Exp Brain Res
DOI 10.1007/s00221-012-3390-3

RESEARCH ARTICLE

Cross-modal prediction in speech depends on prior linguistic experience

Carolina Sánchez-García · James T. Enns · Salvador Soto-Faraco

Received: 24 July 2012 / Accepted: 18 December 2012
© Springer-Verlag Berlin Heidelberg 2013

Abstract The sight of a speaker's facial movements during the perception of a spoken message can benefit speech processing through online predictive mechanisms. Recent evidence suggests that these predictive mechanisms can operate across sensory modalities, that is, vision and audition. However, to date, behavioral and electrophysiological demonstrations of cross-modal prediction in speech have considered only the speaker's native language. Here, we address a question of current debate, namely whether the level of representation involved in cross-modal prediction is phonological or pre-phonological. We do this by testing participants in an unfamiliar language. If cross-modal prediction is predominantly based on phonological representations tuned to the phonemic categories of the native language of the listener, then it should be more effective in the listener's native language than in an unfamiliar one. We tested Spanish and English native speakers in an audiovisual matching paradigm that allowed us to evaluate visual-to-auditory prediction, using sentences in the participant's native language and in an unfamiliar language. The benefits of cross-modal prediction were only seen in the native language, regardless of the particular language or participant's linguistic background. This pattern of results implies that cross-modal visual-to-auditory prediction during speech processing makes strong use of phonological representations, rather than low-level spatiotemporal correlations across facial movements and sounds.

Keywords Audiovisual speech · Speech perception · Predictive coding · Multisensory integration

C. Sánchez-García · S. Soto-Faraco (✉)
Departament de Tecnologies de la Informació i les Comunicacions, Center of Brain and Cognition (CBC), Universitat Pompeu Fabra, c/Roc Boronat 138, 08018 Barcelona, Spain
e-mail: Salvador.Soto@icrea.cat

J. T. Enns
Department of Psychology, University of British Columbia, Vancouver, BC, Canada

S. Soto-Faraco
Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain

Introduction

Speech perception is a multisensory process involving auditory (i.e., linguistic sounds) and visual information (i.e., lip movements) resulting from the speaker's articulations. According to several authors, the listener's perceptual system extracts spatiotemporal correlations across vision and audition and integrates them in order to decode the spoken message more effectively (e.g., Calvert et al. 2004; Arnold and Hill 2001; Schwartz et al. 2004; Grant and Seitz 2000; Grant 2001). Perhaps, the most dramatic demonstration of the role of vision in speech perception is the McGurk effect (McGurk and MacDonald 1976), whereby a spoken syllable (i.e., /ba/) presented auditorily and dubbed onto a videoclip of a speaker articulating an incongruent syllable (i.e., /ga/) often results in the acoustic perception of a fused, in-between, syllable (in this case, /da/). However, unlike the McGurk effect, where visual and auditory information are set in artificial conflict, the normal consequence of seeing articulatory gestures together with their corresponding speech sounds is beneficial, resulting in a comprehension enhancement of speech in noise (Sumby and Pollack 1954) or improvements in discriminating phonemes from a second language (Navarra and Soto-Faraco 2007). Although the exact mechanisms underlying these enhancements are

Published online: 06 February 2013

still unknown, it has been suggested that visual speech information can be used to constrain the perception of upcoming speech sounds via predictive coding mechanisms (van Wassenhove et al. 2005; Skipper et al. 2005; Pickering and Garrod 2006; Sanchez-García et al. 2011).

Predictive coding mechanisms have been shown to operate during speech perception as well as in many other domains (see Bubic et al. 2010 for a review). The general idea underlying these models is that sensory information in the brain flows in a forward fashion that, at one or more stages, is compared with top–down "predictions," projected back from higher levels of information processing. These feedback predictions help to reduce ambiguity among potential interpretations of sensory input, refining and enhancing perception.

The question addressed here is whether, during speech processing, these mechanisms operate primarily at a pre-phonological level of representation, capitalizing on correlations between spatiotemporal dynamics of speech across sensory modalities, or whether they also operate at the phonological level. By phonological level, we mean an abstract level of speech representation, where acoustic input has already been organized around categories tuned to the phonological system of the listener's native language. According to this, speakers of different languages would organize the input according to different sound categories, relevant to their own linguistic experience. There is ample evidence that these phonological categories act as a sieve and have a major impact on the initial parsing of the speech input (e.g., Pallier et al. 2001; Sebastian-Galles and Soto-Faraco 1999; Navarra et al. 2005). For instance, perception and categorization of non-native language contrasts seem to be tuned to the phonemic categories of the native language of the listener, even after a long exposure to, and despite high proficiency levels of, the second language (Sebastian-Galles and Soto-Faraco 1999; Navarra et al. 2005). Furthermore, the lack of sensitivity in discriminating non-native phonemic contrasts which are not present in the native language extends to the perception of second-language words. Lexical representations are based on abstract phonological language-specific representations, established at an early age, and not easily modifiable later on (Pallier et al. 2001).

Facial articulatory movements are often expressed in advance of their acoustic correlates in natural speech (Chandrasekaran et al. 2009). This offers the possibility of predictive processing, whereby leading visual information allows perceivers to extract reliable information from the sight of the speaker's face and use it to anticipate the incoming auditory input, speeding up speech processing. Accordingly, van Wassenhove et al. (2005) reported a significant speed up of the auditory-evoked potential components N1 and P2 when syllables were presented audiovisually versus only auditorily. Interestingly, the size of this latency shift in the auditory-evoked components was proportional to the visual saliency of the phoneme. Van Wassenhove et al. interpreted this finding to suggest that the extraction of phonological information from the orofacial articulators allowed perceivers to create expectations about what was going to be heard, thereby speeding up processing when the input matched the expectation. But other levels of speech processing are also possibly relevant by the time the expectation has been created. For example, predictive mechanisms could rely on simpler spatiotemporal correlations of audiovisual information, even before any phonological information is extracted, as suggested by studies using non-speech stimuli (Stekelenburg and Vroomen 2007) and artificial non-human-related audiovisual events (Vroomen and Stekelenburg 2009). In these studies, Stekelenburg and Vroomen showed that temporal anticipation of the visual cues in non-speech stimuli was sufficient to elicit an amplitude modulation of the auditory-evoked potentials in the audiovisual condition compared to the auditory events presented alone. Moreover, while presenting artificial audiovisual events, the amplitude reduction of the N1 component depended critically on the predictability of the visual cues with respect to the subsequent auditory input. Interestingly, the N1 effect seems to be dependent on the anticipation of the visual information and independent of the audiovisual congruency of the stimulus with regard to content (see also Arnal et al. 2009). This suggests that visuoauditory facilitatory processes can be based on pre-phonological information. In contrast with the N1 component, the P2 auditory component appears to be dependent on the audiovisual congruency of the information, and regarding this later component, a dissociation between a more central P2 effect in non-speech events and a more occipitotemporal distribution in speech has been observed (Stekelenburg and Vroomen 2007). According to Stekelenburg and Vroomen, this dissociation in brain signals might map onto differences between the semantic and phonetic levels, pointing to a relevant role of phonetic representations during cross-modal facilitation.

Further support for cross-modal interactions at a variety of levels of representation comes from brain imaging studies of the anatomo-functional networks that are active during speech perception. For instance, Arnal et al. (2009) proposed that when information between sensory modalities matches, its processing engages preferentially direct connections between visual and auditory areas. In contrast, mismatching information across modalities engages a slower, more indirect network, whereby visual input is integrated and compared with the auditory input via association areas (i.e., the superior temporal sulcus, STS). Following up on this idea and using MEG, Arnal et al. (2011) reported a change in oscillatory brain activity so that when audiovisual events were congruent, predictions were mediated by slow oscillations in high-order speech areas, whereas in the case of incongruent visual and auditory inputs, they were mediated by high-frequency oscillations in early unisensory cortex and the STS. In

line with this idea, Sohoglu et al. (2012) suggested that in the presence of prior written text matching with the subsequent auditory input, abstract phonological predictions are created in the inferior frontal gyrus (IFG), and following a top–down process, this information is conveyed later to the superior temporal gyrus (STG), in the form of acoustic–phonetic predictions. These predictions are compared with neural representations of the incoming speech input. Interestingly, when the prior context was highly predictable of the subsequent input, activity on the STG was reduced, because top–down predictions by themselves were already in agreement with the sensory input. In related work, Schroeder et al. (2008) argued for a modulatory effect of visual input on the auditory cortex, specifically to reset the phase of neural oscillations that would go through a high excitability phase, thus allowing for amplification of the auditory signal when it arrives at the right time. This mechanism is claimed to be responsible for temporal parsing of audiovisual signals, actively driven by a cross-modal modulation of low-frequency neuronal information in early sensory areas during speech perception (Luo et al. 2010).

In summary, previous studies have highlighted the potential for cross-modal facilitation operating at several processing levels. Here, we are interested in addressing whether cross-modal facilitation in speech processing primarily exploits low-level pre-phonological representations, or whether it also occurs at phonological levels. In order to explore this question, we compare the cross-modal transfer of information in native and non-native language perception.

Previous studies looking at prediction in speech perception within a single modality have often used semantically biasing contexts, so that expectations are based on narrowing semantics and prediction is expressed at the levels of word form and lexical entry (deLong et al. 2005; Van Berkum et al. 2005; Dambacher et al. 2009). When it comes to prediction across sensory modalities, some recent evidence suggests that cross-modal predictive processes may be rooted in pre-semantic stages of input analysis instead. For example, Sanchez-García et al. (2011) showed that a visually presented speech context produced an audiovisual processing benefit when the upcoming auditory channel was a continuation of the prior visual speech context. This occurred even though this visual-only context was not semantically biasing, and it conveyed little lexical or semantic information on its own (Bernstein et al. 1998; Soto-Faraco et al. 2007; Altieri et al. 2011). In their study, Sánchez-García et al. used a speeded audiovisual correspondence judgement on a speech fragment to test whether cross-modal transfer of information could benefit the perception of speech during an online task. The audiovisual speech target fragment followed either a visual or an auditory (unimodal) sentence context. This audiovisual target could be a continuation of the unisensory sentence context, either in the same sensory modality (intra-modal) or in the opposite modality (cross-modal). Otherwise, the audiovisual target was completely discontinuous with both modalities of the sentence context. The authors found that participants were able to speed up their response times to the target fragment when a visually presented leading sentence fragment continued into the auditory modality (cross-modally), meaning that participants were able to extract information from visual speech and use it to constrain the ensuing auditory speech fragment, speeding up their audiovisual matching judgments. Interestingly, the cross-modal effect was not found when the leading fragment was presented in the auditory modality. Along similar lines, results from electrophysiological studies report that visual-based prediction not only modulates the early stages of input analysis, such as the auditory-evoked components N1 and P2 (van Wassenhove et al. 2005; Stekelenburg and Vroomen 2007; Besle et al. 2004; Luo et al. 2010; Schroeder et al. 2008), but that it also depends critically on the visual saliency of the articulatory features involved (van Wassenhove et al. 2005).

It is perhaps to be expected that predictive coding across sensory modalities in speech operates at pre-semantic levels of input analysis, given the potential for cross-modal transfer to occur at early levels of representation (see Sanchez-García et al. 2011 for more detailed discussion). However, within this early stage, the distinction between the role of phonological and pre-phonological levels of processing is currently the focus of much debate among theories of audiovisual speech integration (Schwartz et al. 2004; Bernstein 2005). Some authors argue for the importance of a pre-phonological level of representation common to both modalities, prior to phonemic categorization (Summerfield 1987; Schwartz et al. 1998; Rosenblum 2005). The data supporting this view include the enhancement of speech detection in noise from a visual cue related to the auditory signal (Kim and Davis 2003; Bernstein et al. 2004), and the finding that auditory comprehension is improved by temporally correlated visual articulations, even when they do not match perfectly in phonemic content (Schwartz et al. 2004). Alternative claims, in support of a phonological level of processing, argue instead that audiovisual integration follows phonemic identification (i.e., Massaro and Cohen 1983; Massaro 1998). A strong motivation for this view is that the phonological level has been proposed as a convenient common representational code for speech perception (both auditory and visual) and production (Pickering and Garrod 2006; Skipper et al. 2005, 2007; Fowler 2004).

The present study sought to distinguish whether the input analysis contributing to the cross-modal transfer benefit is predominantly phonological or pre-phonological. According to pre-phonological fusion theories (Summerfield 1987; Schwartz et al. 1998), online transfer of visual-to-auditory information can occur based on the spatiotemporal correlations between visual (articulatory) and auditory (acoustic)

inputs, thus before (or at least, independent of) phonological categorization. Temporal anticipation of visual information, characteristic of natural speech, when the articulated utterance contains highly salient visual information, would make it possible to extract cues that would constrain the processing of subsequent information in the auditory modality. According to this account, cross-modal prediction should occur during the perception of speech in any input language. Indirect evidence in support of this hypothesis is that visual-to-auditory prediction can be possibly based solely on temporal cues and occurs independently on the audiovisual congruency of the stimuli (Stekelenburg and Vroomen 2007; Vroomen and Stekelenburg 2009; Arnal et al. 2009). On the other hand, if cross-modal prediction is based on phonological representations and is therefore mediated by the repertoire of phonological categories that are specific to each language, then one would expect stronger cross-modal transfer for a listener's native language, as compared to unfamiliar languages.

To address this question, we tested Spanish and English speakers in their native language, respectively, and compared it with Spanish, English and German native speakers tested in a non-native language (English for Spanish speakers, and Spanish for English and German speakers), in an audiovisual matching task (Sanchez-García et al. 2011). Participants made speeded judgments about whether or not the voice and the lips of the speaker in an audiovisual speech target fragment were in agreement, following a leading visual-only sentence context. Prior visual context could be continuous with the visual modality in the audiovisual target fragment (i.e., intra-modally continuous), continuous with the auditory modality in the audiovisual target fragment (i.e., cross-modally continuous) or with neither of the two modalities in the target fragment (i.e., discontinuous). When participants experienced their native language, we expected that the linguistic continuity from the previous visual context would benefit audiovisual processing of the target fragment, as previously found (Sanchez-García et al. 2011). The critical question here was whether language familiarity would have an additional influence on this benefit. If online cross-modal transfer is mediated by a phonological level of representation, over and above lower levels, then the native language should yield greater benefits in the audiovisual matching task. Alternatively, if cross-modal transfer is supported only by pre-phonological representations, then there should be similar benefits for familiar and unfamiliar languages.

Methods

Participants

Eighteen native Spanish speakers (6 males, mean age 21 years) and 10 native English speakers (3 males, mean age 21.2 years) were tested in their native language. Data from 9 Spanish and 8 English additional participants who failed to meet a performance criterion of 65 % accuracy in the task were not included in the reported analysis (although separate analysis showed that the exclusion of their data did not alter any of our conclusions). A different group of 19 Spanish native speakers (with low to medium knowledge of English language; 7 males, mean age 22.9 years), a group of eight English native speakers (1 male, mean age 20.33 years) and a group of nine German native speakers (2 males, mean age 24.1 years), both without knowledge of Spanish, were tested in their non-native language (English for the Spanish group, and Spanish for the English and German groups). Additional data from 12 Spanish, ten English and 14 German participants were excluded, following the same criterion as above and with the same consequences for the reported analysis.

All participants reported normal audition and normal or corrected-to-normal vision and were naive to the purpose of the experiment. All of them were selected after filling out a questionnaire to assess their experience with the languages being tested (Marian et al. 2007; see Table 1).

Materials and procedure

The stimuli were made of high-resolution audiovisual recordings of a Spanish–English bilingual male speaker uttering forty complete sentences in each language. See "Appendix". Each sentence was edited with Adobe Premiere Pro 1.5, to last 2,400, 2,600, 2,800 and 3,000 ms, including a 560 ms linear fade-in ramp and a 360 ms linear fade-out ramp. Participants viewed the video recordings from a distance of 60 cm on a 17″ CRT computer monitor that displayed a full view of the face of the speaker at the center of the screen. The audio stream was played through two loudspeakers located on each side of the monitor, at an intensity of 65 dB (A) SPL. DMDX software (Forster and Forster 2003) was used to organize the randomization, presentation and timing of the experiments.

Each trial began with a central fixation point (a circle subtending 0.8° of visual angle, 500 ms duration), followed

Table 1 Level of proficiency in the non-native language for each group of participants

Participants native language	Oral expression	Oral comprehension	Written comprehension
Spanish speakers	5.7	6.1	6.6
English speakers	1.29	1.57	1.71
German speakers	0	0	0

Ratings were given by the participants using a scale from zero to ten (0 = none; 10 = perfect) regarding their proficiency in speaking, understanding spoken language and reading

by the presentation of the visual-only sentence context. This leading stream of variable duration (1,600, 1,800, 2,000 or 2,200 ms), was followed by the audiovisual target fragment (800 ms) without any gap. The moment at which the target fragment began was varied randomly in order to create temporal uncertainty and thus to promote participant's sustained level of attention to the context. The task consisted of an audiovisual match/mismatch judgment on the audiovisual target. Following each response or time-out (1,800 ms deadline from target onset), the screen blanked for 800 ms before the next trial began. Participants were asked to respond as quickly and accurately as possible, whether the target fragment had matching or mismatching sound and lip movements, by pressing one of the two keys. Responses were made with the index and middle fingers on two neighboring keys, with the assignment of finger to response counterbalanced across participants.

There were three main types of trials that varied in the linguistic continuity between the visual context and the audiovisual target. On *intra-modal continuous* trials, the visual leading context continued into the visual modality of the target but did not continue into the auditory modality. On *cross-modal continuous* trials, the visual leading context fragment was continuous only with the auditory modality in the target fragment. Finally, on *discontinuous* trials, there was no continuity from context to target in any modality. In each of these three conditions, the correct response was a mismatch. The *cross-modal continuous* and *discontinuous* were the critical trials, while *intra-modal continuous* trials served to control for the nature of the response across the three mismatching trial types.

In addition to these critical trials, there were a number of filler trials in which auditory and visual channels matched in the target fragment, some of which were linguistically continuous and others discontinuous with the leading context. These filler trials were included in order to help compensate the likelihood of the two available responses. On *continuous matching* trials, there was linguistic continuity from visual context fragment to target fragment in both modalities, visual and auditory, while on *discontinuous matching* trials, there was no continuity between the visual leading context fragment and any modality in the target fragment, though in the target fragment audio and video were coincident. Overall, there were 60 % mismatching and 40 % matching trials in the experiment.

During the presentation of the visual context, the auditory channel was replaced by a sequence of rhythmic beats (300 Hz tones, 120 ms duration each, presented at 5 Hz). Regular beats were chosen so the stimuli were comparable in general temporal predictability to the syllabic rhythmical cadence of speech as well as to help maintaining the level of alertness similar to when we listen to someone speaking. Each participant was tested on a total of 200 trials, distributed in five equivalent blocks of 40 trials in which each trial type was equiprobable. These sentences were sampled randomly without replacement for each participant, with context duration varying randomly and equiprobably among the four possible context durations (1,600–2,200). Participants practiced the task on a subset of 20 training sentences prior to testing. Each experimental session lasted approximately 30 min.

As illustrated in Fig. 1, the comparison of greatest interest in the experiment was between the *discontinuous* and the *cross-modal continuous conditions*, both of which ended in an audiovisually mismatching target and involved an identical video splice between context and target fragments. Moreover, the discontinuous condition served as a comparison for both continuous mismatch conditions, in that it required the same response (a mismatch judgment) but the visual leading context provided no information about the message in the target fragment (since it belonged to a different sentence).

After the editing, sentences sometimes contained grammatical discontinuities (owed to the cross-splicing procedure followed, see below); however, it is important to note that, regarding visual predictability as well as grammatical discontinuities, the sentences were equivalent and counterbalanced across conditions.

We used the same stimuli and procedure to test for visual-to-auditory cross-modal prediction in a familiar language in a group of Spanish and a group of English native speakers. To test for cross-modal prediction in an unfamiliar language, we used the English materials with a group of Spanish speakers with low to medium English ability, and the Spanish materials in a group of English speakers and a group of German speakers, neither of whom were familiar with Spanish language.

Results

The median response time (RT) was calculated for each participant and condition, after filtering out errors and trials in which the response latencies were more than 2 standard deviations from the mean for each participant and condition (see Table 2). These RTs were then submitted to an ANOVA including the within-participants factor of context continuity (i.e., cross-modal continuous vs. discontinuous condition) and the between-participants factors of language familiarity (i.e., native vs. non-native) as well as language of the sentences (Spanish vs. English). These are the data shown in Fig. 2. Results from English and German speakers tested in a non-native language (Spanish) were analyzed together. We will refer to them as "English/German speakers" from now on, because their pattern of results did not differ significantly from one another. Moreover, both groups lacked of

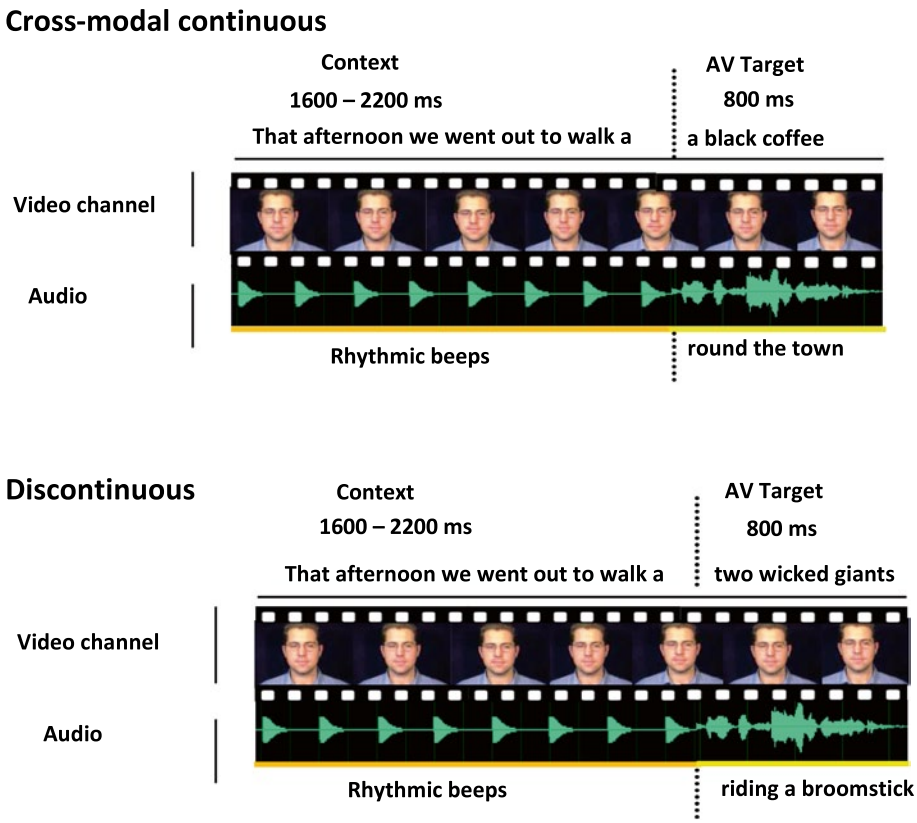

Fig. 1 Illustration of the critical conditions in the experiment. In the *cross-modal continuous* condition, the visual leading context fragment is continuous with the auditory modality (but not the visual modality) in the target fragment, while in the *discontinuous* condition, there is no continuity from visual context to target in any of the two modalities. In both conditions, visual and auditory modalities in the target fragment are mismatching. Context duration is chosen at random and equiprobably between 1,600 and 2,200 ms in steps of 200 ms

any knowledge of Spanish, and so for the aim of our study, their linguistic familiarity with Spanish was equivalent.

The critical finding was a significant two-way interaction between context and language familiarity (F (1,60) = 8.92; $p < 0.01$). This pattern of interaction reveals that participants were able to benefit from the prior informative visual context in the audiovisual matching task, but only in their native language (see Fig. 2). Furthermore, this pattern held independently of the particular language (English or Spanish) or population tested (English/German or Spanish), as indicated by the absence of a significant three-way interaction between context, language familiarity and test language

Table 2 Median response times (RT) and accuracy rate per group of participants and condition

Condition	Continuous matching		Discontinuous matching		Intra-modal continuous		Cross-modal continuous		Discontinuous mismatching	
	Median RTs	Accuracy	Median RTs	Accuracy	Median RTs	Accuracy	Median RTs	Accuracy	Median RTs	Accuracy
Native language group										
Spanish speakers	704	0.87	706	0.92	938	0.87	937	0.84	958	0.85
English speakers	794	0.84	753	0.84	949	0.86	904	0.84	958	0.83
Non-native language group										
Spanish speakers	814	0.92	797	0.9	1,014	0.86	999	0.84	1,002	0.83
English/ German speakers	870	0.81	864	0.9	1,018	0.87	1,046	0.86	1,026	0.85

Median RT was calculated after filtering out errors and trials in which the response latencies were more than 2 standard deviations above or below the mean for that participant and condition. Accuracy rates correspond to the participants who performed >65 % for all conditions in the experiment and were, therefore, included in the analysis. No significant difference was observed between the accuracy rates of the two critical conditions, cross-modal continuous and discontinuous mismatch (in all groups $|t| < 1$)

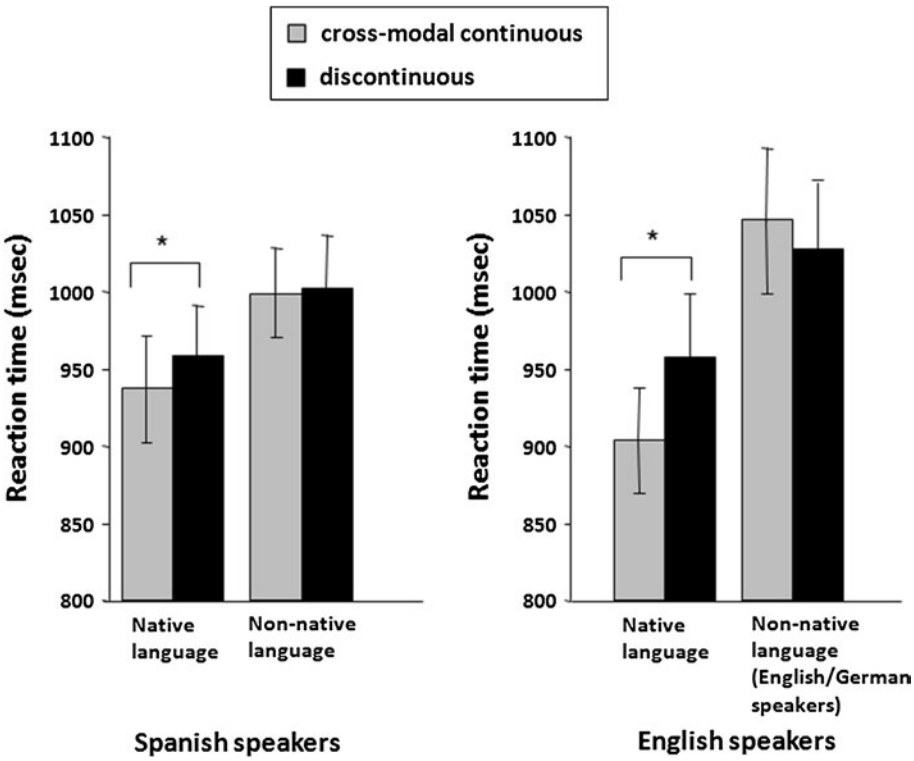

Fig. 2 Median response times for participants tested in their native and non-native languages for the critical conditions (cross-modal continuous and discontinuous). Spanish speakers (*left graph*) were able to detect audiovisual mismatches more rapidly following a cross-modal continuous versus a discontinuous visual leading context, but only when tested in their native language. English speakers (*right graph*) followed the same pattern; they were significantly faster detecting audiovisual mismatches in the cross-modal continuous condition compared with the discontinuous condition, but only when tested in their native language. *Error bars* represent one standard error of the median

(F (1,60) = 0.09; p = 0.76). The two-way interaction between context and test language was nearly significant, (F (1,60) = 3.28; p = 0.075), though note that it does not affect the conclusions in any way.

The significant interactions involving context were explored in greater detail with Student's t-tests comparing the two critical conditions (cross-modal continuous vs. discontinuous) in native and non-native groups of participants. Through these pairwise comparisons, we tested whether the context effect was significant within each of the familiarity and language conditions. These tests indicated that the cross-modal continuous condition led to faster RTs than the discontinuous condition for participants tested in their native language, t (27) = 3.91; $p < 0.01$, but not for participants tested in their non-native language, t (35) = 0.667; $p = 0.50$. Finer grained t-tests of the context effect (cross-modal continuous condition leading to faster RTs than discontinuous condition) for each particular group of participants considered separately showed that this pattern held independently of the test language. That is, the context effect was significant for Spanish participants tested with Spanish materials (t (17) = 2.42; $p < 0.05$) as well as for the English participants tested with English materials (t (9) = 3.36; $p < 0.01$), but not for Spanish speakers tested with English materials, t (18) = 0.17; p = 0.86), nor for English/German speakers tested with Spanish materials, t (16) = −1.45; p = 0.16).

Response accuracy was calculated for all groups of participants: native language (Spanish group = 87 %, English group = 84 %) or non-native language (Spanish group = 87 %, English/German group = 86 %). There were no significant differences in accuracy between the two critical conditions for any of the groups of participants (all $|t| < 1$) (see Table 2). Our conclusions were, therefore, based solely on the significant differences in RT.

Signal detection analyses, based on hits (responding "match" on a matching target) and false alarms (responding "match" on a mismatch target), were also performed to test for possible group differences (shown in Table 3). For participants tested in their native language, there were no differences in sensitivity between any of the critical conditions. Spanish group: cross-modal continuous $d' = 2.58$ and discontinuous $d' = 2.61$, ($|t| < 1$). English group: cross-modal continuous $d' = 2.13$ and discontinuous $d' = 2.10$, ($|t| < 1$). The response criterion parameter was significantly different from zero in both conditions for Spanish speakers (cross-modal continuous, $C = -0.24$, t (17) = −4.56, $p < 0.01$; discontinuous, $C = -0.22$, t (17) = −4.85, $p < 0.01$), indicating a similar response bias in both conditions. For the English group, the criterion was not significantly different from zero (cross-modal continuous, $C = -0.007$, discontinuous, $C = -0.026$, both $|t| < 1$), indicating the absence of a response bias of any kind. The groups tested in their non-native language did not reveal any significant differences in sensitivity. Spanish group: cross-modal continuous $d' = 2.57$ and discontinuous $d' = 2.43$, ($|t| < 1$). English/German group: cross-modal continuous $d' = 2.60$ and

Table 3 Sensitivity (d′) and criterion scores for Spanish and English speakers, tested in their native and non-native language

d′ values				Criterion scores			
Condition	Discontinuous	Cross-modal	Signif.	Discontinuous	Signif.	Cross-modal	Signif.
Native language group							
Spanish speakers	2.61	2.58	n/s	−0.22	<0.01	−0.24	<0.01
English speakers	2.10	2.13	n/s	−0.026	n/s	−0.007	n/s
Non-native language group							
Spanish speakers	2.43	2.57	n/s	−0.18	<0.01	−0.16	<0.05
English/German speakers	2.53	2.60	n/s	−0.14	n/s	−0.11	n/s

For participants tested in their native language, no differences were seen in performance between any of the critical conditions (*n/s* non-significant); criterion was significantly different from zero in both conditions for Spanish speakers, indicating a similar bias toward a positive response in both conditions. For the English group, the criterion was not significantly different from zero, indicating the absence of bias toward any kind of response. For groups tested in their non-native language, there were no significant differences in performance and the criterion measures followed the same pattern as for groups tested in their native language

discontinuous $d' = 2.53$, ($|t| < 1$). The criterion measure followed the same pattern as in the groups tested in their native language. In particular, for the Spanish group tested in English, the criterion revealed a bias toward "match" responses in both conditions: cross-modal continuous ($C = -0.16$; $t\ (18) = -2.50$, $p < 0.05$) and discontinuous condition ($C = -0.18$; $t\ (18) = -3.23$, $p < 0.01$). For the English/German group tested in Spanish, the criterion was not significantly different from zero (cross-modal continuous, $C = -0.11$; discontinuous, $C = -0.14$, all $|t| < 1$), indicating the absence of bias toward any kind of response.

We were surprised to find that during the testing, a large number of participants had difficulties performing the audiovisual matching task correctly, failing to meet a performance criterion of at least 65 % of correct responses, hence the relatively high rejection rate. Those who failed to meet this criterion included, in the non-native language groups, 12 native Spanish from a total of 31, 10 native English speakers from a total of 18, and 14 German speakers from a total of 23; and in the native language groups, 9 native Spanish and 8 native English speakers from a total of 27 and 18, respectively. We excluded their data from our reported analysis on the grounds that we could not be confident of the reliability of the correct RTs collected under those conditions (i.e., they could include a large proportion of lucky guesses).

Discussion

The results of the present study show that participants presented with audiovisual speech in their native language can benefit from prior visual information, responding more rapidly to an audiovisual match/mismatch task when the auditory channel of the target continues from the prior visual-only context. However, this benefit was not observed when participants performed the same task in an unfamiliar language. The primary result leading to this conclusion involved a comparison between a condition in which a visual only speech context was continuous with the auditory channel of the audiovisual target fragment, giving participants a response time advantage from this prior information, versus a condition where the visual context was not continuous with any modality in the target fragment, thereby forcing them to perform the matching task without any relevant leading information.

On one hand, these findings extend previous behavioral evidence of cross-modal prediction in one's own language (Sanchez-García et al. 2011) with a broader range of linguistic backgrounds (Spanish and English speakers) but, more importantly, they go one step further by showing that this predictive process breaks down in a language unfamiliar to the listener. Our finding implies that specific knowledge about the phonological repertoire of the language on the perceiver's side is required to capitalize effectively on anticipatory visual information during audiovisual speech processing. These findings also suggest that during speech perception, cross-modal correlations based on spatiotemporal dynamics between visual and acoustic signals are not by themselves sufficient to support effective cross-modal prediction during speech perception, at least not in all cases. It is also important to note that the present results cannot be due to the particular characteristics of one concrete set of speech materials, nor of one particular participant's group, because the pattern of effects was consistent between languages and across different groups.

It has been reported that phonemic categories are established during the first year of life (Best and McRoberts 2003; Best et al. 1995; Werker and Tees 1984) and are tailored to the specific linguistic input present in the environment. Once these initial representations have been settled, the consensus is that the phonemic repertoire of the native

language works as a "sieve," such that listeners will filter speech input according to these categories when perceiving any language. This favors an efficient categorization of the sounds in one's own language, but at the same time, it reduces the capacity to discriminate phonemes from a different language (Pallier et al. 1997, 2001; Navarra et al. 2005). This phenomenon is called *perceptual narrowing* (Werker and Tees 1984). Interestingly, recent studies using cross-modal matching in infants suggest that perceptual narrowing of the phonemic categories in one's own language in the first year of life is a pan-sensory phenomenon, affecting the capacity to perform cross-modal matching of speech sounds (Pons et al. 2009). Based on the present results, we suggest that the visemic information (visual phonemic) carried by lip movements is also decoded through the sieve of one's own native categories.

Several studies addressing phoneme discrimination in bilinguals have shown that even after an early and intensive exposure to a second language, the first language still exerts a great influence in the processing of the non-native phonemic categories of the second language (Sebastian-Galles and Soto-Faraco 1999; Pallier et al. 1997). Recently, studies with infants have shown differences in the way they deal with the perception of a native and a non-native language, from an early age. Lewkowicz and Hansen-Tift (2012) showed that in infants as young as 12 months old, visual attention begins to shift from the eyes to the mouth of the speaker and again to eyes while listening to a native language. Instead, while perceiving a non-native language, infants' attention shifts from eyes to mouth and not back to the eyes. According to the author's interpretation, this finding suggests that in such case infants continue searching for redundant information. In summary, phonemic categories that are defined early during development affect the way we process languages other than our native one. In line with this idea, speech information extracted from visible articulatory movements might only be exploited in full when the visemic categories belong to the listener's native repertoire. In a non-native language, phonological information is perceived, but it is more difficult to match to any existent category, hence being less efficient in order to ground any kind of online prediction mechanism.

Although there is some overlap between the phonological representations of any two languages, the remaining mismatch between the two phonological repertoires generally makes for a less effective categorization of non-native sounds. This mismatch ranges in extent from slight differences in the boundaries between phonological categories [i.e., voice onset time values between /p/ and /b/ are different between Spanish and English (Lisker and Abramson 1964)] to the deletion or insertion of whole new phonological categories (i.e., English /b/ vs. /v/ are both in the same /b/ category in [Castilian] Spanish (Best 1995)), and it poses considerable difficulties for adults trying to discriminate and identify non-native phonemic categories that overlap with a single phonemic category in their native language (Strange and Jenkins 1978; Werker and Tees 1999; Werker et al. 1981).

Phonology has been proposed as a common representational code for various aspects of speech perception (auditory and visual) as well as speech production (Pickering and Garrod 2006; Skipper et al. 2005, 2007; Fowler 2004). The present results are consistent with the view that audiovisual speech perception involves anticipation processes, whereby the visual correlates of auditory phonemes are useful in the online prediction of speech. The caveat added by the present findings is that this is more effective when observers have already learned the perceptual consequences of the articulatory gesture within a categorical representation.

These findings could be framed within the proposal that there is a strong relationship between speech perception and production. A common articulatory repertoire shared by perceptual and production systems was first proposed in the Motor Theory of speech perception (Liberman et al. 1967; Liberman and Mattingly 1985) and, more recently, by models of speech perception based on the predictive coding theory (Pickering and Garrod 2006; Skipper et al. 2005, 2007; Poeppel et al. 2008; Sams et al. 2005). According to these models, the motor system is recruited by visual speech gestures, in the same way as when producing them (i.e., activation is based on previous experience), sending efferent copies of the expected sensory consequences to sensory cortices (forward model), that are compared with the current perceived input. This makes it possible to speed up the processing of incoming information that matches expectations. For example, Skipper et al. (2007) proposed that the neural bases for this is the mirror neuron system (Rizzolatti et al. 1996), which has been shown to be comparable to the motor-auditory-mediated activation showed by several studies during speech perception (Fadiga et al. 2002; Skipper et al. 2007, 2009). These findings suggest that phonetic recognition during speech perception is possible because speaker and observer share the same articulatory motor repertoire. Although our results do not speak directly to the question of motor involvement in the predictive coding process, they are consistent with this possibility.

On a different level, our results align well with previous findings attesting a direct relationship between experience articulating speech sounds and an increase of visual influence during audiovisual speech perception (Desjardins et al. 1997; Siva et al. 1995). They are also consistent with differences in cross-modal temporal processing between native versus non-native speech (Navarra et al. 2010). For example, in order for audiovisual simultaneity to be perceived, the visual signal must lead the auditory signal in time, and Navarra et al. showed that this asynchrony is more

pronounced in one's own language (i.e., vision has to lead for a greater interval audition for both to be perceived as simultaneous). Navarra et al. argued that when the listener is familiar with the viseme–phoneme correlation, the perceptual system handles visual and auditory information as being more synchronous in time, reducing the apparent temporal separation between the visual and auditory signals (see also Sekiyama and Burnham 2008). This is consistent with the present results, which imply that audiovisual prediction is mostly effective when it occurs within the native phonemic repertoire of the listener.

The generally faster responses on matching trials as compared to mismatching ones across all groups in the present experiments (together with the significant bias to respond "match" for the Spanish participants in their native as well as non-native language) may reflect a strategy whereby participants have an a priori default to make a matching response. From this perspective, checking for disconfirmation (mismatching responses) would take longer than checking for a confirmation (matching responses). Although English speakers did not show this kind of bias, they nonetheless showed the speed up in response times in the matching versus the mismatching conditions. Faster responses during audiovisual matching trials could be related to the engagement of a more compact brain network, in comparison with the more widely distributed one recruited by mismatching conditions (Arnal et al. 2009).

Our central theoretical thrust has been to consider whether visuoauditory predictions occur primarily at a phonological-segmental level as opposed to lower levels, but it is worth considering whether higher levels, such as lexical representations, may also be relevant. Previous studies (Bernstein et al. 1998; Soto-Faraco et al. 2007; Altieri et al. 2011) have already demonstrated that visual information carries little lexical or semantic information on its own. Moreover, in our previous study (Sanchez-García et al. 2011), we explicitly tested this by asking a group of (Spanish) participants to lip-read as much information as they could from the same Spanish materials used here. The result was that subjects could not extract more than 4 % lexical information from the sentences. This makes us suspect that the lexical/semantic level of information may not play a strong role as a possible base for cross-modal facilitation in our experiments. Yet, it is fair to acknowledge that it cannot be excluded with certainty for all participants, given that we did not directly test the participants of this particular study for speech reading, and considering the large individual differences in the ability to speech-read (Bernstein et al. 1998, 2000).

On another count, we should not ignore the possibly important contribution of prosody (i.e., rhythm, stress and intonation) to language processing. The possibility that rhythmical information is used to support cross-modal prediction cannot be ruled out without further direct evidence. Consider, for instance, that in our experiments, the non-native language tested always involved a language of a different rhythmic class than the participant's own language (i.e., Spanish is a syllable-timed language, while English and German are stress-timed languages; Abercrombie 1967). Thus, although we are relatively confident in claiming phonology as an essential mediator of cross-modal prediction, it is more difficult to pinpoint whether it is segmental phonology, prosody, or a combination of the two, what plays a role (see Soto-Faraco et al. 2012, for related discussion). In this respect, it is interesting that previous studies testing visual language discrimination in adults (the ability to tell apart two different languages based on visual speech alone) have shown that it requires prior familiarity with the phonology of, at least, one of the comparison languages (Soto-Faraco et al. 2007). Remarkably, this ability is also present in young infants (4 months olds), but quickly lost during the process of acquisition of a native phonology (8 months of age), unless both comparison languages are relevant in the infant's environment (Weikum et al. 2007). One limitation of our experimental design is that it has so far focused on phonological and lower levels of representation, such that our materials leave little room for higher-level processes to operate. However, we believe that during the perception of speech in more naturalistic settings, it is likely that predictive processes will also operate at higher processing levels (i.e., lexicon), as has been documented in studies of word recognition (see Obleser and Eisner 2008; Norris and McQueen 2008; Levelt 2001).

Finally, although the present findings reveal the important role played by phonological representations in the process of audiovisual speech processing, it is only fair to acknowledge that they do not rule out the possible contributions from pre-phonological representations (Summerfield 1987; Schwartz et al. 1998, 2004). For instance, as highlighted in the Introduction, there is evidence that the perceptual system is sensitive to spatiotemporal matching across modalities between low-level features (Stekelenburg and Vroomen 2007; Arnal et al. 2009), including those where audiovisual interactions are supported by pre-phonological levels of representation (Green 1998; Green and Kuhl 1991; Grant 2001; Kim and Davis 2003; Rosenblum 2005). The specific claim we are making here concerns the possibility of rapid online cross-modal transfer of information from vision to audition, which is admittedly only one aspect of audiovisual processing. It is in this domain, of matching auditory and visual aspects of the speech signal, that our data suggest that the role of preexisting phonological representations is important. If lower level cues were the only basis for performing the present task of cross-modal matching, then the difference between native versus non-native speech input could not have been observed, since the low-level cues for

cross-modal matching were identical for all listeners of our study. This is why we interpret our evidence as supporting for theories that advocate a strong role for phonological levels of representation in audiovisual speech perception.

Appendix

Spanish sentences

1. Pensando que así el queso se mantendría más fresco.
2. Entró en un castillo y vio a un malvado gigante que tenía un caballo blanco.
3. Con un pico tan largo no podía coger los trozos de comida más grandes.
4. Mi hermano y yo estábamos ocultos en una extraña terraza de laureles.
5. Un rato después de la visita yo ya había olvidado esa cara redonda y llena.
6. El héroe del cuento fabricaba una máquina para producir espadas.
7. Le mostré la estatua china de piedra verde que yo había comprado esa misma mañana.
8. Salió para buscar un bar en el que hicieran café de calidad.
9. Me tapé con la manta, me tumbé en el sofá de casa.
10. Enseguida se levanta, guarda la rosa entre sus manos.
11. La fruta relucía en la mesa, parecía de mentira.
12. Los dos viejos amigos se acercaron cada uno a un espejo.
13. No daba nunca las gracias y nunca levantaba la vista para saber quién era el donante.
14. En las afueras de la ciudad, escondida entre los árboles, se encontraba una pequeña casita blanca.
15. Un día mientras el hombre se encontraba ya trabajando en el pueblo.
16. Los cristales estaban tan limpios que reflejaban la luz del sol.
17. Los libros estaban clasificados en diferentes temas.
18. Como el ordenador no funcionaba llamaron a un técnico.
19. La gente tenía que encender sus chimeneas casi toda la noche.
20. Ese hombre era conocido como el cómico más divertido del mundo.
21. El guitarrista era feliz sobre el escenario, tocando hasta la madrugada.
22. Después de caminar todo el día por un camino muy difícil al fin vieron las primeras luces del pueblo.
23. Cuando llegaron a la casa estaba muy sucia.
24. Hay tan pocas cosas por las que estar contento y feliz.
25. Pasó el verano y vino el otoño y el jardín continuó dando flores.
26. Los pocos que quedaban empezaron a reconstruir el castillo.
27. Su gran ilusión era salir de su casa durante el día.
28. En ocasiones, a uno le hace falta vivir una tragedia para volver a poner las cosas en perspectiva.
29. La oficina estaba totalmente vacía cuando ellos se fueron.
30. El hombre que le tenía que vender la bicicleta no cogía el teléfono.
31. El agua de la lluvia, que ha entrado sin parar, se ha colado hasta mis rodillas.
32. Tenían la televisión a un volumen tan alto que no se podían entender.
33. Todos los relojes se pararon a la misma hora.
34. El sol se filtraba entre los rascacielos del centro.
35. Acostado en la cama de matrimonio, un hombre con los ojos húmedos leía una carta que le trajo el correo.
36. La policía llegó a la manifestación de estudiantes tan rápido como pudo.
37. El bosque estaba muy quieto cuando los hombres llegaron.
38. El hombre de blanco le miró desde lejos, sonriendo.
39. Los patos se acercaron a ellos, esperando a que les tiraran un poco de comida.
40. Era inútil abrir los ojos y mirar en todas direcciones.

English sentences

1. Very old studies have classified one thousand species.
2. An old family lived in this house for many years.
3. The voyage took them over the bridge and up the mountain.
4. Another version of this silly story describes a fox.
5. There was but just one exception to this difficult and trying rule.
6. Although poor, they were happy together.
7. When the kind woman knew what had happened to her pet she was very sad.
8. A new student approached the Zen master and asked.
9. Leaving their home every morning before sunrise.
10. He never went to work. He didn't even go out.
11. The strongest thief climbed over the wall and ran.
12. The pain was so great that he could not sleep at night.
13. She is seven years old, her family is very rich.
14. When I first arrived, the entire town was talking about him.
15. Nothing but the prospect of a huge payment induced the corrupt man to have anything to do.
16. The fact that he was prime minister apparently had no effect.
17. That was one bad name to have when I was younger.
18. Sometimes the tone of his voice was harsh and strident.

19. If you were to be here, you could see for yourself.
20. Except at dinner-time my brothers and sisters and I used to see my father very little.
21. So when he tells her that the entire family will be leaving the city the next weekend she is not surprised.
22. He was only twelve years old when a younger brother of his mother came.
23. Later that night, in the dormitory he shared.
24. All of a sudden the room became as bright as the sun.
25. Normally, after their bath, the boys would be put straight into their pyjamas.
26. There was a mutual bond of loyalty and trust.
27. Quite a few girls were sorry to see James married.
28. He sat down next to the kitchen door.
29. Andrew disliked working and believed that at home he was free.
30. The Doctor has always been a good man, she thinks.
31. Peter happily accepted the rich proposal.
32. The household economy of my parents was on a humble plane.
33. A honeybee hive is far more than just a buzz of activity.
34. I am devoted to studying your system of martial arts.
35. The farmer got so old that, sadly, he couldn't work the fields anymore.
36. Many students wondered about this and went to the village.
37. The next morning another poem appeared on the wall in the hallway.
38. That night he was unable to sleep because he feared.
39. Finally, their voyage brought them to a cement wall.
40. On the other side they could hear the sound of a waterfall.

References

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

Altieri NA, Pisoni DB, Townsend JT (2011) Some normative data on lip-reading skills (L). J Acoust Soc Am Lett Ed 130(1):1–4

Arnal LH, Morillon B, Kell CA, Giraud AL (2009) Dual neural routing of visual facilitation in speech processing. J Neurosci 29:13445–13453

Arnal LH, Wyart V, Giraud AL (2011) Transitions in neural oscillations reflect prediction errors generated in audiovisual speech. Nat Neurosci 14(6):797–803

Arnold P, Hill F (2001) Bisensory augmentation: a speechreading advantage when speech is clearly audible and intact. Br J Psychol 92(2):339–355

Bernstein LE (2005) Phonetic perception by the speech perceiving brain. In: Pisoni DB, Remez RE (eds) The handbook of speech perception. Blackwell, Malden, pp 51–78

Bernstein LE, Demorest ME, Tucker PE (1998) What makes a good speechreader? First you have to find one. In: Campbell R, Dodd B, Burnham D (eds) Hearing by eye II: advances in the psychology of speechreading and auditory-visual speech. Psychology Press/Erlbaum, Taylor & Francis, Hove, pp 211–227

Bernstein LE, Demorest ME, Tucker PE (2000) Speech perception without hearing. Percept Psychophys 62(2):233–252

Bernstein LE, Auer ET Jr, Takayanagi S (2004) Auditory speech detection in noise enhanced by lipreading. Speech Commun 44:5–18

Besle J, Fort A, Delpuech C, Giard MH (2004) Bimodal speech: early suppressive visual effects in human auditory cortex. Eur J Neurosci 20:2225–2234

Best CT (1995) A direct realist view of cross-language speech, perception. In: Strange W (ed) Speech perception and linguistic experience. York Press, Timonium, pp 171–204

Best CC, McRoberts GW (2003) Infant perception of non-native consonant contrasts that adults assimilate in different ways. Lang Speech 46:183–216

Best CT, McRoberts GW, LaFleur R, Silver-Isenstadt J (1995) Divergent developmental patterns for infants' perception of two nonnative consonant contrasts. Infant Behav Dev 18:339–350

Bubic A, von Cramon DY, Schubotz RI (2010) Prediction, cognition and the brain. Front Hum Neurosci 4:25

Calvert GA, Spence C, Stein BE (eds) (2004) The Handbook of multisensory processes. The MIT Press, Cambridge

Chandrasekaran C, Trubanova A, Stillittano S, Caplier A, Ghazanfar AA (2009) The natural statistics of audiovisual speech. PLoS Comput Biol 5(7):e1000436

Dambacher M, Rolfs M, Gollner K, Kliegl R, Jacobs AM (2009) Event-related potentials reveal rapid verification of predicted visual input. PLoS ONE 4(3):e5047

DeLong KA, Urbach TP, Kutas M (2005) Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. Nat Neurosci 8:1117–1121

Desjardins RN, Rogers J, Werker JF (1997) An exploration of why preschoolers perform differently than do adults in audiovisual speech perception tasks. J Exp Child Psychol 66:85–110

Fadiga L, Craighero L, Buccino G, Rizzolatti G (2002) Speech listening specifically modulates the excitability of tongue muscles: a TMS study. Eur J Neurosci 15:399–402

Forster KI, Forster JC (2003) DMDX: a windows display program with millisecond accuracy. Behav Res Methods Instrum Comput 35:116–124

Fowler CA (2004) Speech as a supramodal or a modal phenomenon. In: Calvert GA, Spence C, Stein BE (eds) The Handbook of multisensory processing. The MIT Press, Cambridge, pp 189–202

Grant KW (2001) The effect of speechreading on masked detection thresholds for filtered speech. J Acoust Soc Am 109:2272–2275

Grant KW, Seitz PF (2000) The use of visible speech cues for improving auditory detection of spoken sentences. J Acoust Soc Am 108(3):1197–1208

Green KP (1998) In: Campbell R, Dodd B, Burnham D (eds) Hearing by eye II. Psychology Press, Hove (UK), pp 3–25

Green KP, Kuhl PK (1991) Integral processing of visual place and auditory voicing information during phonetic perception. J Exp Psychol Hum Percept Perform 17:278–288

Kim J, Davis C (2003) Hearing foreign voices: does knowing what is said affect visual-masked-speech detection? Perception 32:111–120

Levelt WJ (2001) Spoken word production: a theory of lexical access. Proc Natl Acad Sci USA 98:13464–13471

Lewkowicz DJ, Hansen-Tift AM (2012) Infants deploy selective attention to the mouth of a talking face when learning speech. Proc Natl Acad Sci USA 109:1431–1436

Liberman AM, Mattingly IG (1985) The motor theory of speech perception revised. Cognition 21:1–36

Liberman AM, Cooper FS, Shankweiler DP, Studdert-Kennedy M (1967) Perception of the speech code. Psychol Rev 74:431–461

Lisker L, Abramson AS (1964) A cross-language study of voicing in initial stops: acoustical measurements. Word 20:384–422

Luo H, Liu Z, Poeppel D (2010) Auditory cortex tracks both auditory and visual stimulus dynamics using low-frequency neural phase modulation. PLoS Biol 8(8):e1000445

MarianV, Blumenfeld HK, Kaushanskaya M (2007) The language experience and proficiency questionnaire (LEAP-Q): assessing language profiles in bilinguals and multilinguals. J Speech Lang Hear Res 50(4):940–967

Massaro DW (1998) Perceiving talking faces: from speech perception to a behavioral principle. The MIT Press/Bradford Books series in cognitive psychology, Cambridge

Massaro DW, Cohen MM (1983) Evaluation and integration of visual and auditory information in speech perception. J Exp Psychol Hum Percept Perform 9:753–771

McGurk H, MacDonald J (1976) Hearing lips and seeing voices. Nature 264:746–748

Navarra J, Soto-Faraco S (2007) Hearing lips in a second language: visual articulatory information enables the perception of second language sounds. Psychol Res 71:4–12

Navarra J, Sebastián-Gallés N, Soto-Faraco S (2005) The perception of second language sounds in early bilinguals: new evidence from an implicit measure. J Exp Psychol Hum Percept Perform 31(5):912–918

Navarra J, Alsius A, Velasco I, Soto-Faraco S, Spence C (2010) Perception of audiovisual speech synchrony for native and non-native language. Brain Res 1323:84–93

Norris D, McQueen JM (2008) Shortlist B: a bayesian model of continuous speech recognition. Psychol Rev 115:357–395

Obleser J, Eisner F (2008) Pre-lexical abstraction of speech in the auditory cortex. Trends Cogn Sci 13:14–19

Pallier C, Bosch L, Sebastian-Galles N (1997) A limit on behavioral plasticity in speech perception. Cognition 64:B9–17

Pallier C, Colomé A, Sebastián-Gallés N (2001) The influence of native-language phonology on lexical access: exemplar-based vs. Abstract lexical entries. Psy Sci 12(6):445–450

Pickering MJ, Garrod S (2006) Do people use language production to make predictions during comprehension? Trends Cogn Sci 11:105–110

Poeppel D, Idsardi WJ, van Wassenhove V (2008) Speech perception at the interface of neurobiology and linguistics. Philos Trans R Soc Lond B Biol Sci 363:1071–1086

Pons F, Lewkowicz DJ, Soto-Faraco S, Sebastian-Galles N (2009) Narrowing of intersensory speech perception in infancy. Proc Natl Acad Sci USA 106:10598–10602

Rizzolatti G, Fadiga L, Gallese V, Fogassi L (1996) Premotor cortex and the recognition of motor actions. Cogn Brain Res 3:131–141

Rosenblum LD (2005) Primacy of multimodal speech perception. In: Pisoni DB, Remez RE (eds) The handbook of speech perception. Blackwell Publishing, Malden, pp 51–78

Sams M, Mottonen R, Sihvonen T (2005) Seeing and hearing others and oneself talk. Brain Res Cogn Brain Res 23:429–435

Sanchez-García C, Alsius A, Enns JT, Soto-Faraco S (2011) Cross-modal prediction in speech perception. PLoS ONE 6:e25198

Schroeder CE, Lakatos P, Kajikawa Y, Partan S, Puce A (2008) Neural oscillations and visual amplification of speech. Trends Cogn Sci 12:106–113

Schwartz J, Robert-Ribes J, Escudier P (1998) Ten years after summerfield: a taxonomy of models for audio–visual fusion in speech perception. In: Campbell R, Dodd B, Burnham D (eds) Hearing by eye II: advances in the psychology of speechreading and auditory-visual speech. Psychology Press/Erlbaum (UK), Taylor & Francis, Hove, pp 85–108

Schwartz J, Berthommier F, Savariaux C (2004) Seeing to hear better: evidence for early audio-visual interactions in speech identification. Cogn 93:B69–B78

Sebastian-Galles N, Soto-Faraco S (1999) Online processing of native and non-native phonemic contrasts in early bilinguals. Cognition 72:111–123

Sekiyama K, Burnham D (2008) Impact of language on development of auditory-visual speech perception. Dev Sci 11:306–320

Siva N, Stevens EB, Kuhl PK, Meltzoff AN (1995) A comparison between cerebral-palsied and normal adults in the perception of auditory-visual illusions. J Acoust Soc Am 98:2983

Skipper JI, Nusbaum HC, Small SL (2005) Listening to talking faces: motor cortical activation during speech perception. Neuroimage 25:76–89

Skipper JI, van Wassenhove V, Nusbaum HC, Small SL (2007) Hearing lips and seeing voices: how cortical areas supporting speech production mediate audiovisual speech perception. Cereb Cortex 17:2387–2399

Skipper JI, Susan Goldin-Meadow S, Nusbaum HC, Small SL (2009) Gestures orchestrate brain networks for language understanding. Curr Biol 19:661–667

Sohoglu E, Peelle JE, Carlyon RP, Davis MH (2012) Predictive top-down integration of prior knowledge during speech perception. J Neurosci 32(25):8443–8453

Soto-Faraco S, Navarra J, Weikum WM, Vouloumanos A, Sebastian-Galles N, Werker JF (2007) Discriminating languages by speech-reading. Percept Psychophys 69:218–231

Soto-Faraco S, Calabresi M, Navarra J, Werker JF, Lewkowicz DJ (2012) The development of audiovisual speech perception. In: Bremner AJ, Lewkowicz DJ, Spencer C (eds) Multisensory development. Oxford University Press, Oxford, pp 207–228

Stekelenburg JJ, Vroomen J (2007) Neural correlates of multisensory integration of ecologically valid audiovisual events. J Cogn Neurosci 19:1964–1973

Strange W, Jenkins JJ (1978) Role of linguistic experience in the perception of speech. In: Walk RD, Pick HL (eds) Perception and experience. Plenum, New York, pp 125–169

Sumby WH, Pollack I (1954) Visual contribution to speech intelligibility in noise. J Acoust Soc Am 26:212–215

Summerfield Q (1987) Some preliminaries to a comprehensive account of audio-visual speech perception. In: Dodd B, Campbell R (eds) Hearing by eye: the psychology of lipreading. Lawrence Erlbaum Associates, New York, pp 3–51

Van Berkum JJ, Brown CM, Zwitserlood P, Kooijman V, Hagoort P (2005) Anticipating upcoming words in discourse: evidence from ERPs and reading times. J Exp Psychol Learn Mem Cogn 31:443–467

van Wassenhove V, Grant KW, Poeppel D (2005) Visual speech speeds up the neural processing of auditory speech. Proc Natl Acad Sci USA 102:1181–1186

Vroomen J, Stekelenburg JJ (2009) Visual anticipatory information modulates multisensory interactions of artificial audiovisual stimuli. J Cog Neurosci 22(7):1583–1596

Weikum WM, Vouloumanos A, Navarra J, Soto-Faraco S, Sebastian-Galles N, Werker JF (2007) Visual language discrimination in infancy. Sci 316:1159

Werker JF, Tees RC (1984) Phonemic and phonetic factors in adult cross-language speech perception. J Acoust Soc Am 75(6):1866–1878

Werker JF, Tees RC (1999) Influences on infant speech processing: toward a new synthesis. Annu Rev Psychol 50:509–535

Werker JF, Gilbert JHV, Humphrey K, Tees RC (1981) Developmental aspects of cross-language speech perception. Child Dev 52:349–355