The Language Learning Journal
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/rllj20

Basic to applied research: the benefits of audio-visual speech perception research in teaching foreign languages

Doğu Erdener [a]

[a] Psychology Program, Middle East Technical University - Northern Cyprus Campus, Güzelyurt, Morphou, North Cyprus

Version of record first published: 18 Sep 2012.

To cite this article: Doğu Erdener (2012): Basic to applied research: the benefits of audio-visual speech perception research in teaching foreign languages, The Language Learning Journal, DOI:10.1080/09571736.2012.724080

To link to this article: http://dx.doi.org/10.1080/09571736.2012.724080

The Language Learning Journal
2012, 1–9, iFirst article

Routledge
Taylor & Francis Group

Basic to applied research: the benefits of audio-visual speech perception research in teaching foreign languages

Doğu Erdener

Psychology Program, Middle East Technical University – Northern Cyprus Campus, Güzelyurt (Morphou), North Cyprus

Traditionally, second language (L2) instruction has emphasised auditory-based instruction methods. However, this approach is restrictive in the sense that speech perception by humans is not just an auditory phenomenon but a multimodal one, and specifically, a visual one as well. In the past decade, experimental studies have shown that the audio-visual aspects of speech perception have facilitative effects in L2 acquisition. This article has four theoretical and practical aims: (1) to synthesise the existing evidence from audio-visual speech perception (AVSP) in the context of L2 acquisition; (2) to demonstrate how L1–L2 orthographic matching can be used in L2 instruction; (3) to present ideas on how AVSP research can profitably be used in L2 teaching settings; and (4) to argue for the need for further applied and interdisciplinary research into the issues highlighted here.

Introduction

Speech perception has long been understood as an auditory-only phenomenon; however, it is in reality a multimodal process wherein both auditory and non-auditory modalities interact to create a resultant percept. The most robust form of non-auditory sources of speech information is visual speech information, in the form of lip and face movements. Although for a long time we have known the role of visual speech information in noisy listening conditions, where the availability of such information yields a clearer percept (Sumby and Pollack 1954), the evidence for the effect of visual speech information in clear listening conditions is relatively recent (McGurk and MacDonald 1976). In what is referred to as the McGurk Effect, when perceivers are presented with an auditory input (e.g. /ba/) coupled with a conflicting visual input (e.g. /ga/), they typically report a percept different to the actual inputs (e.g. /tha/ or /da/). This effect has also been shown in word and sentence contexts (e.g. Sams et al. 1998) and has come to be used as a common measure of visual speech influence wherein visually based responses are counted as a metric for visual speech influence. While the research in audio-visual speech perception centres around several themes, the focus of this article is on the perceptual aspects in the context of foreign/non-native language (L2, hereafter) perception and how findings in basic and applied research can profitably be used in educational settings.

*Email: d.erdener@gmail.com

ISSN 0957-1736 print/ISSN 1753-2167 online

http://dx.doi.org/10.1080/09571736.2012.724080
http://www.tandfonline.com

The potential role of visual cues in identifying consonants on the basis of their visual discernibility was investigated prior to the report of the McGurk effect (Binnie, Montgommery and Jackson 1974). In cross-linguistic studies of audio-visual speech perception, we observe two general patterns of results: (1) the level of visual speech influence in a given native language (L1, hereafter) appears to vary from language to language; (2) the visual speech influence for an L2 is usually greater than for L1. For instance, in a McGurk study, Sekiyama and Tohkura (1993) found that native Japanese speakers process visual speech information considerably less than their American counterparts in their respective native languages. However, Sekiyama (1997) also showed that Mandarin speakers were even less susceptible to visual speech than Japanese speakers. Sekiyama asserted that there may be less need to integrate visual speech information in Japanese compared to English as there are presumably fewer visually discernable phonemes and fewer vowels in Japanese than in English. The even weaker visual speech effect in Mandarin (and other tone-based languages), on the other hand, may be due to the observation that tonal information, where structured pitch variations indicate meaning differences, is not visually manifested (Burnham et al. 2000).

The studies by Sekiyama et al. also found that when listeners, irrespective of their L1, observed speakers speaking a foreign language, the amount of visual speech influence considerably increased. Sekiyama called this the foreign speaker effect. Other studies supported this finding in relation to other languages such as Dutch, German (Reisberg, McLean and Goldfield 1987), Korean (Davis and Kim 2001) and Spanish (Ortega-Llebaria, Faulkner and Hazan 2001). This finding clearly has important implications for L2 teaching. The foreign speaker effect was observed in further studies. Using a McGurk effect identification task, Chen and Hazan (2009) tested Mandarin- and English-speaking children aged 8 and 9 years and adults. They found that, irrespective of age, the effect of visual speech information was greater when participants interacted with a non-native speaker. There is also evidence suggesting language-specific influences in the foreign speaker effect. In another McGurk study, Hazan, Kim and Chen (2010) investigated the weighting of auditory and visual speech information when attending to native and non-native speech in clear and noisy listening conditions. They found that native English speakers showed a stronger weighting for visual information than their Mandarin counterparts. Wang, Behne and Jiang (2009) presented native Korean, Mandarin and English speakers with stimuli made up of labiodentals (e.g. /f/ as in *f*light, non-Korean), interdentals (e.g. /θ/, as in *th*ick, non-Korean and non-Mandarin) and alveolars (/s/ as in *s*till) in auditory–visual, auditory-only and visual-only listening conditions. Despite the fact that native English speakers performed better than the other two groups, both Korean and Mandarin perceivers showed native-like performance for labiodentals, which have a relatively higher degree of visibility than interdentals and alveolars, for which these groups showed poorer performance.

Some phonemes are thus harder to attain for certain speaker groups. Hazan et al. (2006), in their second experiment, investigated the well-known /r/–/l/ contrast, which is hard to achieve without explicit training or experience; this is particularly the case for native Japanese and Korean speakers as the /r/ and /l/ phonemes are not contrastively present in these two phonological repertoires. In this study, Hazan and her colleagues found no immediate benefit for the Japanese participants from either audio-visual or auditory training; however, Korean speakers showed small but significant benefit from auditory–visual training. On the same contrast, Hardison

(2003) found a significant effect for auditory–visual vs. auditory-only training with Japanese and Korean learners of English and showed how factors such as vowel position, word position and speaker might affect the attainment of this contrast.

A key factor in L2 speech perception is the degree to which a given L2 phoneme is relevant to one's L1 phonemic repertoire. The early speech perception development research of the past four decades has shown that early in life, human infants can discriminate most, if not all, speech contrasts in the world's languages based on their acoustic/phonetic properties (Jusczyk 1995). This innate skill is human infants' basic phonetic infrastructure to acquire *any language* readily as their L1. From around 6 months, however, this language-general skill is attenuated; first for vowels at around 6 months (Kuhl et al. 1992), then for consonants towards the end of first year of life (Miller and Eimas 1995). In this way, speech perception is shaped in a language-specific manner in accordance with the phonological repertoire of L1. For instance, while Japanese infants can discriminate the /r/–/l/ contrast perfectly during their first year, they gradually lose this ability thereafter (Christophe and Morton 1994).

Psycholinguistic research over the past four decades has yielded a large body of evidence on how this early speech perception attenuation affects our L2 acquisition. One model deriving from this work is Flege's Speech Learning Model (SLM) which is built on the assumption that the degree of attainment of a speech contrast in L2 depends partly on the degree to which L1 and L2 phonemic repertoires are compatible (Flege 2002). If there is sufficient phonemic difference between two phonemes then the perceiver will report them as different. For example, in the case of *r*ock vs. *l*ock, a Turkish perceiver for whom /r/ and /l/ are two separate phonemes will assimilate these two non-native instantiations of /r/ and /l/ into two separate native categories, whereas for a Japanese speaker they will be assimilated into the same category in the absence of sufficient exposure (see Best's (1994) Perceptual Assimilation Model) or experience in an L2, as evidenced by production studies (Ingvalson, McClelland and Holt 2011).

There are two key implications from the above studies for L2 teaching: firstly, L2 speech processing, like L1 speech processing, is a multimodal event, particularly an audio-visual one; secondly, given the individual differences highlighted in several studies (e.g. Hazan et al. 2010), as well as cross-linguistic differences in the amount of visual speech use, a one-size-fits-all approach to developing learners' L2 speech perception and production is clearly inadequate. This is the key argument which this article seeks to develop. Particularly in the past decade, there has been a growing body of research on the relationship between the multimodal nature of speech perception and foreign language acquisition in terms of both perceptual and productive aspects. The next section will review the current state of this research.

Audio-visual speech perception and second language acquisition

A growing body of research has studied the relationship between the modalities of L2 processing. In one such study, Ortega-Llebaria et al. (2001) tested native Spanish speakers learning English on their perception of English consonants. English differs significantly from Spanish in its phonological repertoire and most native speakers of Spanish confuse English phonemes that are not present in Spanish in both perception and production tasks. For example, English /t/ and /d/ are usually assimilated to the Spanish /t/, while English /ð/ (as in *that*) and /d/ are confused with and assimilated into the Spanish /d/. Ortega-Llebaria et al. (2001) exposed their participants to a group of English consonants and vowels in two training conditions: auditory–visual, in which phonemes to be learnt were taught by means of an interactive conversational agent that provided both auditory and visual information for these phonemes, and auditory-only wherein subjects only heard these phonemes. The results showed that providing audio-visual speech information greatly reduced consonant perception errors compared to the auditory-only condition. In a more recent study, Hazan and her colleagues tested Spanish and Japanese learners on the effect of phonetic information within the visual speech cues on the learning L2 English speech contrasts. The learners were presented with British English-native labial and labiodental speech contrasts made up of /p/, /b/ and /v/ in video-only, audio-only and audio-visual listening conditions. The results yielded a clear advantage for the audio-visual condition; training with visual and auditory information was the most effective for developing perception of these relatively difficult contrasts (Hazan et al. 2006).

The apparent benefits of exposing L2 learners to speech in visual and other language-appropriate modalities (e.g. orthographic) are not just limited to perception but also production in a foreign language. Hazan et al. (2005) trained Japanese participants on the perception and production of the /v/–/b/ contrast using two training modalities, auditory and auditory–visual. They found that auditory–visual training was more effective than auditory-only training on perception and production of this contrast; however, unlike earlier studies (Hardison 2003), the auditory–visual training improved neither perception nor production of the /r/–/l/ contrast. Presumably, unlike the visually salient /v/–/b/ contrast, the /r/–/l/ contrast lacked visible discernibility that appears to be a prerequisite for visual benefit. This suggests certain non-native phonemes require more focused training than others, depending on the L1–L2 phonological compatibility (see Flege 2003).

Such L1–L2 compatibility is not limited only to phonological repertoires but also applies to other substructures as well, such as orthography. Erdener and Burnham (2005), for example, tested monolingual native speakers of Australian English and Turkish on Spanish and Irish speech stimuli. English and Irish have relatively opaque orthographies characterized by inconsistent one-to-many letter-sound correspondences (e.g. the letter 'a' corresponds to three different phonemes in the words *cadet*, *apple* and *art*), while Turkish and Spanish have much more regular letter-sound correspondences wherein almost every letter refers to one sound. Two groups of monolingual English- and Turkish-speaking participants were presented with Spanish (a regular orthography) and Irish (a relatively opaque orthography) stimuli in four presentation conditions: auditory-only (AO – subjects only heard the target stimuli), auditory–visual (AV – subjects heard the stimuli and saw the speaker's face), auditory–orthographic (A-Orth – subjects heard and read the stimuli) and auditory–visual–orthographic (AV-Orth – subjects heard the stimuli, saw the speaker's face and saw the stimuli presented in written form in a caption). In each condition, subjects were asked to attend to each stimulus and repeat it. The dependent variable was the mean number of errors made in each condition as rated by native speakers of Spanish and Irish, respectively.

The results yielded two patterns of outcomes. First, in conditions where visual speech information was present, the production performance was superior to when visual information was absent, confirming earlier findings with perceptual tasks. Second, there was an interesting pattern of results when orthographic information was presented: whether or not visual information was present, Turkish subjects appeared to rely on orthographic information more than their Australian counterparts as revealed by the number of errors in the orthographic conditions (A-Orth and AV-Orth). This, apparently, was a particularly good strategy with the Spanish stimuli (transparent orthography) but not with Irish stimuli (opaque orthography). In contrast, Australian subjects seemed to ignore the orthographic input and relied mostly on auditory and visual sources, maintaining comparable performances for Spanish and Irish stimuli across orthographic conditions (Figure 1).

The results from Erdener and Burnham (2005) indicate two key points for L2 training. First, as suggested by earlier studies of perception, visual speech information can help learners not only to discriminate non-native phonemes but also to produce them more accurately. In this regard, L2 teaching approaches should incorporate the use of audio-visual speech input more systematically. In particular, previous research has shown that hyper-articulation of speech enhances the resultant percept (Lees and Burnham 2005), a finding that fits nicely in L2 teaching practice (see below). Second, other non-auditory aspects of L2, such as orthography, can be used advantageously to promote accurate speech perception and production in specific language teaching contexts. For instance, native speakers of languages with a transparent writing system, such as Turkish, will benefit from L2 orthographies with similar sound-letter transparencies, such as Spanish or Italian. However, the value of L2 written presentation is less clear in teaching speakers of an orthographically opaque language such as English.

Applying audio-visual speech perception research in L2 acquisition

As indicated above, one of the aims of applied research in audio-visual speech perception is to develop effective speech technologies in various domains such as L2

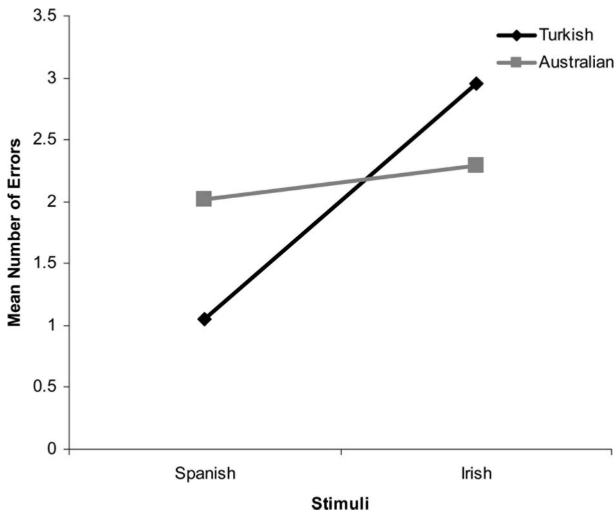

Figure 1. The overall correct mean performance in producing Spanish and Irish stimuli by Turkish and Australian participants in AV–orthographic and auditory–orthographic conditions combined (adapted from Erdener and Burnham 2005).

education. In what follows, we develop our thinking on how the findings detailed above, linked to further research, can be used in online and classroom L2 settings.

Classroom and online teaching of pronunciation

Pronunciation teaching can be carried out face-to-face or online. In presenting models for pronunciation and engaging interactively with students, teachers need to be aware that the observation of hyper-articulated speech has been shown to result in clearer speech perception, in both L1 and L2. Apparently, when teaching young children, the speech style of teachers typically changes to what Håkansson (1987) has referred to as *teacherese*, a distinct hyper-articulated speech style. Hyper-articulated speech style is marked with exaggerated articulatory movements that enhance the visual speech information provided to the perceiver and yield better visual speech detection (Lees and Burnham 2005).

L2 instruction can also be delivered through online audio-visual material such as instructional videos. While some commercially available online computer-aided language learning (CALL) systems claim to improve L2 pronunciation (e.g. www.L2accent.com 2011), the feedback they provide comes from speech processing software. Research suggests that the speech clarity and accent judgments made by CALL systems still do not match those made by human listeners (Sauro 2009; Müller et al. 2009). Thus while CALL systems can usefully support appropriate input and feedback from face-to-face teachers, they are not yet sophisticated enough to replace them. Efforts are nevertheless under way to develop interactive systems that will eventually teach learners speech contrasts that exist in L2 but not in their L1 (Wik and Hjalmarsson 2009). Attaining these differences is important in realising word-based semantic differences both perceptually and in production (e.g. Japanese learners of English attaining the /r/ vs. /l/ difference as in, *rock* and *lock*, respectively).

Teaching L2 with orthography

While the focus of this article has been on the beneficial aspects for L2 learning of audio-visual speech perception, we have also touched on the value of orthographic presentation, at least in cases where L1 and L2 orthographies are relatively transparent. It appears that speakers of L1s with transparent orthography usually outperform speakers whose L1 has an opaque orthography on tests of phonological awareness (Oktay and Aktan 2002). Phonological awareness is the ability to manipulate speech elements, measured through mental operations such as phoneme deletion, counting number of syllables, etc. despite some preliminary findings, the relationship between L1 and L2 orthographies in the context of L2 instruction warrants further research. As such, we also have yet to understand the underlying phonological properties of audio-visual speech perception.

Conclusions

L2 research on the applicability of speech perception theories has typically focused on either native English speakers or speakers who are learning English. Recently, however, research has started to focus on a wider range of language learning contexts; for example, work on the application of speech perception models such as

Best's (1994) Perceptual Assimilation Model has usefully focused on native speakers of Turkish (Komurcu and Yildiz 2011). While the continuation of basic and applied research in testing these models with new languages is of paramount importance, it is also important that the language teaching profession builds awareness, in the light of the evidence presented here, of the range of potential pedagogic tools supported by speech perception research. It is also important that the auditory–visual aspects of speech perception relating to language acquisition, be it L1 or L2, should be implemented not just into L2 curricula but also into standardised language tests. The audio-visual speech perception research has yet to mature in terms of interdisciplinary collaboration between psychologists, engineers, linguists and education scientists in order for such implementations to take place and sophisticated and effective tests to be developed. In a growing number of studies, nevertheless, the use of such technologies and their benefits in L2 teaching are reported. Using an interactive animated agent, Massaro and Light (2003) trained and tested Japanese learners of English on the /r/ and /l/ contrast. The animated agent, Baldi, not only showed the outer mouth and lip movements but the movements of articulators inside the mouth. Results showed significant improvement both in learners' speech perception and production. Such systems and the development of other multipurpose animated interactive systems (e.g. The Thinking Head Project 2012) in the context of L2 acquisition (Anderson et al. 2009) set examples of how the development of interactive agents can pave the way for better teaching and learning tools for both self-paced and classroom learning in this area.

Acknowledgements

This manuscript was prepared with the support of Middle East Technical University – Northern Cyprus Campus Scientific Research Project Fund (ODTÜ-KKK Bilimsel Araştırma Proje Fonu, SOSY-12) provided to the author. The author expresses his gratitude to *The Language Learning Journal* Editorial Board and the anonymous reviewers for their invaluable contribution to this article.

References

Anderson, T., R. Leibbrandt, T. Lewis, M. Luerssen and D. Powers. 2009. Poverty of stimulus in language education. Paper presented at the HCSNet Summerfest, 30 November–4 December, in University of New South Wales, Sydney, Australia.

Best, C.T. 1994. The emergence of native-language phonological influences in infants: A perceptual assimilation model. In *The Development of Speech Perception*, ed. C. Goodman and H. Nusbaum, 167–224. Cambridge, MA: The MIT Press.

Binnie, C.A., A.A. Montgommery and P.L. Jackson. 1974. Auditory and visual contributions to the perception of consonants. *Journal of Speech and Hearing Research* 17: 619–30.

Burnham, D., V. Ciocca, C. Lauw, S. Lau and S. Stokes. 2000. Perception of visual information for Cantonese tones. In *Proceedings of the Eighth Australian International Conference on Speech Science and Technology*, ed. M. Barlow, 86–91. Canberra: Australian Speech Science and Technology Association, Inc.

Chen, Y. and V. Hazan. 2009. Developmental factors and the non-native speaker effect in auditory–visual speech perception. *Journal of the Acoustical Society of America* 126, no. 2: 858–65.

Christophe, A. and J. Morton. 1994. Comprehending baby-think. *Nature* 370: 250–1.

Davis, C. and J. Kim. 2001. Repeating and remembering foreign language words: implications for language teaching systems. *Artificial Intelligence Review* 16, no. 1: 37–47.

Erdener, V.D. and D.K. Burnham. 2005. The role of audiovisual speech and orthographic information in nonnative speech production. *Language Learning* 55, no. 2: 191–228.

Flege, J.E. 2002. Interactions between the native and second language phonetic systems. In *An Integrated View of Language Development:Papers in Honor of Henning Wode*, ed. P. Burmeister, T. Piske and A. Rohde, 217–44. Trier, Germany: Wissenschaftlicher Verlag.

Håkansson, G. 1987. *Teacher Talk: How Teachers Modify their Speech when Addressing Learners of Swedish as a Second Language*. Lund, Sweden: Lund University Press.

Hardison, D.M. 2003. Acquisition of second-language speech: effects of visual cues, context, and talker variability. *Applied Psycholinguistics* 24, no. 4: 495–522.

Hazan, V., J. Kim and Y. Chen. 2010. Audiovisual perception in adverse conditions: language, speaker and listener effects. *Speech Communication* 52, nos. 11–12: 996–1009.

Hazan, V., A. Sennema, A. Faulkner and M. Ortega-Llebaria. 2006. The use of visual cues in the perception of non-native consonant contrasts. *Journal of the Acoustical Society of America* 119, no. 3: 1740–51.

Hazan, V., A. Sennema, M. Iba and A. Faulkner. 2005. Effect of audiovisual perceptual training on the perception and production of consonants by Japanese learners of English. *Speech Communication* 47, no. 3: 360–78.

Ingvalson, E.M., J.L. McClelland and L.L. Holt. 2011. Predicting native English-like performance by native Japanese speakers. *Journal of Phonetics* 39, no. 4: 571–84.

Jusczyk, P. 1995. Language acquisition: speech sounds and the beginning of phonology. In *Speech, Language, and Communication*, ed. J.L. Miller and P.D. Eimas, 263–301. New York: Academic Press.

Komurcu, B. and Y. Yildiz. 2011. An assessment of the auditory speech perceptual skills of young Turkish learners of English: perceptual assimilation model. *Procedia – Social and Behavioral Sciences* 15: 3511–5.

Kuhl, P.K., K.A. Williams, F. Lacerda, K.N. Stevens and B. Lindblom. 1992. Linguistic experience alters phonetic perception in infants by 6 months of age. *Science* 255: 606–8.

Lees, N. and D. Burnham. 2005. Facilitating speech detection in style!: the effect of visual speaking style on the detection of speech in noise. In *Proceedings of the Auditory–Visual Speech Processing International Conference*, ed. E. Vatikiotis-Bateson, D. Burnham and S. Fels, 23–8. Parksville, BC: Causal Productions.

Massaro, D.W. and J. Light. 2003. 'Read my tongue movements': bimodal learning to perceive and produce non-native speech /r/ and /l/. *EUROSPEECH-2003*: 2249–52.

McGurk, H. and J. MacDonald. 1976. Hearing lips and seeing voices. *Nature* 264: 746–8.

Miller, J. and P. Eimas. 1995. Speech perception: from signal to word. *Annual Review of Psychology* 46: 467–92.

Müller, P., F. de Wet, C. van der Walt and T. Niesler. 2009. Automatic assessment of oral language proficiency and listening comprehension. *Speech Communication* 51, no. 10: 864–74.

Oktay, A. and E. Aktan. 2002. A cross-linguistic comparison of phonological awareness and word recognition in Turkish and English. *International Journal of Early Years Education* 10, no. 1: 37–48.

Ortega-Llebaria, M., A. Faulkner and V. Hazan. 2001. Auditory–visual L2 speech perception: effects of visual cues and acoustic–phonetic context for Spanish learners of English. *Speech, Hearing and Language: Work in Progress* 13: 39–51.

Reisberg, D., J. McLean and A. Goldfield. 1987. Easy to hear but hard to understand: a lip-reading advantage with intact auditory stimuli. In *Hearing by Eye: The Psychology of Lip-Reading*, ed. R. Campbell and B. Dodd, 97–114. Hillsdale, NJ: Erlbaum Associates.

Sams, M., P. Manninen, V. Surakka, P. Helin and R. Kättö. 1998. McGurk effect in Finnish syllables, isolated words, and words in sentences: effects of word meaning and sentence context. *Speech Communication* 26, nos. 1–2: 75–87.

Sauro, S. 2009. Computer-mediated corrective feedback and the development of L2 grammar. *Language Learning and Technology* 13, no. 1: 96–120.

Sekiyama, K. 1997. Cultural and linguistic factors in audiovisual speech processing: the McGurk effect in Chinese subjects. *Perception and Psychophysics* 59, no. 1: 73–80.

Sekiyama, K. and Y. Tohkura. 1993. Inter-language differences in the influence of visual cues in speech perception. *Journal of Phonetics* 21, no. 4: 427–44.

Sumby, W.H. and I. Pollack. 1954. Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America* 26, no. 2: 212–5.

The Thinking Head Project. 2007. Retrieved May 9 2012, from the Thinking Head website, http://thinkinghead.edu.au/.
Wang, Y., D. Behne and H. Jiang. 2009. Influence of native language phonetic system on audio-visual speech perception. *Journal of Phonetics* 37: 344–56.
Wik, A. and P. Hjalmarsson. 2009. Embodied conversational agents in computer assisted language learning. *Speech Communication* 51, no. 10: 1024–37.