

Fig. 9. Mean scores (% correct) for edge-'words' and middle-'words' from Experiment 6. In (a), edge-'words' occurred at the left edge of IPs, while in (b), edge-'words' occurred at the right edge of IPs. Edge-'words' are efficiently segmented, while middle-'words' are segmented with much less efficiency. Error bars represent 95% confidence limits of the means.

$p < .05$, $d = 0.71$. The middle-'words' in the two conditions were less well recognized, left edge: 44.64%, *SD* 24.37, $t(13) = -0.82$, $p = .43$, $d = 0.22$ and right edge: 47.92%, *SD* 19.09, $t(11) = -0.38$, $p = .71$, $d = 0.11$.

Pooling the data in an ANOVA with factors Edge (left or right) and Position (edge-'word' or middle-'word') revealed a significant effect of Position, $F(1,24) = 11.99$, $p = .002$, $\eta^2 = 0.24$. The Edge condition was not significant ($p > .6$, $\eta^2 = 0.003$), and neither was the interaction ($p > .3$, $\eta^2 = 0.02$). Post-hoc (Scheffe) tests revealed that the edge-'words' were recognized better than the middle-'words' for the left edge group, $p < .01$, $d = 1.36$, but not for the right edge group, $p = .09$, $d = 0.86$.

Since the ANOVA revealed no main effect of Edge or interaction between Edge and Position, we collapsed the data from the left- and right-edge groups. The combined data revealed that edge-'words' were recognized better than chance, mean 70.19% (*SD* 20.3) $t(25) = 5.1$, $p < .0001$, $d = 0.99$, while the middle-'words' were not, mean 46.15% (*SD* 21.7), $t(25) = -0.9$, $p > .3$, $d = 0.18$. In addition, the score for the combined data for the edge-'words' was significantly different from the score for the combined data for the middle-'words,' $t(50) = 4.12$, $p < .001$, $d = 1.14$.

A separate ANOVA compared the results from this Experiment (edges of Japanese contours) with Experiment 3 (edges of Italian contours). The factors were Language (Japanese or Italian), Edge (left or right) and Position (edge-'word' or middle-'word'). There was a main effect of Position, $F(1,96) = 32.5$, $p \leqslant .0001$ and a main effect of Language, $F(1,96) = 6$, $p = .016$, while the factor Edge was not significant. None of the two- or three-way interactions were significant. A post-hoc (Scheffe) test revealed that participants in the Italian condition performed better than those in the Japanese condition (overall, 9.7% greater accuracy in the Italian condition), $p = .016$, $d = 0.73$.

8.3. Discussion

The results from Experiments 5 and 6 using Japanese IPs replicate the pattern of results obtained in Experiments 1 and 3 with Italian IPs. Notice that Japanese is geographically, historically, and structurally very different from Italian. Despite these dissimilarities how-

ever, the overt realization of IPs from both languages contain cues that signal 'phrases' in otherwise fluent speech. In our experimental paradigm, these are indexed both by an advantage of IP-internal 'words' over straddling 'words,' as well as an advantage for edge-'words' over middle-'words.'

Comparing Experiments 3 and 6 (edge-'words' against middle-'words' with Italian or Japanese IPs) revealed a significantly better performance with Italian IPs. The better performance with Italian IPs was not observed while comparing Experiments 1 and 5 (internal 'words' against straddling 'words' with IPs from the two languages). Thus, although in some tasks familiarity with native prosody results in an advantage, nevertheless, the overt realization of IPs from both languages appear to contain cues that signal 'phrases' in otherwise fluent speech.

9. General discussion and conclusions

The aim of the present paper is to explore empirically how TP information and phrasal prosody interact with one another to extract word-like entities from fluent speech streams. We have developed an experimental paradigm that allows us to explore segmentation of fluent speech in streams consisting of sequences of syllables with superimposed prosodic contours. Our streams are thus comparable with speech that one would experience when confronted with a foreign language. In this novel paradigm, trisyllabic pseudo-'words' are interspersed with what we refer to as 'noise' syllables, that is, syllables that are not used to construct the 'words.' 'Noise' syllables occur at random, and are matched for frequency with the syllables that constitute the 'words.' This novel paradigm allows for a flexibility in the placement of target words in the investigation of the relative contributions of statistical and prosodic information.

Italian adult participants were exposed to such speech streams, which were followed by an auditory test phase with prosodically 'flat' test items. We showed that participants recognize 'words' internal to Italian IP contours, but not 'words' straddling two contours; both 'word' types were recognized in the absence of prosody (Experiments 1 and 2, respectively). In addition, in Experiment 3, 'words' at the edges of Italian IP contours were better recognized than 'words' in their middles. These results demonstrate that Italian IP prosody is utilized by Italian adults in processing fluent speech streams.

In Experiment 4 we discovered that with visual presentation of the test items, in contrast to the auditory presentation as in Experiment 1, straddling 'words' were successfully recognized, suggesting that statistical computations might be carried out automatically and independently of prosodic boundaries. Thus, prosody appears to act as a filter, suppressing sequences that span prosodic boundaries.

Finally, we found that the effects of prosody in the auditory modality (Experiments 1 and 3) were not due to the familiarity of the participants with Italian IPs: similar results were obtained when Italian participants were confronted with Japanese IPs (Experiments 5 and 6). This suggests that prosody contains universal cues that can carve fluent speech into smaller constituents. These cues can be used to filter sequences that straddle the boundaries of such constituents. Thus, prosody might serve to organize the speech stream in a way that is available even to very young infants (see below).

Taken together, the data can be accounted for by separating encoding and retrieval, corresponding to the familiarization and the test phases respectively. During encoding (familiarization), we suggest that the speech stream is converted into an abstract, syllabic

representation and distributional properties are computed over this syllabic representation. Sequences with high transition probabilities are proposed as possible lexical candidates. In parallel, suprasegmental cues highlight edges that define prosodic constituents. We found that 'words' that straddle IPs are not recognized. Interestingly, minimally sliding the straddling 'words' so as to align them with the edges of prosodic contours (e.g., the transformation of the material from Experiments 1 to 3) leads to their successful recognition. This suggests that IPs serve to break up the speech stream into smaller units.

Throughout the paper we have used the terms 'statistical computations' and 'transition probabilities' interchangeably. However, TPs capture only one kind of statistical regularity in the input. For example, several authors have suggested that keeping track of the frequencies of bi- or tri-syllables might help extract words from fluent speech (e.g., Swingley, 2005; Perruchet & Vinter, 1998; see, e.g., Brent & Cartwright, 1996; Christiansen, Allen, & Seidenberg, 1998 for other possibilities). In streams used in the experiments in this paper, both TPs and the frequencies of tri-syllables would result in the 'words' being extracted. Thus, we do not make any commitment to the exact nature of the statistical computations that might be used to extract 'words' from the middles of IPs in our experiments. Indeed, there might be multiple sources of statistical information that are used in conjunction.

How does prosody affect the word-like entities extracted using statistical computations? As we proposed above, though syllabic and prosodic information might be computed in parallel, at some stage they must interact (see Fig. 5(b)). We propose that during encoding, TP computations yield syllabic sequences that are potential 'words,' whose representations are linked to their episodic, prosodic properties. When a straddling 'word' is presented during test, its associated illicit prosody, namely one formed by the edges of two prosodic contours, causes it to be suppressed. The reason for its suppression, as suggested in the introduction, is that in natural language, prosodic units are invariably aligned with the edges of words. That is, sequences misaligned with the edges of phrasal prosodic constituents are not acceptable as possible lexical candidates. At this stage, we cannot be certain whether the suppression of straddlers results from experience with natural speech or from innate perceptual biases; see below for a discussion on this issue.

In Experiment 4, participants were tested with visually presented test items. The results show that, in contrast to Experiment 1, in Experiment 4, participants were able to recognize both middle and straddling items. Below, we discuss how the different results from the two experiments help clarify further the mechanism by which the filtering effect of prosody is obtained.

It has been proposed that the effectiveness of a cue in triggering recognition of an associated item is strongest when they are both stored at the same time (the encoding specificity hypothesis, Thomson & Tulving, 1970). An extension of encoding specificity is that there are strong contextual effects during encoding (see Bouton, Nelson, & Rosas, 1999 for a review). For example, Godden and Baddeley (1975) found that the extrinsic environment during the encoding of a list of words (on land or underwater in their experiment) had an effect on recall such that it was most effective when encoding and recall environments were the same. In the acoustic domain it has been shown that perceptual features (e.g., male or female voice) of spoken words aid in their subsequent recognition (e.g., Goldinger, 1996, 1998; Palmeri, Goldinger, & Pisoni, 1993). Indeed, based on several lines of evidence, Baddeley, 2000 proposed the existence of an episodic buffer, capable of providing a temporary store of incoming information and information from long-term memory in multimodal codes (see also Baddeley, 2001, 2003; Morey & Cowan, 2005).

Thus, we propose that during the auditory test phase, when a participant hears a test item, s/he will activate both a syllabic skeleton (a bare syllabic representation devoid of prosody and accidental acoustic properties) and the associated prosodic representation. We propose two distinct representations, the segmental (including syllables, as structured groups of segments) and the suprasegmental. It is thus in line with linguistic evidence that suggests the (quasi-)independence of the two levels (e.g., Goldsmith, 1976; McCarthy, 1982). When the conjunction of the two results in a syllabic skeleton misaligned with a prosodic boundary, the participant will reject the word.

In the visual test phase, the phonological representation as the bare syllabic skeleton will be predominantly activated. Given the visual modality of presentation, the acoustic, episodic properties will be less activated, and thus will interfere less with the recognition of the test items.

Note that while the 'words' (as syllable sequences) are phonological representations, what we refer to as the suprasegmental component might be either phonological or acoustic/phonetic. That is, at this stage it is not clear whether it is the 'low-level,' acoustic properties of IPs (like pitch decline) that causes a perception of edges, or whether adults actually perceive (abstract) phonological units. Indeed, the 'phrases' in the artificial speech streams we create can be regarded either as abstract phonological units, or as acoustic groupings without reference to phonological representations of any kind. In either case, mis-alignment with the boundaries of such units is proposed to cause the rejection of the straddling 'words.'

It might be that larger phonological units, like IPs, have their origins in more primitive physiological mechanisms (as suggested by Lieberman, 1967; see also Ohala et al., 2004 and Nespor et al., submitted for publication). Thus, phonological units would, by and large, be in correspondence with physiologically determined acoustic/phonetic units.

Our results suggest that TPs are computed across the sequence of segments, but the output of these computations are filtered by the suprasegmental component. If so, our findings have implications for language acquisition. Developmental studies have indicated that infants of around 11 months of age are able to integrate multiple cues for segmentation (Johnson & Jusczyk, 2001; Morgan & Saffran, 1995). More recently, Thiessen and Saffran (2003) examined the role of lexical stress and TPs in an experiment with 7- and 9-month-olds. These experiments were based on the observation that, while 2-month-old infants can dishabituate to a change in stress pattern (Jusczyk & Thompson, 1978), it is only by 7–9 months of age that (English speaking) infants use stress patterns in segmenting speech, prefer the predominant stress pattern of their native language, and treat strong-weak sequences (trochees) as forming units (Echols, Crowhurst, & Childers, 1997; Jusczyk et al., 1993, 1999). The question that these authors raised was whether the computation of distributional properties (like TPs over syllables) might show developmental changes. In particular, they tested the hypothesis that younger infants might rely more on statistical cues than on prosodic (stress) cues, when the two are in conflict, and obtained results that agree with such a hypothesis.

In contrast to a suprasegmental cue like word stress, whose location is language specific,[7] we have focused on phrasal prosodic constituents that may be signaled at least in part

[7] For example, Hungarian has primarily word-initial stress, Polish has penultimate stress and Turkish has word-final stress.

by universal properties. The prosodic patterns associated with IPs that we consider in this article might be a good candidate for such universal properties. Indeed, previous work has suggested that infants can, for example, detect IP boundaries (Hirsh-Pasek et al., 1987) and newborns discriminate bisyllables that are word internal from bisyllables that span phrasal boundaries (Christophe et al., 1994, 2001).

Thus, we might distinguish two kinds of suprasegmental cues to speech segmentation: those that need to be learnt based on the input and those that might be innately recognized. Many of the cues that have been studied in infant speech segmentation in addition to word stress, like allophone distributions or word-onset versus word-medial consonantal clusters (Hohne & Jusczyk, 1994; Jusczyk, Luce, & Charles-Luce, 1994; Jusczyk, Hohne, & Bauman, 1999; Mattys et al., 1999) must all be learnt from the input (see also Brent & Cartwright, 1996; Batchelder, 2002 for implementations of using such distributional cues in speech segmentation). This is not to say that stressed syllables, which are perceptually salient for newborns do not play any role in language acquisition (see, for example, Echols, 1993). However, the advantage of universal prosodic cues like, possibly, some aspects of IP prosody, is that these can be used in conjunction with statistical cues at a very early stage in language acquisition.

Indeed, results from Experiments 5 and 6 (using Japanese IPs) suggest that some aspects of prosody might be universal in nature. In these experiments, Italian participants did not choose contour-straddling 'words,' and showed an enhanced detection of 'words' at the edges of Japanese IP contours. Thus, Italian participants use the prosodic cues from Japanese IPs in a similar manner to how they use such cues from Italian IPs. However, these results by themselves do not necessarily predict that infants would also rely on these cues in a like manner. The realization of IPs in all natural languages might have some universal properties (like the decline in pitch), but the relationship between such properties and word boundaries might *not* be innately specified. It might be that, with several years of experience, Italian adults learn to associate IP boundaries with word boundaries, and can thus utilize similar cues from the prosody of a foreign language to constrain lexical search.

A second possibility is that the perception of edges, provided by prosody, is innately utilized to constrain lexical search. This would suggest that even very young infants, with limited language experience, would benefit from the prosody of speech to place constraints on where, in continuous speech, the lexical items might be present. As mentioned before, this filtering effect might be due to either a phonological or an acoustic analysis of the speech stream. Syllabic sequences that span 'phrasal' boundaries might not be considered as candidates for the mental lexicon. Alternatively, on subsequent encounters with such straddling sequences, learners might be biased to prefer parses that disrupt the 'phrase'-straddling sequences. Further experiments with infants are needed to clarify the role that prosody plays in the early stages of speech segmentation.

In conclusion, any prosodic grouping of speech, by virtue of being aligned with the edges of words, would aid in word segmentation. Although the present work is carried out with adults, our findings may contribute to and clarify early stages of language acquisition by showing how multiple cues interact. We suggest that the TP system is encapsulated; in our experiments the TP computation system operates over the segmental organization of speech. In addition, prosody might filter the output of such a TP system. Finally, we hypothesize that these effects of prosody might be mediated through episodic memory processes.

Appendix A.

The following Italian sentences were spoken by a single female Italian speaker. In each set of sentences, a phrase corresponding to a single intonational phrase (IP) was embedded. The portions corresponding to the intonational phrases are underlined.

1. È gia tardi, devi andare a scuola. Il latte è caldo, bevilo. Tutte le mattine la solita storia! (''It's already late, you have to leave for school. The milk is hot. Drink it up. Every morning it's the same old story!'')
2. Ti ho comprato lo sciroppo per la tosse. Bevilo tutto. Ti fa bene. (''I've bought you some syrup for your cough. Drink it all up. It will do you good.'')
3. Ascolta. Bevilo lentamente. È molto caldo. (''Listen. Drink it slowly. It's very hot.'')
4. Mi sembri un bambino di due anni. Non ridere quando bevi. Ti sbrodoli tutto. (''You look like a 2-year-old child. Don't laugh while drinking. You're making a mess all over you.'')
5. Ti ho preparato un po' di Scotch. Mettici il ghiaccio e bevi. Ma non troppo, visto che devi guidare. (''I've fixed you some Scotch. Put in some ice and drink it. But don't overdo it, since you have to drive.'')
6. Mi vergogno di te. Guarda come bevi. Sembri un bambino. (''I'm ashamed of you. Look how you drink. You look like a child.'')
7. Questo libro mi è molto caro. Lo metto via. Altrimenti rischio di perderlo. (''This book is very dear to me. I'll put it away. I might lose it otherwise.'')
8. Dicono che questo libro è molto bello. Lo penso anch'io. Anche se devo ancora finire di leggerlo. (''They say that this is a very nice book. I think so too. But I am yet to finish reading it.'')
9. Giovanni vuole il mio libro. Lo terrò nascosto. Altrimenti, se lo prende, rischia di non restituirmelo più. (''John wants my book. I'm going to keep it hidden. If he takes it, he might never return it.'')

The following Japanese sentences were spoken by a single female Japanese speaker. In each set of sentences, a phrase corresponding to a single intonational phrase (IP) was embedded. The portions corresponding to the intonational phrases are underlined.

1. Keito-wa kibun-ga warukatta. Kusuri-wo nonda. Kedo isha-ni itta. Kanojo-ha ima kibun-ga yokunatteiru. (''Keito was not feeling well. She took medicine. But she went to the doctor. She is feeling better now.'')
2. Keito-ha atama-ga itakatta. Isha-ni itta. Kanojo-ha kusuri-wo moratta. (''Keito had headache. She went to the doctor. She got medicine'').
3. Keito-ha netsu-ga atta. Kusuri-wo nonda kedo, isha-ni itta. Kanojo-ha haien datta. (''Keito had fever. Though she took medicine, she went to the doctor. She contracted pneumonia.'').
4. Keito-ha isshoukenmei-ni benkyousita. Nyuusi-ni shippaisita. Kedo isha-ni naritakatta. Kanojo-ha saido chousensuru. (''Keito studied hard. She failed in an entrance examination. But she would like to be a doctor. She will challenge it again.'')
5. Keito-ha igaku-ni kyoumi-ga atta. Isha-ni naritakatta. Kanojo-ha isshoukenmei-ni benkyousita. (''Keito was interested in medicine. She would like to be a doctor. She studied hard.'')

6. Keito-ha benkyousinakatta. Nyuusi-ni shippaisitakedo, isha-ni naritakatta. Kanojo-ha benkyousihajimeta. (''Keito didn't study. Though she failed in an entrance examination, she would like to be a doctor. She started to study.'')
7. Keito-ha Fukutsuu-ga sita. Naottato omotta. Kedo isha-ni denwasita. Kanojo-ha sugu isha-ni itta. (''Keito had abdominal pain. She thought that it got better. But she called a doctor. She went to the docter immediately.'')
8. Keito-ha kega-wo sita. Isha-ni denwasita. Isha-ha byouin-he ikuyouni itta. (''Keito was injured. She called a doctor. He said that she should go to the hospital.'')
9. Keito-ha keiren-wo okoshita.Naottato omotta kedo, isha-ni denwasita. Isha-ha annsei-ni suruyouni itta. (''Keito went into convulsions. She thought that they got better, but she called the doctor. He said that she should lie quietly.'')

References

Aslin, R. N., Saffran, J., & Newport, E. (1998). Computation of conditinal probability statistics by human infants. *Psychological Science, 9*, 321–324.

Baddeley, A. (1990). *Human memory*. Hove, UK: Lawrence Erlbaum Associates Ltd.

Baddeley, A. (2000). The episodic buffer: a new component of working memory? *Trends in Cognitive Sciences, 4*, 417–423.

Baddeley, A. (2001). Comment on cowan: the magic number and the episodic buffer. *Behavioral and Brain Sciences, 24*, 117–118.

Baddeley, A. (2003). Working memory: looking back and looking forward. *Nature Reviews Neuroscience, 4*, 829–839.

Bagou, O., Fougeron, C., & Frauenfelder, U. (2002). Contribution of prosody to the segmentation and storage of ''words'' in the acquisition of a new mini-language. In B. Bel & I. Marlien (Eds.), *Proceedings of the Speech Prosody 2002 conference* (pp. 59–62). Aix-en-Provence: Laboratoire Parole et Langage.

Batchelder, E. (2002). Bootstrapping the lexicon: a computational model of infant speech segmentation. *Cognition, 83*(2), 167–206.

Beckman, M. E., & Pierrehumbert, J. (1986). Intonational structure in Japanese and English. *Phonology Yearbook, 3*, 255–309.

Bertoncini, J., Bijeljac-Babic, R., Jusczyk, P., Kennedy, L., & Mehler, J. (1988). An investigation of young infants' perceptual representations of speech sounds. *Journal of Experimental Psychology: General, 117*(1), 21–33.

Bertoncini, J., & Mehler, J. (1980). Language perception in the newborn infant: some observations. *Reproduction Nutrition Development, 20*(3B), 859–869.

Bertoncini, J., & Mehler, J. (1981). Syllables as units in infant perception. *Infant Behavior and Development, 4*, 247–260.

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

Bolinger, D. (1964). Intonation as a universal. *Proceedings of the 9th international congress of linguistics, Cambridge, 1962* (pp. 833–844). The Hague: Mouton.

Bouton, M. E., Nelson, J. B., & Rosas, J. M. (1999). Stimulus generalization, context change, and forgetting. *Psychological Bulletin, 125*, 171–186.

Brent, M., & Cartwright, T. (1996). Distributional regularity and phonotactic constraints are useful for segmentation. *Cognition, 61*(1–2), 93–125.

Brown, G., Preece, T., & Hulme, C. (2000). Oscillator-based memory for serial order. *Psychological Review, 107*, 127–181.

Burgess, N., & Hitch, G. (1999). Memory for serial order: a network model of the phonological loop and its timing. *Psychological Review, 106*, 551–581.

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

Christophe, A., & Dupoux, E. (1996). Bootstrapping lexical acquisition: the role of prosodic structure. *The Linguistic Review, 13*, 383–412.

Christophe, A., Dupoux, E., Bertoncini, J., & Mehler, J. (1994). Do infants perceive word boundaries? An empirical study of the bootstrapping of lexical acquisition. *Journal of the Acoustical Society America, 95*(3), 1570–1580.

Christophe, A., Gout, A., Peperkamp, S., & Morgan, J. (2003). Discovering words in the continuous speech stream: the role of prosody. *Journal of Phonetics, 31*, 585–598.

Christophe, A., Mehler, J., & Sebastián-Gallés, N. (2001). Perception of prosody boundary correlates by newborn infants. *Infancy, 2*, 385–394.

Christophe, A., & Morton, J. (1998). Is Dutch native English? Linguistic analysis by 2-month-olds. *Developmental Science, 1*(2), 215–219.

Christophe, A., Nespor, M., Guasti, M. T, & van Ooyen, B. (1997). Reflections on phonological bootstrapping: its role in lexical and syntactic acquisition. In G. Altmann (Ed.), *Cognitive models of speech processing: A special issue of language and cognitive processes*. Mahwah, NJ: Lawrence Erlbaum.

Christophe, A., Peperkamp, S., Pallier, C., Block, N., & Mehler, J. (2004). Phonological phrase boundaries constrain lexical access: I. adult data. *Journal of Memory and Language, 51*, 523–547.

Cooper, W., & Paccia-Cooper, J. (1980). *Syntax and speech*. Cambridge, MA: Harvard University Press.

Cutler, A., Dahan, D., van, W., & Donselaar (1997). Prosody in the comprehension of spoken language: a literature review. *Language Speech, 40*(Pt 2), 141–201.

Cutler, A., McQueen, J. M, Norris, D., & Somejuan, A. (2001). The roll of the silly ball. In E. Dupoux (Ed.). *Language brain and cognitive development: Essays in honor of Jacques Mehler* (pp. 181–194). Cambridge, MA: The MIT Press.

Dahan, D., & Brent, M. (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*(2), 165–185.

Dutoit, T. (1997). *An introduction to text-to-speech synthesis*. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Ebbinghaus, H. (1964). *Memory: A contribution to experimental psychology*. Oxford, England: Dover.

Echols, C. (1993). A perceptually-based model of children's earliest productions. *Cognition, 46*(3), 245–296.

Echols, C., Crowhurst, M., & Childers, J. (1997). Perception of rhythmic units in speech by infants and adults. *Journal of Memory and Language, 36*, 202–225.

Endress, A., Scholl, B., & Mehler, J. (2005). The role of salience in the extraction of algebraic rules. *Journal of Experimental Psychology: General*.

Fisher, C., & Tokura, H. (1996). Prosody in speech to infants: direct and indirect acoustic cues to syntactic structure. In J. L. Morgan & K. Demuth (Eds.), *Signal to syntax: Bootstrapping from speech to grammar in early acquisition* (pp. 343–363). Hillsdale, NJ, England: Lawrence Erlbaum Associates, Inc.

Fowler, C., Smith, M., & Tassinary, L. (1986). Perception of syllable timing by prebabbling infants. *Journal of the Acoustical Society America, 79*(3), 814–825.

Friederici, A., Steinhauer, K., & Pfeifer, E. (2002). Brain signatures of artificial language processing: evidence challenging the critical period hypothesis. *Proceedings of the Natural Academic Science United States of America, 99*(1), 529–534.

Gerken, L., Jusczyk, P., & Mandel, D. (1994). When prosody fails to cue syntactic structure: 9-month-olds' sensitivity to phonological versus syntactic phrases. *Cognition, 51*(3), 237–265.

Godden, D., & Baddeley, A. (1975). Context-dependent memory in two natural environments: on land and under water. *British Journal of Psychology, 66*, 325–331.

Goldinger, S. (1996). Words and voices: episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology: Learning, Memory and Cognition, 22*(5), 1166–1183.

Goldinger, S. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review, 105*(2), 251–279.

Goldsmith, J. (1976). *Autosegmental phonology*. Unpublished doctoral dissertation, MIT, New York.

Gomez, R. (2002). Variability and detection of invariant structure. *Psychological Science, 13*(5), 431–436.

Gout, A., Christophe, A., & Morgan, J. L. (2004). Phonological phrase boundaries constrain lexical access: II. Infant data. *Journal of Memory and Language, 51*, 547–567.

Guasti, M. T (2002). *Language acquisition: The growth of grammar*. Cambridge, MA: The MIT Press.

Harris, Z. (1955). From phoneme to morpheme. *Language, 31*, 190–222.

Hayes, B. (1989). The prosodic hierarchy in meter. In P. Kiparsky & G. Youmans (Eds.), *Phonetics and phonology, Vol 1: Rhythm and meter* (pp. 201–260). San Diego: Academic Press.

Henson, R. (1998). Short-term memory for serial order: the Start End model. *Cognitive Psychology, 36*, 73–137.

Hirsh-Pasek, K., Kemler, D., Nelson Jusczyk, P., Cassidy, K., Druss, B., et al. (1987). Clauses are perceptual units for young infants. *Cognition, 26*(3), 269–286.

Hitch, G., Burgess, N., Towse, J., & Culpin, V. (1996). Temporal grouping effects in immediate recall: a working memory analysis. *Quarterly Journal of Experimental Psychology, 49A*, 116–139.

Hohne, E., & Jusczyk, P. (1994). Two-month-old infants' sensitivity to allophonic differences. *Perception & Psychophysics, 56*(6), 613–623.

Houston, D., Jusczyk, P., Kuijpers, C., Coolen, R., & Cutler, A. (2000). Cross-language word segmentation by 9-month-olds. *Psychonomic Bulletin & Review, 7*(3), 504–509.

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

Jusczyk, P. (1989). *Perception of cues to clausal units in native and non-native languages*. Paper presented at the biennial meeting of the Society for Research in Child Development: Kansas City, Missouri.

Jusczyk, P., Cutler, A., & Redanz, N. (1993). Infants' preference for the predominant stress patterns of English words. *Child Development, 64*(3), 675–687.

Jusczyk, P., Hohne, E., & Bauman, A. (1999). Infants' sensitivity to allophonic cues for word segmentation. *Psychonomic Bulletin & Review, 61*(8), 1465–1476.

Jusczyk, P., Houston, D., & Newsome, M. (1999). The beginnings of word segmentation in English-learning infants. *Cognition Psychology, 39*(3–4), 159–207.

Jusczyk, P., Pisoni, D., & Mullennix, J. (1992). Some consequences of stimulus variability on speech processing by 2-month-old infants. *Cognition, 43*(3), 253–291.

Jusczyk, P., & Thompson, E. (1978). Perception of a phonetic contrast in multisyllabic utterances by 2-month-old infants. *Perception & Psychophysics, 23*(2), 105–109.

Jusczyk, P. W., Luce, P., & Charles-Luce, J. (1994). Infants' sensitivity to phonotactic patterns in the native language. *Journal of Memory and Language, 33*, 630–645.

Kager, R. (1999). *Optimality theory*. Cambridge: Cambridge University Press.

Kemler, D., Nelson Hirsh-Pasek, K., Jusczyk, P., & Wright-Cassidy, K. (1989). How the prosodic cues in motherese might assist language learning. *Journal of Child Language, 16*(1), 55–68.

Klatt, D. (1976). Linguistic uses of segmental duration in English: acoustic and perceptual evidence. *Journal of the Acoustical Society America, 59*(5), 1208–1221.

Ladd, D. (1996). *Intonational phonology*. Cambridge: Cambridge University Press.

Lieberman, P. (1967). *Intonation, perception and language*. Cambridge, MA: MIT Press.

Marslen-Wilson, W., & Tyler, L. (1980). The temporal structure of spoken language understanding. *Cognition, 8*(1), 1–71.

Mattys, S., Jusczyk, P., Luce, P., & Morgan, J. (1999). Phonotactic and prosodic effects on word segmentation in infants. *Cognitive Psychology, 38*(4), 465–494.

McCarthy, J. (1982). *Formal problems in semitic phonology and morphology*. Unpublished doctoral dissertation, MIT, New York.

Mehler, J. (1981). The role of syllables in speech processing: infant and adult data. *Philosophical Transactions of the Royal Society, 295*, 333–352.

Mehler, J., Dupoux, E., Nazzi, T., & Dehaene-Lambertz, G. (1996). Coping with linguistic diversity: the infant's viewpoint. In J. Morgan & K. Demuth (Eds.), *Signal to syntax: Bootstrapping from speech to grammar in early acquisition* (pp. 101–116). Hillsdale, NJ, England: Lawrence Erlbaum Associates, Inc.

Mehler, J., Dupoux, E., & Segui, J. (1990). Constraining models of lexical access: the onset of word recognition. In G. T. M. Altmann (Ed.), *Cognitive models of speech processing: Psycholinguistic and computational perspectives* (pp. 236–262). Cambridge, MA: The MIT Press.

Mehler, J., Jusczyk, P., Lambertz, G., Halsted, N., Bertoncini, J., & Amiel-Tison, C. (1988). A precursor of language acquisition in young infants. *Cognition, 29*(2), 143–178.

Miller, G. (1956). The magical number seven plus or minus two: some limits on our capacity for processing information. *Psychol. Rev., 63*(2), 81–97.

Morey, C., & Cowan, N. (2005). When do visual and verbal memories conflict? the importance of working-memory load and retrieval. *Journal of Experimental Psychology: Learning, Memory and Cognition, 31*(4), 703–713.

Morgan, J., & Saffran, J. (1995). Emerging integration of sequential and suprasegmental information in preverbal speech segmentation. *Child Development, 66*(4), 911–936.

Morgan, J. L. (1994). Converging measures of speech segmentation in preverbal infants. *Infant Behavior and Development, 17*, 387–400.

Nazzi, T., Bertoncini, J., & Mehler, J. (1998). Language discrimination by newborns: toward an understanding of the role of rhythm. *J. Exp. Psychol. Hum. Percept. Perform., 24*(3), 756–766.

Nespor, M., Avesani, C., Donati, C., & Shukla, M. (submitted for publication). Different phrasal prominence realizations in vo and ov languages? Cognition.

Nespor, M., & Vogel, I. (1986). *Prosodic phonology*. Dordrecht: Foris.

Ng, H., & Maybery, M. (2002). Grouping in short-term verbal memory: Is position coded temporally? *Quarterly Journal of Experimental Psychology, 55A*, 391–424.

Ohala, J., Dunn, A., & Sprouse, R. (2004). Prosody and phonology. In B. Bel & I. Marlien (Eds.), *Speech prosody 2004*, Nara, Japan (pp. 161–163). ISCA Archive.

Onnis, L., Monaghan, P., Christiansen, M., & Chater, N. (2004). Variability is the spice of learning, and a crucial ingredient for detecting and generalising in nonadjacent dependencies. In *Proceedings of the 26th Annual Conference of the Cognitive Science Society*.

Palmeri, T., Goldinger, S., & Pisoni, D. (1993). Episodic encoding of voice attributes and recognition memory for spoken words. *Journal of Experimental Psychology: Learning, Memory and Cognition, 19*(2), 309–328.

Peña, M., Bonatti, L., Nespor, M., & Mehler, J. (2002). Signal-driven computations in speech processing. *Science, 298*(5593), 604–607.

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

Pierrehumbert, J., & Hirschberg, J. (1990). The meaning of intonational contours in the interpretation of discourse. In P. Cohen, J. Morgan, & M. Pollack (Eds.), *Intentions in communication*. Cambridge, MA: The MIT Press.

Ramus, F., Hauser, M., Miller, C., Morris, D., & Mehler, J. (2000). Language discrimination by human newborns and by cotton-top tamarin monkeys. *Science, 288*(5464), 349–351.

Saffran, J. (2001). Words in a sea of sounds: the output of infant statistical learning. *Cognition, 81*(2), 149–169.

Saffran, J., Aslin, R., & Newport, E. (1996). Statistical learning by 8-month-old infants. *Science, 274*(5294), 1926–1928.

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

Saffran, J., Newport, E., Aslin, R. N., Tunick, R., & Barrueco, S. (1997). Incidental language learning: listening (and learning) out of the corner of your ear. *Psychological Science, 8*, 101–195.

Sansavini, A., Bertoncini, J., & Giovanelli, G. (1997). Newborns discriminate the rhythm of multisyllabic stressed words. *Developmental Psychology, 33*(1), 3–11.

Selkirk, E. (1984). *Phonology and syntax: The relation between sound and structure*. Cambridge, MA: The MIT Press.

Selkirk, E. (1996). The prosodic structure of function words. In J. L. Morgan & K. Demuth (Eds.), *Signal to syntax: Bootstrapping from speech to grammar in early acquisition* (pp. 187–213). Hillsdale, NJ, England: Lawrence Erlbaum Associates, Inc..

Shattuck-Hufnagel, S., & Turk, A. (1996). A prosody tutorial for investigators of auditory sentence processing. *Journal of Psycholinguistic Research, 25*(2), 193–247.

Soderstrom, M., Seidl, A., Kemler Nelson, G., Deborah Jusczyk, W., & Peter (2003). The prosodic bootstrapping of phrases: Evidence from prelinguistic infants. *Journal of Memory and Language, 492*, 249–267.

Steinhauer, K. (2003). Electrophysiological correlates of prosody and punctuation. *Brain Langauge, 86*(1), 142–164.

Steinhauer, K., Alter, K., & Friederici, A. (1999). Brain potentials indicate immediate use of prosodic cues in natural speech processing. *Nature Neuroscience, 2*(2), 191–196.

Steinhauer, K., & Friederici, A. (2001). Prosodic boundaries, comma rules, and brain responses: the closure positive shift in ERPs as a universal marker for prosodic phrasing in listeners and readers. *Journal of Psycholinguistic Research, 30*(3), 267–295.

Swingley, D. (2005). Statistical clustering and the contents of the infant vocabulary. *Cognitive Psychology, 50*, 86–132.

Thiessen, E., & Saffran, J. (2003). When cues collide: use of stress and statistical cues to word boundaries by 7- to 9-month-old infants. *Developmental Psychology, 39*(4), 706–716.

Thomson, D., & Tulving, E. (1970). Associative encoding and retrieval: weak and strong cues. *Journal of Experimental Psychology, 86*, 255–262.

Vaissière, J. (1995). Phonetic explanations for cross-linguistic prosodic similarities. *Phonetica, 52*, 123–130.

Watson, D., & Gibson, E. (2004). The relationship between intonational phrasing and syntactic structure in language production. *Language and Cognitive Processes, 19*(6), 713–755.

Wightman, C., Shattuck-Hufnagel, S., Ostendorf, M., & Price, P. (1992). Segmental durations in the vicinity of prosodic phrase boundaries. *Journal of the Acoustical Society America, 91*(3), 1707–1717.