Speech segmentation is facilitated by visual cues

Toni Cunillera[a,b], Estela Càmara[c,d], Matti Laine[b], Antoni Rodríguez-Fornells[d,e]

[a]Department of Basic Psychology, Faculty of Psychology, University of Barcelona, 08035, Barcelona, Spain

[b]Department of Psychology, Åbo Akademi University, FIN-20500 Åbo, Finland

[c]Department of Neuropsychology, Otto-von Guericke University, 39106, Magdeburg, Germany

[d]Department of Physiology II, Faculty of Medicine, Campus de Bellvitge – IDIBELL, University of Barcelona, 08907, L'Hospitalet de Llobregat (Barcelona), Spain

[e]Institució Catalana de Recerca i Estudis Avançats (ICREA)

Toni Cunillera: tcunillera@ub.edu
Estela Càmara: ecamara@ub.edu
Matti Laine: matlaine@abo.fi
Antoni Rodríguez-Fornells: antoni.rodriguez@icrea.es

Running head: Audio-visual speech segmentation

Type of article: Full Article

Word count: 5076

Correspondence to:
Toni Cunillera
Faculty of Psychology
University of Barcelona
Passeig de la Vall d'Hebron 171, 08035,
Barcelona, SPAIN.
Phone: +34 93 312 51 45
Fax: +34 93 402 13 63
E-mail: tcunillera@ub.edu

Keywords: Speech segmentation; Temporal contiguity; Audio-visual speech; Statistical learning; Language Learning.

Acknowledgements

TC was supported by a fellowship from the University of Barcelona and the Finnish Government Scholarship Pool (HH-05-3276). ML was supported by a grant from the NEURO Research Program of the Academy of Finland. ARF was supported by research grants of the Spanish Government (MCYT) (with EC Fondos FEDER SEJ2005-06067/PSIC). The authors would like to thank the anonymous reviewers for their insightful comments.

Abstract

Evidence from infant studies indicates that language learning can be facilitated by multimodal cues. We extended this observation to adult language learning by studying the effects of simultaneous visual cues (non-associated object images) on speech segmentation performance. Our results indicate that segmentation of new words from a continuous speech stream is facilitated by simultaneous visual input that it is presented at or near syllables that exhibit the low transitional probability indicative of word boundaries. This indicates that temporal audiovisual contiguity helps in directing attention to word boundaries at the earliest stages of language learning. Off-boundary or arrhythmic picture sequences did not affect segmentation performance, suggesting that the language learning system can effectively disregard non-informative visual information. Detection of temporal contiguity between multimodal stimuli may be useful in both infants and second language learners not only for facilitating speech segmentation, but also for detecting word-object relationships in natural environments.

Introduction

Language learners are faced with different challenges when acquiring a language. At first, they must solve the speech segmentation problem by identifying word candidates in the continuous auditory stream of the new language. It has been shown that, among other cues embedded in the speech signal, both infants and adults are sensitive to the distribution of regularities of speech and can exploit this information in language learning. This important learning mechanism has been coined as statistical learning. This mechanism involves the ability to learn from different regular patterns via different sensory modalities (Conway & Christiansen, 2006). Concerning speech, statistical learning has been demonstrated to assist language learning at the phonetic level in infants (Jusczyk & Luce, 1994), as well as to be a sufficient cue for children and adults to segment word candidates in fluent speech (Aslin et al., 1998; Saffran et al., 1996a). For instance, the computation of the transitional probabilities of syllables (i.e., the likelihood of one syllable to be followed by another one) has been shown to be useful for language learners in the location of word boundaries (see for example Saffran et al., 1996a; Saffran et al., 1996b).

In language learning, however, there are often multiple cues besides transitional probabilities of syllables that can help to solve the word identification problem. In a recent review, Kuhl (2004) proposed that joining visual attention to an object that is named by an adult might help infants to segment words from ongoing speech. This idea is based on previous results (Baldwin, 1991; Baldwin, 1993; see also Tomasello & Barton, 1994) which found that 18 month-old children tend to follow the speaker's eye gaze to infer the referent of a novel word. Thus, at the earliest stages of language acquisition, a plausible strategy for learning words is to extract referents from direct visual observations of objects, scenes, or events that are guided by joining visual

attention to an object that is named by an adult (Kuhl, 2004). This is further supported by the demonstration that receptive vocabulary skills are related to an infant's tendency to follow the gaze of an adult (Baldwin, 1995; Brooks & Meltzoff, 2002).

In convergence with the idea that multiple cues could be used to enhance speech segmentation, several studies have provided evidence that infants are sensitive to audiovisual synchrony in speech (see e.g., Dodd, 1979; Gogate & Bahrick, 1998; Gogate et al., 2000). For example, infants gaze longer at a speaking face when the audio and visual sources are synchronous than when they are not, and they are sensitive to asynchronies as small as 400 ms (Dodd, 1979). Similarly, Kuhl and Meltzoff (1982; 1984) reported that 4-month-old infants gazed longer at video images that were vocally compatible with an auditory signal when compared to incompatible ones. Related to these studies, Aronson and Rosenbloom (1971) reported that 10-day-old infants showed distress when their mother's voice was heard to emanate from a location distal to her face. In a more recent study, Hollich, Newman, and Jusczyk (2005) showed that 7.5-month-olds were able to segregate speech in a noisy environment when seeing a video of the talker's face synchronized with the target passage. However, they were not able to accomplish this task when the video was unsynchronized or when there was a static face during the familiarization phase. When the synchronized face was substituted by a synchronized signal from an oscilloscope, their performance was also facilitated. The authors interpreted this finding in favor of the existence of a special sensitivity in infants to synchronized multimodal information that helps them to segregate the target speech signal from other sound sources in a noisy environment. Moreover, blind children have trouble acquiring certain phonemic distinctions (Mills, 1987), highlighting the importance of vision in language acquisition.

It thus seems that infants' word comprehension develops from the early detection of intersensory associations between auditory speech patterns (words) and visible objects or actions. Accordingly, language learning may depend on the dynamic and reciprocal interaction between intersensory perception, selective attention, and memory mechanisms. Consider, for example, the synchronous appearance of a car and the mouth movements and vocalizations of a caretaker together with the corresponding sounds: this information can be used as multimodal cues by an infant to isolate the word "car" from the continuous speech stream and ultimately to comprehend speech.

At the theoretical level, a recent model of language learning claims that infants' sensitivity to joint visual and auditory attention, together with their imitative abilities, may explain their capability to appreciate the communicative intentions of other persons (Tomasello, 2003). Temporal contiguity in the form of simultaneous appearance of an object and a word (its label) can be argued to play a central role at the early phase of infant word learning and constitutes an important element in the emergentist coalition model of word learning (Hollich et al., 2000). One of the tenets of this model is that infants' word learning relies on a perceptual subset of the available cues in the coalition, and social cues, like the eye gaze direction of others, are recruited later on during development. Thus, temporal contiguity, together with perceptual salience, would guide word learning early on in child development, followed by a shift at 12 months of age towards a greater dependency on social cues, like following adults' eye gaze or handling of objects (Golinkoff & Hirsh-Pasek, 2006).

The studies reviewed above emphasize the importance of the presence of multimodal cues that can initially guide infants' selective attention and enhance speech segmentation (e.g., Bloom, 1998; Gogate et al., 2000; Hollich et al., 2005). The use of multimodal information in language processing has also been documented in adults

learning a second language (Davis & Kim, 2001). In fact, a number of studies have demonstrated that adult listeners increase their identification rate of speech sounds when they also have access to visual information such as the dynamics of the facial articulators (Rosenblum & Saldana, 1996). In addition, in noisy environments, the intelligibility of speech increases when the speaker's face is present (Dodd, 1977; Macleod & Summerfield, 1987; Sumby & Pollack, 1954). Using non-degraded auditory information, Reisberg et al. (1987) observed that when listening to a speaker with a strong foreign accent or to a passage with a complex semantic message, seeing the speaker's lips helped language comprehension (see also Dodd, 1977; Sanders & Goodrich, 1971). In a similar vein, Thompson and Ogden (1995) reported that participants' memory of spoken sentences using native language materials was facilitated by showing the face of the speaker. The impact of "visible speech" in processing a foreign language has also been documented in adult language learners, which could serve to compensate for the weaker information accrued in the lexicon. Reisberg et al. (1987) showed that seeing the speaker's lips improved the performance of two groups of second language learners (native English speakers learning French or German). Interestingly, this effect was larger for second language learners than for native speakers. Also, in a more recent study in which adults were asked to repeat and memorize phrases of a language that they had not heard before (a foreign language), their performance improved when at the learning phase they had visual access to a video of the lower part (from the nose to the chin) of the speaker's face (Davis & Kim, 2001).

Given these findings on the impact of visible speech in first and second language processing, it is possible that visual information may also aid adults in initially segmenting the words of a new language. This hypothesis favors the idea that all available visual and auditory cues might be employed in order to better understand

speech in noisy situations or to learn a new language (e.g., Davis & Kim, 2001). Moreover, as it has been successfully implemented in a computational multisensory language interface (Yu & Ballard, 2004), it is possible that language learning could take place in an unsupervised mode with the collection of acoustic signals in concert with multisensory information from other sensory modalities, such as the speaker's eye gaze direction, head and hand movements, etc. The fundamental idea is that to acquire a language, the learner can make use of non-speech contextual information to facilitate speech segmentation.

In order to study this issue in relation to second language learning, we investigated whether adults' speech segmentation was facilitated by visual cues (images of objects) that were synchronized to the onset/offset of the words embedded in an artificial language stream (see Figure 1 for a summary of all experimental conditions). We hypothesized that speech segmentation in adults would benefit from the temporal contiguity of visual and auditory information and that this facilitation would occur even when there was no association between the novel words and object images.

Experiment 1

In this experiment, we explored whether speech segmentation is facilitated by synchronously presented visual object images. Participants were exposed to a continuous auditory speech signal composed of nonsense words. This artificial language stream is, at first, usually perceived as a long string of syllables but, after a short period of exposure, the nonsense words can be segmented from the syllable stream by computing the transitional probabilities of the syllables (Aslin et al., 1998; Saffran et al., 1996b). In the critical experimental condition, visual stimuli were added and delivered in synchrony with the word onsets and offsets in the auditory language stream (see

Figure 1). The visual information consisted of real drawings of objects, presented one at a time, with each one remaining on the screen for the entire duration of each word in the acoustic stream. The pictures were presented in a pseudorandom order, and thus they were not associated with specific words in the speech stream. In this way, we ensured that the only useful visual information provided by the object images was the temporal contiguity between words and pictures.

Methods

Participants

Fifty-two students at the University of Barcelona participated in the study. Participants were randomly assigned to one of the two conditions: auditory language stream and audio-visual language stream. All participants were native speakers of Spanish or Catalan and all of them received extra course credits for their participation.

Stimuli

Twenty-four different consonant-vowel syllables were used to create two language streams. For each stream, four trisyllabic nonsense words were concatenated to form a continuous speech stream. The acoustic streams were first created by using the speech synthesizer MBROLA (Dutoit et al., 1996) and then the duration of the streams was adjusted to a millisecond precision using the Cooledit software. The use of the artificial language learning methodology enables us to control for potential segmentation cues, such as word-stress or coarticulation. Thus, all phonemes had the same duration (116 ms) and pitch (200 Hz; equal pitch rise and fall, with pitch maximum at 50% of the phoneme) in the language streams. The only reliable cue that could help to discover word boundaries was the statistical structure of the language. In

all streams the transitional probability of the syllables forming a word was 1.0, while for syllables spanning word boundaries it was 0.33. The duration of the acoustic stream was 2 min 24 sec and 768 msec. The duration of each word was 696 msec and each one was repeated 52 times along the stream.

In addition, for each language stream 24 part-words were created by recombining the syllables of the 4 words. Thus, twelve part-words were made by concatenating the last two syllables of a word and the first one of another (part-words 2-3-1), and the other twelve were made by concatenating the last syllable of a word and the first two syllables of another (part-words 3-1-2).

Visual stimuli consisted of 2 x 4 black-and-white familiar objects (Snodgrass & Vanderwart, 1980) comparable in terms of word frequency in Spanish (on the average ~19 per million words), name agreement (~91%), imageability (mean 6.1), familiarity (mean 6.1), and concreteness (mean 5.8) (the last three variables were rated on a 1-7 scale where 7 denotes highest imageability, familiarity, or concreteness).

****
Figure 1 about here
****

Procedure

Each participant was exposed to either a single auditory speech stream or a single audio-visual speech stream delivered through headphones at a comfortable sound pressure level. All participants were instructed to listen carefully to the syllable stream and to identify novel words appearing in it. To ensure that the participants in the audio-visual condition were paying attention to both stimulus streams, they were instructed to try to associate the novel words with the pictures that were simultaneously presented on the screen.

Pictures in the audio-visual condition were presented in pseudorandom order, with the constraint that each picture appeared equally often with each word in the acoustic stream (13 times). In other words, there were no associative relationships between the pictures and the words. The visual stimuli were displayed on a white background with each picture extending for a ~3.8 x 3.8° visual angle and were presented at the center of the screen at an average viewing distance of 75 cm. Picture and word durations were equal (696 msec); that is, pictures were presented in a perfect onset-offset synchrony with words. Auditory streams and picture pools were counterbalanced across participants and conditions.

Immediately after the auditory or audio-visual stream in each experimental condition, a test phase was presented. The test consisted of a standard auditory two-alternative-forced-choice (2AFC) test. Test items were comprised of the four words of each stream and four part-words randomly selected from the pool of 24 part-words of the same stream (two part-words corresponding to the syllable structure 2-3-1 and two to the syllable structure 3-1-2; see the Stimuli section for more details). Words and part-words were exhaustively combined, rendering a total of 16 pairs presented in random order. After hearing each test item pair, the participants were asked to decide, by pressing a button corresponding to the first or the second item of the pair, which item was a word of the language stream. The presentation of the items of a pair was separated by a 400 msec pause.

It should be noted that the frequencies of words and part-words are not equated in this sort of paradigm, and words appear much more often than part-words in the language stream (the word/part-word ratio is 26/9). Nevertheless, it has been demonstrated that the same results (words selected more often than part-words in the

recognition test) are found when controlling for word and part-word frequency in the stream (e.g., Aslin et al., 1998; Graf et al., 2007).

Results and discussion

The mean percentages of correctly segmented words were as follows (see Figure 2): 68.75 ± 18.29% for the auditory condition and 85.58 ± 12.72% for the audio-visual condition. Both values were significantly different from chance (50%), p values < 0.001. Significantly more words were successfully segmented by the participants who were exposed to the audio-visual streams than by those exposed to the auditory speech streams ($t(50) = -3.85$, $p < 0.001$).

In support of the hypothesis, the participants' recognition performance showed a clear-cut beneficial effect of combined auditory and visual information on word segmentation when compared to purely auditory speech input. We hypothesize that this reflects multimodal sensory integration used in language learning. However, it might also simply result from heightened attention/motivation due to the presence of additional visual stimulation. To rule out the latter alternative, we ran an additional experiment.

****
Figure 2 about here
****

Experiment 2

We designed a new experiment in which the duration of the image exposure was varied while the same number of picture exposures as in the previous experiments was maintained. This manipulation yielded an audio-visual arrhythmic condition (see Figure

1). However, each picture along the visual stream continued to be synchronized with the onset of one syllable (either the first, the second, or the third syllable) of each word.

Methods

Participants

A new group of twenty-four native speakers of Spanish or Catalan students at the University of Barcelona participated in the study. All participants received extra course credits for their participation.

Stimuli

The audio-streams, words, part-words, pictorial stimuli, and overall setup were the same as in Experiment 1.

Procedure

The procedure was the same as in Experiment 1 except that in the present experiment each picture in the visual stream was synchronized equally often with the second syllable onset (232 msec from word onset) and with the third syllable onset (464 msec from word onset). Moreover, each picture was displayed in the visual stream for 464, 696, and 928 msec, i.e., for the duration of two, three, and four syllables, respectively (see Fig. 1). In addition, a constraint was introduced in the setting so that two consecutive pictures with the same duration were not permitted in the visual sequence. The same speech segmentation test as that used in Experiment 1 was given to participants.

Results and discussion

The mean percentage of correctly segmented words was 71.88 ± 19.2% (see Figure 2). This percentage was different from chance (50%), $p < 0.001$. The arrhythmic condition did not differ from the audio-alone condition of Experiment 1 ($t(48) = -0.6$, $p > 0.5$), but when the arrhythmic condition was compared to the synchronous condition of Experiment 1, a statistically significant difference was observed (arrhythmic vs. synchronous: $t(48) < 3.0$, $p < 0.01$).

These results rule out the possibility that the observed facilitation of speech segmentation in the audiovisual condition of Experiment 1 was due to general attentional/motivational effects of multimodal stimulation. Rather, the changing visual stimuli presumably catch attention, which helps in determining word onset/offset when coinciding with changes in transitional probabilities of syllables. It is worth noting here that the arrhythmic condition did not interfere with speech segmentation as compared to the audio-alone baseline. In other words, visual attention was not driving speech segmentation performance, but it was effective only when it provided congruent, useful information for the task at hand.

While the present results show a theoretically important audiovisual synchrony effect on speech segmentation performance, in real-life learning situations audiovisual information is not synchronized at the millisecond level. One would thus expect to find a temporal window within which coinciding auditory and visual information could facilitate speech segmentation. To explore this issue, we ran an additional experiment using a less than perfect word-picture synchrony.

Experiment 3

This audiovisual speech segmentation experiment involved a systematical displacement of the visual stream so that it was delayed from the onset of the auditory

language stream by one or two syllables (see Figure 1). Thus, novel words and pictures were no longer synchronized. On one hand, if the effect encountered in the first experiment was merely a laboratory finding that is obtained only when there is a perfect synchrony of the auditory and visual information, it should have been abolished here, as was the case with the arrhythmic audiovisual condition of the second experiment. On the other hand, if the auditory and visual information can interact within a certain temporal window, the facilitatory effect should be observed even when the visual information is somewhat displaced in time.

Methods

Participants

Another forty-eight students at the University of Barcelona participated in the study. All participants were native speakers of Spanish or Catalan and received extra course credits for their participation. Participants were randomly assigned to one of the two conditions: one-syllable asynchrony and two-syllable asynchrony.

Stimuli

The audio-streams, words, part-words, pictures, and overall setup were the same as in Experiments 1 and 2.

Procedure

The procedure was otherwise identical to Experiment 1, but in two separate conditions, the visual stimuli were synchronized with the onset of the second syllable (asynchronous-2nd-syllable: 232 msec from word onset and 464 msec from word offset) or with the onset of the third syllable (asynchronous-3rd-syllable: 464 msec from word

onset and 232 msec from word offset) of each word in the auditory stream (see also Fig. 1). Each picture remained for a constant duration of 696 msec (three syllables) along the visual stream. Finally, the same speech segmentation test was administered to participants as in Experiment 1, but with the main constraint that only one type of part-words was used in each condition. Thus, when the visual stream was synchronized with the onset of the second syllable, four 2-3-1 part-words (i.e., the ones synchronized with the visual stimuli in that condition) were exhaustively paired with words to create the test items. In the same way, in the asynchronous-3rd-syllable condition, 3-1-2 part-words were used to create the test pair items.

Results and discussion

The mean percentage of correctly segmented words for each experimental condition was as follows (see Figure 2): 73.96 ± 19.8% for the asynchrony-2nd-syllable condition and 79.95 ± 13.5% for the asynchrony-3rd-syllable condition. Both percentages were different from chance (50%), $p < 0.001$. The two conditions did not differ from each other ($t(46) < -1.22$, $p > 0.2$). The present results and the ones from Experiment 1 (the audio-alone and the synchronous conditions) were compared by a between-group one-way ANOVA. The results revealed a clear task effect ($F(3,96) = 5.13$, $p < 0.01$). Further pairwise t-tests showed that the difference between the auditory-alone and the asynchronous-3rd-syllable condition was significant ($t(48) = -2.44$, $p < 0.02$). The audio-alone condition did not differ from the asynchronous-2nd-syllable condition ($t(48) = -0.97$, $p > 0.3$). The asynchronous-3rd-syllable condition was also not different from the synchronous condition ($t(48) = -1.52$, $p > 0.1$).

These results indicate that the word-picture synchronous and asynchronous-3rd-syllable conditions led to the highest word segmentation performance. Why would the

word segmentation performance in the asynchronous-3rd-syllable condition be closer to perfect synchrony than that of the asynchronous-2nd-syllable condition? In the former condition, picture onset is closer to the next word boundary. Therefore, we suggest that picture onset synchronized with the last syllable highlights an upcoming low-probability syllable transition indicating a word boundary. In the asynchronous-2nd-syllable condition, the visual cue onset highlights the middle syllable of a possible word, an irrelevant position for detecting word boundaries. It should be noted that an opposite explanation, i.e., the auditory information capturing attention and directing it to the visual stimuli, cannot account for the present pattern of results.

In conclusion, the audiovisual facilitation effect we report does not hinge upon perfect synchrony. Instead, there appears to be a time window (of at least 232 ms in the case of the present manipulation) within which relevant cross-modal information can be integrated with speech-related segmentation cues. This is in line with studies that have investigated the effects of lip synchrony on speech recognition. Interestingly, Hashimoto and Kumashiro (2004) found that a delay up to 120 ms (corresponding to the mean duration of the mora, the Japanese equivalent of the syllable) did not disrupt the lip-reading advantage. They concluded that visual and auditory information in speech is integrated on a syllabic time scale. It has also been shown that a strict temporal synchrony between visual and the auditory speech stimuli is not necessary for the McGurk effect to occur (Munhall et al., 1996; Soto-Faraco & Alsius, in press).

General discussion

We sought to study the importance of temporal contiguity of visual information in speech segmentation. The present three experiments show that audio-visual temporal contiguity helps in segmenting words from the continuous auditory stream, but only

when the audio-visual information is synchronized with word onset/offset or when the visual information changes close to the word offset. This must be a perceptual/attentional effect, as the visual information in the audio-visual condition in Experiment 1 provided only word onset-offset cues: the images appeared in random order and had thus no relationship to the specific words.

The present pattern of results thus emphasizes the importance of temporal contiguity of visual information in speech segmentation and bears relevance to the study of how the integration of different types of information or multimodal cues facilitates language learning (Hollich et al., 2005; Hollich et al., 2000) and specifically speech segmentation. Furthermore, the present results provide the basis for a new paradigm that can be extended to study other specific aspects of multimodal language learning in perfect laboratory control settings in adults and infants.

Although infants and adults are able to track computational probabilities across syllables and are able to segment artificial speech when this statistical information is the only available segmentation cue (Saffran et al., 1996a; Saffran et al., 1996b), it is evident that in natural learning contexts, multiple and multimodal cues are used to segment real speech (Hollich et al., 2000). This corresponds well with the everyday experience when learning a new language. Thus, the temporal contiguity between auditory and visual information such as lip movements (visible speech) or a teacher's gaze to objects, pictures, or other persons in a context provides cues that facilitate speech segmentation. The present results support previous findings in which native language processing (Dodd, 1977; Reisberg et al., 1987; Sanders & Goodrich, 1971; Thompson & Ogden, 1995) or foreign language learning (Davis & Kim, 2001; Reisberg et al., 1987) was facilitated with visible speech.

The facilitatory temporal contiguity effect on speech segmentation may rely on domain-general capabilities that also benefit language learning (Bloom, 2002). The temporal contiguity of the stimuli probably acts as an attentional cue that highlights the words embedded in the speech stream. We observed this effect even though in our experiments, the visual cues were void of any associative relationship with the specific words. Importantly, the facilitation effect does not require perfect millisecond-level synchrony to appear. We observed a significant increase in the number of segmented words when the visual cue appeared together with the last syllable of each word in the audio-visual stream. This indicates that visual cues facilitate speech segmentation when the cue is near to an upcoming word boundary. This is also in line with previous studies in language learning suggesting that learners pay more attention to the end of words and benefit more from salient syllables (i.e., syllable carrying word-stress) placed at the end of words (see for example Cunillera et al., 2008; Echols & Newport, 1992; Echols, 1993; Saffran et al., 1996b).

An important aspect of the present experiments is to understand the underlying mechanism responsible for the facilitation of speech segmentation when redundant intersensory information is provided. In principle, it is possible that the attentional cue provided by the synchrony between the onset/offset of each picture and word facilitates the computation of statistical probabilities across word boundaries. Alternatively, this attentional cue could also act independently of the statistical computation process, simply helping to identify word boundaries. However, the results from the arrhythmic and asynchronous audiovisual conditions speak against an independent attentional process that bypasses statistical learning. If that were the case, interference would have been observed in the arrhythmic and asynchronous conditions because the onset-offset of the pictures would have captured incorrect syllable transitions as the onset-offset of

the words. The results depicted in Figure 2 do not show any interference in the arrhythmic or asynchrony condition as compared to the audio-alone baseline.

Our interpretation of the present results favors a model in which segmentation of continuous speech is facilitated *only* when visual and auditory information temporally coincide within a given time window that encompasses at least two hundred milliseconds (van Wassenhove et al., 2007), possibly the syllable preceding the onset of a word. When these cues do not co-occur within this time range, participants might favor a default statistical learning mode, which provides enough information to be able to isolate words from the speech stream based on transitional probabilities. Disregarding visual cues when they do not temporally match the information present in the speech signal itself might be important in everyday language learning situations. Imagine a situation where a teacher is speaking about a static object without pointing or gazing at it, or when unrelated visual cues come and go in an asynchronous fashion. In such situations, filtering out unnecessary visual information would be important in order to be able to segment the speech correctly. However, in other cases, such as lip-reading and speech, visual and auditory information tend to coincide, and therefore, the system would benefit from the temporal contiguity between both visual and auditory cues (Rosenblum & Saldana, 1996). As we have seen, this strategy is in fact used in infant-directed communication (i.e., "multimodal motherese", Gogate et al., 2000), in which mothers provide multimodal redundant information. In the same way, if one wanted to teach someone a new word for a visual object and that object was present, one would most likely point to it when pronouncing its name. Interestingly, it has even been found that the head movements that naturally co-occur with speech improve auditory speech perception (Munhall et al., 2004).

It is also important to consider that intersensory temporal synchrony does not require the different audio-visual components to occur at the same instant in time, as the perceptual system tolerates a certain amount of temporal discrepancy. This audio-visual discrepancy window is larger when a visual stimulus is presented before an auditory one (112 ms) than in the reverse presentation, audio-visual order (65 ms) (Lewkowicz, 1996). These temporal synchrony windows are very similar to the ones obtained by McGrath and Summerfield (1985). These authors presented adult participants with liplike figures that mimicked the opening of lips with a tone that appeared either before or after the opening of the lips, with systematically varied intervals. Asynchrony was detected only when the auditory event preceded the visual event by about 79 ms, while for the reverse (visual-auditory) order, the integration window was about 138 ms (see also Dixon and Spitz, 1980).

The reason for these differences in the intersensory temporal synchrony window has to be related to the faster processing of auditory information in the central nervous system. Thus, if the auditory information is presented faster than the visual information, the temporal window that allows the creation of a unified perceptual experience is reduced when compared to visual-audio presentations. Interestingly, Lewkowicz (1996) has shown that the size of the audio-visual temporal synchrony window is larger in infants than in adults (audio-visual: 350 ms; visual-audio: 450 ms). These differences might reflect infants' inexperience with temporal discriminations and their slower rate of transmission of information in the nervous system (Lewkowicz, 1996). However, the advantage for infants of having a larger time synchrony window is that it could facilitate the identification of certain relationships in the environment that are presented temporally closely but with a certain degree of asynchrony. For example, Gogate et al. (2000) have shown that mothers of prelexical infants (between 5 to 8 months old) tend

to use multimodal communication styles to teach labels for novel objects and actions (see also Zukow-Goldring, 1997). In particular, named labels are produced in synchrony with the actions exerted on the objects. This synchrony might help infants to detect and infer word-referent relations. Curiously enough, this tendency by mothers to use multimodal communication styles is reduced when the infant becomes older (e.g., at 21-30 months).

These lines of evidence suggest that infants might take advantage of a larger intersensory temporal window when compared to adults, which likely provides more capacity to unify perceptually relevant elements in the context. For example, it has been shown that, between two-and-a-half and four months of age, infants attended more to synchronous than asynchronous visible lip movements and audible speech patterns (Dodd, 1979). In addition, at 4 months, infants are able to recognize the correspondence between the sight of a bouncing object and a sound (Spelke, 1979). In line with this, the Intersensory Redundancy Hypothesis (IRH, Bahrick and Lickliter, 2000; 2002) claims that overlapping of information provided by different senses helps in focusing attention on critical aspects of the environment. The redundancy, which includes synchrony, rhythm, tempo, etc. across more than one sensory modality, is considered to be an advantage for the perceptual processing system involved in learning. It is also a possible cornerstone of perceptual development, allowing learners to selectively attend to related aspects of the multimodal information found in the input that represent unitary events, and at the same time, ignore the information from unrelated events nearby (Bahrick & Lickliter, 2002; Gibson & Pick, 2000). Similarly, it is easy to conceive that mothers, who use a slower tempo when speaking to their infants ("motherese" or "infant-directed speech") (Fernald & Simon, 1984), might provide supportive multimodal information (e.g., visible speech, gestures) that could be integrated in wider time-synchrony

windows. This multimodal information might help the process of identifying the boundaries of the words and, ultimately, the speech segmentation process. A similar idea has been proposed by Hollich et al. (2005) in order to explain their results of better speech segmentation in 7.5-month-old infants when synchronous visual information was provided. The authors suggest that infants might tolerate larger temporal asynchronies compared to adults, increasing their capacity to segregate speech streams and segment speech especially in noisy environments.

Thus, the infant's initial sensitivity to multimodal information provides an economical way of guiding perceptual processing to focus on meaningful, unitary events. It would be reasonable to assume that the advantage of exploiting crossmodal redundant information is preserved throughout the life span and that an adult acquiring a second language might be able to exploit the redundancy of information found in the learning environment to boost the perception of unitary speech units.

Finally, a variable that we manipulated in the present study was the visual rhythmicity. It is important to note that the auditory stream we applied provided no rhythmic properties besides the syllabic pattern, as the speech streams were synthesized with a constant syllabic length and a flat stress-pattern. Rhythm is considered by linguistics to be paramount for distinguishing one family of languages from another (Pike, 1945; Abercrombie, 1967; Ramus & Mehler, 1999), with Spanish, like most of the Romance languages, being classified as a syllabic-timed language. Rhythmicity is probably the first source of information that language learners can detect in the speech signal, as newborns are able to discriminate their native language from a non-native one when the two languages belong to different rhythmic classes (Mehler et al., 1988). In addition, Ramus and coworkers (2000) found that Cotton-top tamarin monkeys were able to discriminate continuous speech from two rhythmically distinct languages (see

also Tincoff et al., 2005), indicating that rhythmicity is a general property of the auditory signal detected by, at least, the mammalian auditory system.

However, the exact properties that correspond to the perception of rhythm in the speech signal are still not well understood. Rhythm could emerge from the succession of syllables, vowels, stress patterns, pitch features, or any repeated perceptual changes detected in the speech input (Ramus & Mehler, 1999; Ramus et al., 1999). In spite of the lack of a coherent definition of speech rhythmicity, several studies have explored the role of rhythm in speech segmentation. The underlying idea is that rhythm might aid in acquiring some phonological properties and that speakers of different languages might use different segmentation units, with rhythm being the cue that guides infants to select the proper unit (Cutler et al., 1986; Otake et al., 1993). Other studies have shown that rhythmicity provides critical cues for segmenting utterances into constituents such as clauses, phrases, and words. For example, infants at the age of 6 to 7 months can exploit overall rhythm to predict clause and phrase boundaries (Hirsh-Pasek et al., 1987), and at the age of 9 months, infants coordinate the statistical and the rhythmic structures of speech input to identify possible "word-like" multisyllabic rhythmic units (Morgan & Saffran, 1995). It seems that infants evolve from detecting large rhythmic linguistic units as clauses to finally achieving the detection of multisyllabic words, the smallest meaningful rhythmic units. This might be possible due to the progressive development of a more flexible attentional system that enables faster changes of attentional allocation.

It is evident that, in real-life word learning, multimodal cues do not appear in perfect (millisecond) synchrony. Moreover, it is possible to perceive synchronicity even if a perfect multimodal synchrony is lacking, as has been demonstrated in studies on the phenomena coined as perceptual centres (P-centres; Morton et al., 1976). The P-centres

are subjective moments of occurrence based on properties of regularity and synchrony found in production and perception (Scott, 1998). Thus, the present setup does not claim ecological validity but was rather designed to test the role of multimodal perception in speech segmentation in a strictly controlled situation. The demonstration that the beneficial effect of audiovisual synchrony exists also in the absence of perfect synchrony and cannot explained by more general attentional/motivational factors paves the way to research in the context of more natural language learning. More studies are also needed to further specify the time window for integration of auditory and visual information in speech segmentation. For instance, an interesting experiment would be one in which the auditory stream is accompanied by visual lip movement (visible speech) pronouncing the same syllabic stream.

In summary, temporal contiguity of intersensory information will probably sharpen and tune the efficacy of the underlying learning mechanisms, in this case the statistical learning process. Furthermore, the detection of temporal synchrony might also be very useful for both infants and second language learners not only to increase speech segmentation but also to detect word-object relations in natural environments. Our results support the importance of visual cues in language learning and, in line with the emergentist coalition model of infants' acquisition of native language (Hollich et al., 2000), also emphasize the potential importance of temporal contiguity at the early phase of second language learning in adults.

References

Abercrombie, D. (1967). *Elements of general phonetics*, Chicago: Aldine.

Aronson, E. & Rosenbloom, S. (1971). Space perception in early infancy - Perception within common auditory-visual space. *Science, 172,* 1161-1163.

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

Bahrick, L. E. & Lickliter, R. (2000). Intersensory redundancy guides attentional selectivity and perceptual learning in infancy. *Developmental Psychology, 36,* 190-201.

Bahrick, L. E. & Lickliter, R. (2002). Intersensory redundancy guides early perceptual and cognitive development. In R. Kail (Eds.), Advances *in child development and behavior* (pp.153-187). New York: Academic Press.

Baldwin, D. A. (1991). Infants' contribution to the achievement of joint reference. *Child Development, 62,* 875-890.

Baldwin, D. A. (1993). Infants' ability to consult the speaker for clues to word reference. *Journal of Child Language, 20,* 395-418.

Baldwin, D. A. (1995). Understanding the link between joint attention and language. In C.Moore & P. J. Dunham (Eds.), *Joint Attention: Its Origins and Role in Development* (pp. 131-158). Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Bloom, L. (1998). Language acquisition in its developmental context. In W.Damon, D. Kuhn, & R. S. Siegler (Eds.), *Handbook of child psychology* (pp. 309-370). New York: Wiley & Sons.

Bloom, P. (2002). Mindreading, communication and the learning of names for things. *Mind & Language, 17,* 37-54.

Brooks, R. & Meltzoff, A. N. (2002). The importance of eyes: How infants interpret adult looking behavior. *Developmental Psychology, 38,* 958-966.

Conway, C. M. & Christiansen, M. H. (2006). Statistical learning within and between modalities - Pitting abstract against stimulus-specific representations. *Psychological Science, 17,* 905-912.

Cunillera, T., Gomila, A., & Rodriguez-Fornells, A. (2008). Beneficial effects of word final stress in segmenting a new language: evidence from ERPs. *BMC Neuroscience, 9,* 23.

Cutler, A., Mehler, J., Norris, D., & Segui, J. (1986). The syllable's differing role in the segmentation of French and English. *Journal of Memory and Language, 25,* 385-400.

Davis, C. & Kim, J. (2001). Repeating and remembering foreign language words: Implications for language teaching systems. *Artificial Intelligence Review, 16,* 37-47.

Dixon, N. F. & Spitz, L. (1980). The detection of auditory visual desynchrony. *Perception, 9,* 719-721.

Dodd, B. (1977). The role of vision in the perception of speech. *Perception, 6,* 31-40.

Dodd, B. (1979). Lip reading in infants - Attention to speech presented in-synchrony and out-of-synchrony. *Cognitive Psychology, 11,* 478-484.

Dutoit, T., Pagel, N., Pierret, F., Bataille, O., & van der Vreken, O. (1996). The MBROLA project: Towards a set of high-quality speech synthesizers free of use for non-commercial purposes. In (pp. 1393-1396). Philadelphia.

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

Echols, C. H. & Newport, E. L. (1992). The role of stress and position in determining first words. *Language Adquisition, 2,* 189-220.

Fernald, A. & Simon, T. (1984). Expanded intonation contours in mothers' speech to newborns. *Developmental Psychology, 20,* 104-113.

Gibson, E. J. & Pick, A. D. (2000). *An ecological approach to perceptual learning and development*. New York: Oxford University Press.

Gogate, L. J. & Bahrick, L. E. (1998). Intersensory redundancy facilitates learning of arbitrary relations between vowel sounds and objects in seven-month-old infants. *Journal of Experimental Child Psychology, 69,* 133-149.

Gogate, L. J., Bahrick, L. E., & Watson, J. D. (2000). A study of multimodal motherese: The role of temporal synchrony between verbal labels and gestures. *Child Development, 71,* 878-894.

Golinkoff, R. M. & Hirsh-Pasek, K. (2006). Baby wordsmith - From associationist to social sophisticate. *Current Directions in Psychological Science, 15,* 30-33.

Graf, E. K., Evans, J. L., Alibali, M. W., & Saffran, J. R. (2007). Can infants map meaning to newly segmented words? Statistical segmentation and word learning. *Psychological Science, 18,* 254-260.

Hashimoto, M. & Kumashiro, M. (2004). [Intermodal timing cues for audio-visual speech recognition]. *J.UOEH., 26,* 215-225.

Hirsh-Pasek, K., Kemler-Nelson, D. G., Jusczyk P. W., Cassidy, K. W., Druss, B., & Kennedy, L. (1987). Clauses are perceptual units for young infants. *Cognition, 26,* 269-286.

Hollich, G., Newman, R. S., & Jusczyk, P. W. (2005). Infants' use of synchronized visual information to separate streams of speech. *Child Development, 76,* 598-613.

Hollich, G. J., Hirsh-Pasek, K., Golinkoff, R. M., Brand, R. J., Brown, E., Chung, H. L. et al. (2000). Breaking the language barrier: an emergentist coalition model for the origins of word learning. *Monogr Soc.Res.Child Dev., 65,* i-123.

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

Kuhl, P. K. (2004). Early language acquisition: Cracking the speech code. *Nature Reviews Neuroscience, 5,* 831-843.

Kuhl, P. K. & Meltzoff, A. N. (1982). The bimodal perception of speech in infancy. *Science, 218,* 1138-1141.

Kuhl, P. K. & Meltzoff, A. N. (1984). The intermodal representation of speech in infants. *Infant Behavior & Development, 7,* 361-381.

Lewkowicz, D. J. (1996). Perception of auditory-visual temporal synchrony in human infants. *Journal of Experimental Psychology-Human Perception and Performance, 22,* 1094-1106.

Macleod, A. & Summerfield, Q. (1987). Quantifying the contribution of vision to speech perception in noise. *British Journal of Audiology, 21,* 131-141.

McGrath, M. & Summerfield, Q. (1985). Intermodal timing relations and audio-visual speech recognition by normal-hearing adults. *Journal of the Acoustical Society of America, 77,* 678-685.

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

Mills, A. E. (1987). The development of phonology in the blind child. In B.Dodd & R. Campbell (Eds.), *Hearing by eye: The psychology of lip reading* (pp. 145-162). Hillsdale, NJ: Lawrence Erlbaum Associates.

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

Morton, J., Marcus, S. M., & Frankish, C. (1976). Perceptual centres (P-centres). *Psychological Review, 83,* 405-408.

Munhall, K. G., Gribble, P., Sacco, L., & Ward, M. (1996). Temporal constraints on the McGurk effect. *Perception & Psychophysics, 58,* 351-362.

Munhall, K. G., Jones, J. A., Callan, D. E., Kuratate, T., & Vatikiotis-Bateson, E. (2004). Visual prosody and speech intelligibility: head movement improves auditory speech perception. *Psychological Science, 15,* 133-137.

Otake, T., Hatano, G., Cutler, A., & Mehler, J. (1993). Mora or syllable? Speech segmentation in Japanese. *Journal of Memory and Language, 32*, 258-278.

Pike, K. L. (1945). *The intonation of American English*, Ann Arbor, MI: University of Michigan Press.

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

Ramus, F. & Mehler, J. (1999). Language identification with suprasegmental cues: A study based on speech resynthesis. *Journal of the Acoustical Society of America, 105*, 512-521.

Ramus, F., Nespor, M., & Mehler, J. (1999). Correlates of linguistic rhythm in the speech signal. *Cognition, 73*, 265-292.

Reisberg, D., McLean, J., & Goldfield, A. (1987). Easy to hear but hard to understand: A lip-reading advantage with intact auditory stimuli. In B. Dodd & R. Campbell (Eds.), *Hearing by eye: The psychology of lip-reading* (pp. 97-114). Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Rosenblum, L. D. & Saldana, H. M. (1996). An audiovisual test of kinematic primitives for visual speech perception. *Journal of Experimental Psychology-Human Perception and Performance, 22*, 318-331.

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

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

Sanders, D. A. & Goodrich, S. J. (1971). The relative contribution of visual and auditory components of speech to speech intelligibility as a function of three conditions of frequency distortion. *Journal of Speech, Language, and Hearing Research, 14*, 154-159.

Scott, S. K. (1998). The point of P-centres. *Psychological Research, 61*, 4-11.

Snodgrass, J. G. & Vanderwart, M. (1980). 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, & Alsius, A. Deconstructing the McGurk-MacDonald Illusion. *Journal of Experimental Psychology-Human Perception and Performance, in press.*

Spelke, E. S. (1979). Perceiving bimodally specified events in infancy. *Developmental Psychology, 15*, 626-636.

Sumby, W. H. & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America, 26,* 212-215.

Thompson, L. A. & Ogden, W. C. (1995). Visible speech improves human language understanding: Implications for speech processing systems. *Artificial Intelligence Review, 9,* 347-358.

Tincoff, R., Hauser, M., Tsao, F., Spaepen, G., Ramus, F., & Mehler, J. (2005). The role of speech rhythm in language discrimination: further test with a non-human primate. *Developmental Science, 8,* 26-35.

Tomasello, M. (2003). *Contructing a Language*. Cambridge, Massachusetts: Harvard University Press.

Tomasello, M. & Barton, M. (1994). Learning words in non-ostensive contexts. *Developmental Psychology, 30,* 639-650.

van Wassenhove, V., Grant, K. W., & Poeppel, D. (2007). Temporal window of integration in auditory-visual speech perception. *Neuropsychologia, 45,* 598-607.

Yu, C. & Ballard, D. H. (2005). A multimodal learning interface for grounding spoken language in sensory perceptions. *ACM Transactions on Applied Perceptions, 1,* 57–80.

Zukow-Goldring, P. (1997). A social ecological realist approach to the emergence of the lexicon: educating attention to amodal invariants in gesture and speech. In C. Dent-Read & P. Zukow-Goldring (Eds), *Evolving explanations of development: ecological approaches to organism-environment systems* (pp. 199-252). Washington, DC: American Psychological Association.

Figure captions

Figure 1. Illustration of the procedure used for language exposure in the different experimental conditions. In all conditions, auditory information was presented (the uppermost row shows an auditory stream composed by four words). In the first experiment, the auditory alone condition (*audio*) was compared to the audiovisual synchrony condition (*synchronous*). In the latter condition, the onset of the picture perfectly matches the onset of each word. In the second experiment, the duration of each picture varied and each picture was synchronized with the first, second and the third syllable of an auditory word (*arrhythmic*). The right-hand column depicts the mean percentage of correct segmented words in each condition.

Figure 2. Distributions of the percentages of correctly segmented nonsense words in the auditory 2AFC test administered after the auditory and audio-visual conditions (Experiment 1: *Audio alone* condition vs. *synchronous* condition; Experiment 2: *Arrhythmic condition*). Each point corresponds to an individual participant score and stars denote the mean values for each condition. All conditions were significantly different from chance (50%, all p's < 0.001).

Appendix 1. The artificial languages used in the different conditions.

Language 1:

Words: PIRUTA, BAGOLI, TOKUDA, DEMUSI

Part-words 3-1-2: RUTABA, RUTATO, RUTADE, GOLIPI, GOLITO, GOLIDE, KUDAPI, KUDABA, KUDADE, MUSIPI, MUSIBA, MUSITO

Part-words 2-3-1: LIPIRU, DAPIRU, SIPIRU, TABAGO, DABAGO, SIBAGO, TATOKU, LITOKU, SITOKU, TADEMU, LIDEMU, DADEMU

Language 2:

Words: PABELA, DINEKA, KOTUSA, JISODU

Part-words 3-1-2: BELADI, BELAKO, BELAJI, NEKAPA, NEKAKO, NEKAJI, TUSAPA, TUSADI, TUSAJI, SODUPA, SODUDI, SODUKO

Part-words 2-3-1: KAPABE, SAPABE, DUPABE, LADINE, SADINE, DUDINE, LAKOTU, KAKOTU, DUKOTU, LAJISO, KAJISO, SAJISO

Pictures:

Pool 1: BEAR, SKIRT, ORANGE, GLASSES

Pool 2: FLAG, SWEATER, LEAF, BOOT