Available online at www.sciencedirect.com
SCIENCE DIRECT®
COGNITION
ELSEVIER
Cognition 97 (2005) B25–B34
www.elsevier.com/locate/COGNIT

Brief article

Speech segmentation by statistical learning depends on attention

Juan M. Toro, Scott Sinnett, Salvador Soto-Faraco*

Grup de Recerca Neurociencia Cognitiva, Departament de Psicologia Bàsica, Parc Cientific, Universitat de Barcelona, Pg. Vall d'Hebron, 171, 08035 Barcelona, Spain

Received 19 November 2004; accepted 27 January 2005

Abstract

We addressed the hypothesis that word segmentation based on statistical regularities occurs without the need of attention. Participants were presented with a stream of artificial speech in which the only cue to extract the words was the presence of statistical regularities between syllables. Half of the participants were asked to passively listen to the speech stream, while the other half were asked to perform a concurrent task. In Experiment 1, the concurrent task was performed on a separate auditory stream (noises), in Experiment 2 it was performed on a visual stream (pictures), and in Experiment 3 it was performed on pitch changes in the speech stream itself. Invariably, passive listening to the speech stream led to successful word extraction (as measured by a recognition test presented after the exposure phase), whereas diverted attention led to a dramatic impairment in word segmentation performance. These findings demonstrate that when attentional resources are depleted, word segmentation based on statistical regularities is seriously compromised.
© 2005 Elsevier B.V. All rights reserved.

Keywords: Language; Attention; Speech Segmentation; Statistical Learning

1. Introduction

One of the challenges faced by the perceptual system when processing spoken language is to segment a continuous acoustic flow into discrete lexical units.

* Corresponding author. Tel.: +34 93 3125158; fax: +34 93 4021363.
E-mail address: ssoto@ub.edu (S. Soto-Faraco).

0022-2860/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.cognition.2005.01.006

This challenge is even greater during language acquisition in pre-linguistic children, as lexically based strategies are not available. Several recent studies have highlighted the important role that learning mechanisms based on the detection of statistical regularities play in order to achieve speech segmentation (Saffran, Aslin, & Newport, 1996; Saffran, Newport, & Aslin, 1996)—as well as in other aspects of language development (Newport & Aslin, 2004; Gómez, 2002; Maye, Werker, & Gerken, 2002; Saffran, 2001, 2002; Saffran & Thiessen, 2003). An important question is, however, whether word extraction based on statistical learning requires attentional resources or, if it occurs automatically whenever the appropriate type of input (i.e. speech) is present, regardless of the listener's attention.

Several results support the idea that speech segmentation based on statistical learning can indeed occur in the absence of explicit instructions to the observer. For instance, human infants (Saffran, Aslin et al., 1996), monkeys (Hauser, Newport, & Aslin, 2001), and even rats (Toro & Trobalón, in press), none of which can possibly be instructed to attend to the stimuli, can segment speech streams using statistical computations. In fact, Saffran, Newport, Aslin, Tunick, and Barrueco (1997) measured speech segmentation based on statistical learning in adults and children while the observers performed an unrelated drawing task (i.e. in a completely incidental situation). The authors reported effective segmentation of the speech stream in both populations, even though attention was not explicitly directed towards the stimuli. However, none of these findings are conclusive as to the possible role of attention in speech segmentation by statistical learning. As it has emerged from the long lasting debate between early and late selection theorists in attention, task irrelevant information may undergo some processing regardless of the instructions (or lack of instructions) given to the observer, especially if the main task is not demanding (Lavie, 1995; Rees, Frith, & Lavie, 2001; Rees, Russell, Frith, & Diver, 1999). For example, several processes that have been traditionally thought of as automatic (such as word reading or visual motion after-effects) can be affected, or even prevented altogether, if attentional resources are completely depleted (Rees, Frith, & Lavie, 1997; Rees et al., 1999; Sinnett, Costa, & Soto-Faraco, submitted). Consequently, although Saffran et al. (1997, p. 102) suggest that the incidental situation of their task prevented participants from focusing their attention on the speech stream, there is no actual guarantee that participants did not occasionally direct their attention to the irrelevant speech stream while they were performing the free drawing task. Moreover, it is difficult to know the degree to which attention was engaged in the free drawing task used in that study. Note, for example, that overall performance in Saffran et al. (1997) seems to be poorer than in other studies using very similar materials in nonincidental situations (e.g. Saffran, Newport et al., 1996).

The goal of the present study was to directly test the extent to which attention is necessary for speech segmentation by statistical learning. We compared performance on a standard speech segmentation task when participants could focus attention on the speech stream, or when they diverted attention to a difficult auditory (Experiments 1 and 3) or visual (Experiment 2) concurrent task.

2. Experiment 1. Diverting auditory attention from the speech stream

2.1. Participants

Forty undergraduate students of the University of Barcelona participated in Experiment 1 in exchange for course credit. All reported normal hearing and normal or corrected-to-normal vision.

2.2. Stimuli and apparatus

The materials for the present experiments were a replica from the artificial language formed by four trisyllabic nonsense words (*tupiro*, *golabu*, *bidaku*, *padoti*) used by Saffran, Aslin & Newport (1996, Language A). These four words were concatenated in a pseudo-random sequence without any immediate word repetitions. Transitional probability between the syllables forming a word was 1.0, whereas transitional probability between syllables spanning word boundaries was 0.33. The resulting stream was synthesized using text-to-speech MBROLA software (Dutoit, Pagel, Pierret, Bataille, & van der Vrecken, 1996) with a Spanish male diphone database[1] at 16 kHz. Each syllable lasted 232 ms and, of crucial importance, there were no acoustic markers between words in the stream. The only available cue for word segmentation was the statistical distribution of syllables within and between words. The recognition test administered at the end contained the four words mentioned above, plus four part-words made by the concatenation of the third syllable of a word and the first two syllables of another word (*tibida*, *kupado*, *rogola*, *butupi*).

A stream of noises of common objects selected for high familiarity and clarity by three independent judges was also created (i.e. car engine, door slamming, etc)[2]. All sounds were adjusted to lengths between 400 and 500 ms and to an average amplitude equivalent to the word stream in dB. Inter-stimulus intervals (ISI) of 250 ms were introduced between the sounds; giving an SOA (Stimulus Onset Asynchrony) ranging from 650 to 750 ms. The sound stream was mixed with the word stream to create two overlapping (but uncorrelated) auditory streams.

2.3. Procedure

Participants were divided in two groups (Passive listening condition, and High-attention load condition), and told that they would listen to a nonsense language and a stream of sounds. Participants in the passive listening condition ($n=20$) were simply asked to listen to what was played. Participants in the high-attention load condition ($n=20$) were asked to press a button ("B" on the keyboard) each time they detected a repetition in the sound stream (one every four sounds, in average). In order to succeed in this task, close attention to the sound stream is required. No participants in any of the two

[1] Available at http://tcts.fpms.ac.be/synthesis/mbrola.html.
[2] Available at http://www.a1freesoundeffects.com.

groups were given any explicit information regarding the speech stream, nor were they told to attempt to figure out what the words were. After a monitoring phase of 7 min duration, all participants were given a 2 alternative forced choice (2AFC) recognition test. They heard a word and a part-word (500 ms ISI, order counterbalanced) and judged whether the first or the second item was more likely to have been a word in the language they had just heard. The next test pair was presented after the participant pressed a response key, or after a 5 s deadline. Following previous applications of this paradigm, we included eight test trials in order to avoid unnecessary repetitions of test items.

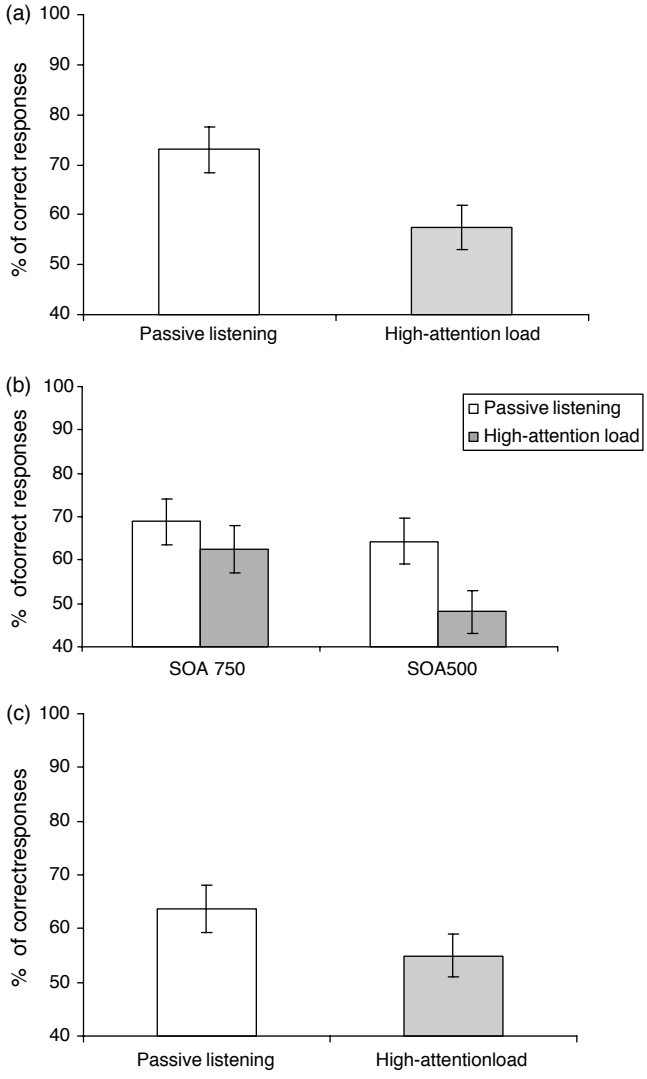

Fig. 1. Percentage of correct responses (and standard error) in the 2AFC word recognition test for the Passive listening and the High-attention load groups in Experiments 1–3 (graphs a–c, respectively).

2.4. Results

Performance in the sound repetition task for the High-attention load group was 62% hits and 5% false alarms. The percentage correct in the test for word segmentation was significantly different across the two attention conditions ($t(19) = -2.17$, $P < 0.05$). The Passive listening group performed at 73% (SE = 4.5), significantly better than chance ($t(19) = 5.11$, $P < 0.001$) (see Fig. 1a). The high-attention group performed at 58% (SE = 4.4), not different from chance ($t(19) = 1.79$, $P = 0.089$). These results suggest that when attentional resources are diverted to a secondary task the detection of statistical regularities is severely taxed.

However, one potential alternative account for the present result is that, rather than compromising the extraction transitional regularities, divided attention simply prevented participants from perceptually segregating the speech stream from the sound stream, thus making word segmentation impossible altogether because of masking. In Experiment 2 we wanted to ensure that the observed results were not a consequence of some kind of sensory masking produced by the noises. We used a visual stream of pictures instead of the sounds, thereby disabling any low-level sensory masking.

3. Experiment 2. Diverting visual attention from the speech stream

3.1. Participants

Eighty undergraduate students of the University of Barcelona participated in Experiment 2 in exchange for course credit. All reported normal hearing and normal or corrected-to-normal vision.

3.2. Stimuli and apparatus

The same artificial language and test items as in Experiment 1 were used. A rapid serial visual presentation (RSVP) stream of pictures, chosen from the Snodgrass and Vanderwart picture database (Snodgrass & Vanderwart, 1980), replaced the sound stream. Half of the pictures (at random) were rotated 30° to the right, and the other half 30° to the left. For half of the participants picture duration was 500 ms and ISI was 250 ms (750 ms SOA), whereas for the rest of participants picture duration was 250 ms (500 ms SOA). The RSVP of pictures was displayed on a computer monitor concurrently with the auditory artificial language.

3.3. Procedure

We manipulated SOA (500 vs. 750 ms) and attention condition (Passive listening or high attention load) between participants. All participants were told that they would hear an artificial language and would also see some pictures on the computer screen. Participants in the Passive listening groups ($n = 20$ each) were asked to listen and watch to what they heard and saw. Participants in the high-load conditions ($n = 20$ each) were asked

to press a button ("B" on the keyboard) each time they detected a picture repetition. After 7 min, all participants were given a recognition test identical to that in Experiment 1.

3.4. Results

In the high attention load conditions, performance in the picture repetition task was 60% hits and 2% false alarms in the 750 ms SOA group, and 62% hits and 8% false alarms in the 500 ms SOA group. In the word recognition test, both 750 ms SOA groups (the passive listening and high-load group) performed at an equivalent level ($t(19)=1.0$, $P=0.33$), all above chance (69%, SE=4.2, $t(19)=4.5$, $P<0.001$; and 63%, SE=5.4, $t(19)=2.3$, $P<0.05$, respectively). In the 500 ms SOA condition, however, the passive listening group obtained a recognition rate of 65% (SE=3.8; significantly better than chance, $t(19)=3.8$, $P<0.001$), outperforming the high-load group (48%, SE=5.1), which obtained a score no better than chance ($t(19)=-0.37$, $P=0.716$) (see Fig. 1b).

When the visual distracter task was presented at a rate of 750 ms per item (the same as in Experiment 1), participants were still able to extract the statistical regularities present in the language stream even when asked to direct their visual attention to the picture stream (suggesting the possibility that they may have been able to switch attention between the two streams when attention was not fully engaged in the primary task). However, when the difficulty of the distracter task was increased, by speeding up the presentation rate to an SOA of 500 ms, no evidence for statistical learning was seen in the diverted attention condition. As in Experiment 1, the present data strongly suggest that speech segmentation based on statistical learning is not an attention-free process[3].

In Experiments 1 and 2 we introduced a manipulation that involved diverting attention from the speech stream to a different stream (auditory or visual, respectively). One classic finding in attention literature is that resources can be more successfully divided amongst parts of the same object than across different objects (i.e. Duncan, 1979). Indeed, a recent study on visual perceptual learning by Baker, Olson, and Behrmann (2004) suggests that the extraction of statistical correlations among elements of separate objects can be prevented when attention is not directed to both objects, but they can be extracted when attention is directed to one of the two elements within a single object. Can it be the case that word extraction can only be prevented if attention is diverted from the speech stream to a different stream?

In order to further characterize the role of attention on speech segmentation, we created a situation where the distracting task had to be performed on the speech stream itself.

[3] One controversial issue in literature about the role of attention on sequence learning (Hsiao & Reber, 1998 for a review) is the possibility that the secondary task reduces working memory capacity, thereby explaining the poorer performance in the main task. This does not seem to be the case here, as our secondary task implies minimal (or null, in Experiment 3) working memory load. Nevertheless, we ran a further control study identical to Experiment 2 with 750 ms SOA but using a 2-back task (i.e. detect non-consecutive picture repetitions), which is considerably more demanding in terms of working memory than the 1-back task of Experiment 2. Participants performing the 2-back task obtained a word recognition rate of 62% (SE 4.6), that was significantly better than chance ($t(19)=2.67$, $P<0.05$) and equivalent to the passive condition result. So this increase in memory demands did not produce any effects, suggesting that it is attention, not working memory, the process responsible for the results reported here.

In Experiment 3 we introduced pitch variations in some randomly selected syllables in the stream and asked participants to detect these changes. Thus, participants' attention was directly placed on the speech stream, yet focused on an acoustic feature different from the one carrying the statistical regularities affording the extraction of words.

4. Experiment 3. Diverting auditory attention within the same speech stream

4.1. Participants

Forty undergraduate students of the University of Barcelona participated in Experiment 3 in exchange for course credit. All reported normal hearing and normal or corrected-to-normal vision.

4.2. Stimuli and apparatus

The pitch of some pseudo-randomly selected syllables (one every ten, on average) was changed by 20 Hz above or below the baseline fundamental frequency (200 Hz). Care was taken to ensure that pitch changes did not consistently signal a given syllable, or syllables in any given word position (either word initial, medial or final). For this experiment no additional stream (either of sounds or pictures) was presented. Otherwise, the stimuli remained as in Experiment 1.

4.3. Procedure

Participants were divided in two groups ($n=20$, each); Passive listening and High-attention load. Participants in the Passive listening condition were told they would hear an artificial language and their task was to simply listen. Participants in the High-attention load condition were told they would hear an artificial language with slight pitch changes. Their task was to press a button ("B" on the keyboard) each time they detected a change in pitch. After 7 min participants in both groups were given a 2AFC test like the one used in Experiments 1 and 2.

4.4. Results

The mean percentage of correct word recognition in the 2AFC test[4] was significantly different across both conditions ($t(19)=-2.14$, $P<0.05$). The Passive listening group performed at 64% (SE=3.8), different from chance ($t(19)=3.48$, $P<0.01$). The High-attention load group performed at 55% (SE=4.5), not different from chance ($t(19)=1.11$, $P=0.282$, see Fig. 1c). The main finding of Experiment 3 was that

[4] No data could be collected for the detection of pitch changes in Experiment 3 because of technical difficulties. However, we ensured beforehand that the pitch detection task was difficult enough via a pilot study. Moreover, this does not create any problem in interpreting the word segmentation data, as the results of Experiment 3 show, again, that the attentional manipulation was effective in reducing segmentation performance to chance levels.

segmentation by statistical learning faltered even if the concurrent distracter task involved focusing attention on the speech stream itself. Furthermore, a secondary finding to emerge from this experiment was that (in the passive listening group) participants were able to segment the speech stream despite the uninformative (uncorrelated) nature of pitch (a fundamental component of word stress; see Toro & Sebastián-Gallés, 2004 for a related finding). Given the prominent role attributed to statistical learning in language acquisition literature, it is interesting to note that recent studies have revealed a reversal in the relative weight of stress and distributional regularities as cues for word segmentation. At 7 months of age stress seems to be more prevalent than statistics, whereas at 9 months statistics dominates over stress (Johnson & Jusczyk, 2001; Thiessen & Saffran, 2003).

5. General discussion

The main conclusion to emerge from this study is that, under conditions of inattention, speech segmentation based on statistical learning can be seriously compromised. As in previous studies, we found successful word segmentation when participants were not given a particular task other than to listen. Yet, when attention was diverted to a difficult unrelated task, word segmentation performance dropped to chance levels. This effect occurred when attention was diverted to a different stream in the same sensory modality (auditory) or even to a different feature within the very same speech stream[5]. When attention was diverted to a different modality (vision), speech segmentation also fell to chance levels, although only when speeding up the presentation rate of the distracter stream (arguably making the distracter task more difficult). This result coincides with the traditional view that it is more difficult to find attention costs across modalities than within modalities (i.e. Duncan, Martens, & Ward, 1997; Soto-Faraco & Spence, 2002; Treisman & Davies, 1973; Wickens, 1984).

Saffran et al. (1997) showed that statistical learning took place while participants performed an incidental task completely unrelated to the speech input. Their concurrent task, however, could hardly allow for a strict control of attentional load because it consisted of self-paced free drawing. Our results from Experiment 2 (750 ms SOA) replicate Safran's (1997) findings, as statistical learning was found in an incidental situation, when the speech stream was combined with a visual distracter task. However, when the concurrent task was more demanding (500 ms SOA) or within the same sensory modality as the speech stream, performance on the segmentation task dropped to chance levels. Thus, the present results clearly show that at least some attentional resources must be available and directed to the speech stream in order to segment it. This conclusion is along the same line as other results in the field of language perception, specifically, on word reading (Rees et al., 1999) and audiovisual speech integration (Alsius, Navarra, Campbell, & Soto-Faraco, in press), both basic processes that traditionally have been

[5] There is the possibility that divided attention does not completely prevent, but only slows down, statistical learning. This is difficult to test directly, as any effects of divided attention could always be attributed to short exposure periods. In any case, our results suffice to show that, with the same amount of exposure in all conditions, divided attention does strongly affect (in principle, inhibiting) the segmentation of the speech stream.

thought of as attention-independent. Thus, our data add to these recent demonstrations and call for some discretion in the interpretation of previous results indicating segmentation under incidental situations.

In conclusion, the present results show that even if speech segmentation based on statistical learning can occur without the need of explicitly instructing the listener to focus on the speech stream, it is clear that some degree of attention is needed to attain word extraction successfully. Another question is, for instance, whether certain features of speech are in fact a very salient stimulus for the human, thus, potentially recruiting whatever attentional resources left available that are not being used in other processes, or even capturing attention already allocated to less salient stimuli. This is especially important in the context of language development. Even though it is not necessary that the infant's attention is solely devoted to any linguistic input in the environment (Saffran et al., 1997), the speech signal seems to have a privileged status for infants as young as 2.5 months old (Vouloumanos & Werker, 2004). Even other variables such as those present in personal interactions can eventually capture infant's attention and help improve performance in linguistic tasks (Kuhl, Tsao, & Liu, 2003).

Acknowledgements

This research was supported by a grant from Nissan Motors Co., the Human Frontier Science Program (HFSP) grant RGP68/2002, the Spanish Ministerio de Educación y Ciencia grants SEJ2004-07680-C02-01/PSIC and TIN2004-04363-C03-02, and a Spanish MECD fellowship AP2000-4164. Correspondence concerning this article should be addressed to SS-F (email: ssoto@ub.edu).

References

Alsius, A., Navarra, J., Campbell, R., Soto-Faraco, S. (in press). Audiovisual speech integration falters under high attentional demands. Current Biology.

Baker, C., Olson, C., & Behrmann, M. (2004). Role of attention and perceptual grouping in visual statistical learning. *Psychological Science*, *15*, 460–466.

Duncan, J. (1979). Divided attention: The whole is more than the sum of its parts. *Journal of Experimental Psychology: Human Perception and Performance*, *5*, 216–228.

Duncan, J., Martens, S., & Ward, R. (1997). Restricted attentional capacity within but not between sensory modalities. *Nature*, *387*, 808–810.

Dutoit, T., Pagel, V., Pierret, N., Bataille, F., & van der Vrecken, O. (1996). *The MBROLA project: Towards a set of high-quality speech synthesizers free of use for non-commercial purposes*. Philadelphia: ICSLP.

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

Hauser, M. D., Newport, E. L., & Aslin, R. N. (2001). Segmentation of the speech stream in a non-human primate: Statistical learning in cotton-top tamarins. *Cognition*, *78*, B53–B64.

Hsiao, A., & Reber, A. (1998). The role of attention in implicit sequence learning. In M. Stadler, & P. Frensch (Eds.), *Handbook of implicit learning* (pp. 471–494). Thousand Oaks: Sage, 471–494.

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.

Kuhl, P., Tsao, F., & Liu, H. (2003). Foreign language experience in infancy: Effects of short-term exposure and social interaction on phonetic learning. *Proceedings of the National Academy of Sciences of the USA*, *100*, 9096–9101.

Lavie, N. (1995). Perceptual load as a necessary condition for selective attention. *Journal of Experimental Psychology: Human Perception and Performance*, *21*, 451–468.

Maye, J., Werker, J., & Gerken, L. A. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, *82*, B101–B111.

Newport, E., & Aslin, R. (2004). Learning at a distance: I. Statistical learning of non-adjacent dependencies. *Cognitive Psychology*, *48*, 127–162.

Rees, G., Frith, C., & Lavie, N. (1997). Modulating irrelevant motion perception by varying attentional load in an unrelated task. *Science*, *278*, 1616–1619.

Rees, G., Frith, C., & Lavie, N. (2001). Perception of irrelevant visual motion during performance of an auditory task. *Neuropsychologia*, *39*, 937–949.

Rees, G., Russell, C., Frith, C., & Driver, J. (1999). Inattentional blindness vs. inattentional amnesia for fixated but ignored words. *Science*, *286*, 2504–2507.

Saffran, J. (2001). The use of predictive dependencies in language learning. *Journal of Memory and Language*, *44*, 493–515.

Saffran, J. (2002). Constraints on statistical learning. *Journal of Memory and Language*, *47*, 172–196.

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

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

Saffran, J., Newport, E., Aslin, R., Tunick, R., & Barrueco, S. (1997). Incidental language learning: Listening (and learning) out of the corner of your ear. *Psychological Science*, *8*, 101–105.

Saffran, J., & Thiessen, E. (2003). Pattern induction by infant language learners. *Developmental Psychology*, *39*, 484–494.

Sinnett, S., Costa, A., Soto-Faraco, S. (submitted). Inattentional blindness across sensory modalities. Quarterly Journal of Experimental Psychology (A).

Snodgrass, J., & Vanderwart, M. (1980). A standardized set of 260 pictures: Norms for name agreement, image agreement, familiarity, and visual complexity. *Journal of Experimental Psychology: Human Learning and Memory*, *6*, 174–215.

Soto-Faraco, S., & Spence, C. (2002). Modality-specific auditory and visual temporal processing deficits. *Quartely Journal of Experimental Psychology (A)*, *55*, 23–40.

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

Toro, J. M. & Sebastián-Gallés, N. (2004). Are language-specific stress patterns applied to speech segmentation? Paper presented at V Congreso de la Sociedad Española de Psicología Experimental (SEPEX). Madrid, March 25th–27th.

Toro, J. M. & Trobalón, J. (in press). Statistical computations over a speech stream in a rodent. Perception & Psychophysics.

Treisman, A. M., & Davies, A. (1973). Divided attention between ear and eye. In S. Kornblum, *Attention and performance IV* (vol. 4) (pp. 101–117). New York: Academic Press, 101–117.

Vouloumanos, A., & Werker, J. (2004). Tuned to the signal: The privileged status of speech for young infants. *Developmental Science*, *7*, 270–276.

Wickens, C. (1984). Processing resources in attention. In R. Parasuraman, & D. Davies (Eds.), *Varieties of attention* (pp. 63–102). London: Academic Press, 63–102.