1st International Black Sea Conference on Language and Language Education
September 22-23, 2017, Ondokuz Mayıs University, SAMSUN

AUDITORY-VISUAL SPEECH PERCEPTION AND SECOND LANGUAGE INSTRUCTION

Dr. Doğu Erdener
Middle East Technical University, Northern Cyprus Campus, Psychology Program
vdogu@metu.edu.tr

ABSTRACT

Speech perception is not just an auditory phenomenon but an auditory-visual one: we process lip and face movements during speech as demonstrated by research in the past four decades as well as paradigms such as the McGurk Effect – an illusory experience in which conflicting auditory and visual (face and lip movements) speech information results in a percept that is available in neither modality. This paper has three purposes: (1) to present the current state of cross-language studies in auditory-visual speech perception from an applied perspective; (2) how studies in auditory-visual speech perception are relevant to both second language (L2) research and instruction; (3) how the knowledge from this research is / can be applied to grammatically and phonotactically distinct languages such as English and Turkish – the latter of which is in need of more research than presently available. While the necessity to study Turkish as an L2 is elucidated here from an experimental psychology point of view, the need for this scrutiny is highlighted from both an applied and instructional stance as well. In this illusory effect,

INTRODUCTION

Auditory vs. auditory-visual speech perception

For a considerable length of time speech perception has been understood as a phenomenon that is exclusively and solely auditory-based. Likewise, traditional methods of language instruction embraced mostly auditory-based methods of instruction although use of video sources are common. However, past forty years of research has witnessed the emergence of the fact that speech perception is not just an auditory phenomenon but a an auditory-visual one – that is, we process not only what we *hear* but what we *see* in the form of face and lip movements during speech. We also should, to some extent and anecdotally, realise that speech is also visual especially in impoverished listening conditions (Sumby & Pollack, 1954) as well as in clear listening conditions as initially demonstrated by a phenomenon that has come to be known as the *McGurk Effect* (McGurk & MacDonald, 1976). In a classic demonstration of the *McGurk Effect*, perceivers are presented with a combined stimulus of auditory /ba/ and visual /ga/, two conflicting information. Most people who get this effect report to perceive a "da", a response that does not exist physically. This effect basically shows that what we see influences what we hear, eventually and evidently influencing what we perceive. The McGurk has also been used to date as an index of visual speech influence: broadly speaking, the greater the "da"-like responses, the greater the effect of visual speech information. So, what exactly is the relevance of visual speech information as such in the context of second language (L2, hereafter) acquisition?

First and foremost, the fact that visual speech information (i.e., when the face of the talker is visible to the listener) enhances the comprehension of the conveyed message has been known to us for a very long time now (Cotton, 1935) whether in a syllable (e.g., McGurk & MacDonald, 1976), or word and sentence context (Sams, Manninen, Surakka, Helin & R. Kättö, 1998). Further to this, if a spoken phrase is hyperarticulated, its perception is enhanced compared to when a perceiver is exposed to a hypo-articulated speech (Lees & Burnham, 2005). A form of hyperarticulated speech style is, in fact, deployed in L2 classes by foreign language teachers and termed as the *Teacherese* (Håkansson, 1987). However, Håkansson (1987) suggests that most of

the time this hyperarticulated speech style is used implicitly with hardly any conscious awareness as is the case when we talk to foreigners. What is coined as foreigner-directed speech (FDS), like Teacherese, is marked by an exaggerated and extended pitch range and set of contours (Uther, Knoll & Burnham, 2007; but also see Biersack, Kempe & Knapton, 2005).

What we can learn from animals and babies

Although 30 years have passed since Håkansson's presenting of the concept of *Teacherese*, there is still a paucity of research into how we benefit from this implicitly deployed speech style especially in a learning environment. Developmental evidence suggests that another naturally elicited speech style with also exaggerated pitch contours known as *infant-directed speech* (IDS) has a special linguistic function. In a study by Burnham and colleagues it was demonstrated that while we speak to both babies and non-human animals (pet-directed speech, PDS) with elevated and exaggerated pitch patterns, when focused on vowels, the pitch pattern is less exaggerated when speaking to animals than to human infants – revealing a more hyperarticulated and clearer speech style, presumable, according to Burnham et al. (2002), serving a linguistic function: helping infant become familiarised with their native language phonology.

Whilst FDS (along with IDS and PDS) is marked by heightened pitch and that it has some visibly discernible correlate (Lees & Burnham, 2005), our current knowledge – which thankfully, is cumulative enough to write up this paper – is bordered by ever-growing literature on the relationship between auditory-visual speech perception and L2 (or non-native speech if subjects in a study are not actual learners of an L2) acquisition to which we now turn our attention.

Auditory-visual speech perception and L2 acquisition

Let us say what we will say at the end: we benefit from visual speech information in the L2 (or non-native speech) context irrespective of whether we have awareness of this knowledge. So how does this happen and in which language contexts?

We started up the issue of auditory-visual speech perception with *McGurk Illusion*. However, this illusion is not present homogeneously across the World's languages; on the contrary, languages differ with respect to the strength of visual speech influence within themselves. Comparing native speakers of English and Japanese over a series of McGurk stimuli, Sekiyama and Tohkura (1993) found that Japanese speakers make less use of (or are less influenced by) visual speech information compared to their English-speaking counterparts. Intriguingly, a similar group of Japanese speakers made relatively *greater* use of visual speech than native Mandarin speakers in a subsequent study (Sekiyama, 1997). These results were attributed to both cultural and linguistic factors the former of which was contestable. Linguistically speaking, the presence of greater number of consonants and consonant clusters in English relative to Japanese and the paucity of discernibility of visual speech information in lexical tones in tonal languages like Mandarin and Cantonese (Burnham et al. 2000) may have led to these differences. In both studies, though, the visual speech effect was far greater when listening to a foreign talker (e.g., English speakers listening to a Japanese speaker). This finding was consistently present in both pure experimental (e.g., Davis & Kim, 2001) and experimental /applied studies in the L2 context (Ortega-Llebaria, Faulkner & Hazan, 2001; Hazan, Kim & Chen, 2010) and this was irrespective of age (Chen & Hazan, 2009).

Whether or not a given phoneme is present in one's L1 can be a determinant of whether that phoneme will be perceived and produced in a native-like manner. Flege (2002), for instance, states 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. If a sufficient degree phonemic difference exists between two L2 phonemes, then the perceivers will perceive them to be different. For example, in the case of *rock vs. lock*, a native Turkish speaker for whom the phonemes /r/ and /l/ are two

separate categories, will assimilate these two non-native instantiations of /r/ and /l/ into two separate native categories; however, a native Japanese speaker will assimilate them into a single native category (presumably, /ɾ/) in the absence of sufficient exposure to or experience with that L2 (also see Best's (1994) Perceptual Assimilation Model), as also evidenced by production studies (Ingvalson, McClelland & Holt 2011). In a study along this theme, 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. /ɵ/, as in *th*ick, non-Korean and non-Mandarin) and alveolar (/s/ as in *s*till) in auditory–visual, auditory-only and visual-only listening conditions. Within-subject results showed that both Korean and Mandarin perceivers showed native-like performance for labiodentals, which have a relatively higher degree of visibility than interdentals and alveolar, for which these groups showed poorer performance. Thus provision of visual speech information appears to pave the way for clearer perception and production of L2 phonemes even if the learners are not aware of this. Visual discernibility of

Raising awareness of visual speech information in the context of L2 learning can be done in both actual classroom settings and technology-based tools. In one study whereby Ortega-Llebaria et al. (2001) used an interactive conversational agent programmed to teach non-native phonemes and found that the learners' error rates in perception were significantly reduced in the auditory-visual training condition compared to the auditory-only form of exposure.

As the above evidence and ever-growing literature on auditory-visual speech perception and its relevance to L2 instruction simply tells us that we need to make use of and exploit all relevant sources of perceptual information available to L2 learners – namely both visual and auditory. It should, however, be noted that there are post-perceptual / cognitive sources of information available to L2 learners, a prominent one of which is orthographic information. A limited amount of research has so far shown that the benefit of orthographic information in learning L2 depended on the orthographic features of L2, namely transparency – the extent to which there is coherence between graphemes and phonemes - of L2 as well as L1.

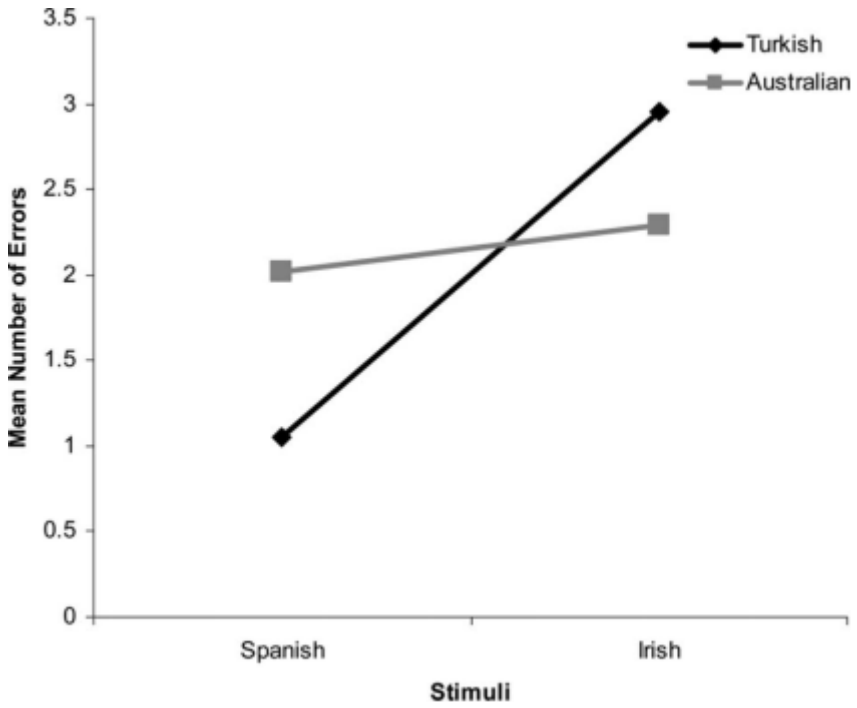

Figure 1. The collapsed mean correct ratings by native speakers for Spanish and Irish stimulus productions by Turkish and Australian participants in the orthographic experimental conditions (data adapted from Erdener & Burnham, 2005 and used in Erdener, 2016).

In one bridging study by Erdener and Burnham (2005), native speakers of English (a language with opaque orthography predominantly featuring inconsistent phoneme-grapheme correspondences) and Turkish (a language with transparent orthography marked by consistent phoneme-grapheme correspondences) over a series of Irish (opaque orthography) and Spanish (transparent orthography) stimuli in four experimental conditions: auditory, auditory-visual, auditory-visual-orthographic and auditory-orthographic. The dependent variable was the number of production errors by participants whose task was to repeat the each stimulus. While a comparison of the performances in the auditory and auditory-visual conditions yielded a similar result as previous studies did revealing an advantage of visual speech information. Further to this, the comparison of performances in the orthographic conditions, namely, auditory orthographic and auditory-visual-orthographic, showed this: whether or not visual information was present, native Turkish speakers relied on orthographic information more than their native English-speaking counterparts as revealed by the number of errors with much better performance with Spanish stimuli than with Irish ones – a good strategy for Spanish (transparent orthography) but not for Irish (opaque orthography). In contrast, English speakers seemed to ignore the orthographic input and relied mostly on auditory and visual sources, with similar performances for Spanish and Irish stimuli across the orthographic conditions (Figure 1).

In short, these results reveal that there are language-general features that can (and should) be implemented and used in L2 instruction process (e.g., visual speech information) and language-specific factors that pertain to certain specificities of a given language (e.g., Spanish having a transparent orthography that facilitates learning among at least those with L1s with a transparent orthography). Next, we will slightly shift our attention to how the abovementioned research can profitably be used in teaching phonotactically and orthographically (and logographically in some cases) distinct languages such as English, Turkish, etc.

CONCLUSION

Auditory-visual speech perception: Teaching Turkish as an L2

There are only two published studies to the best of the author with respect to Turkish in the context of auditory-visual speech perception (Erdener & Burnham, 2005; Erdener, 2015) one of which essentially showed that native Turkish speakers are prone to the McGurk Effect and significantly make use of visual speech information when available (Erdener, 2015).

When it comes to teaching Turkish as a foreign language, most available materials, like their counterparts in English and other languages, are in conventional formats such as printed books and audio material. Most material available in circulation and classrooms focus on grammar (Hengirmen, 2001; Arslan, 2011; Ketrez, 2012), or a combination of grammar, photographic and audio material (TÖMER, 2012). In the light of current research available, inclusion of dynamic, video material allows for better perception and production (e.g., Erdener & Burnham, 2005) of the material to be learnt.

Given the paucity of auditory-visual teaching material as well as the very limited research on Turkish, it is worth to think about what special aspects of Turkish we should study. A most difficult aspects of Turkish especially learners speaking Western languages as their native language is its quite complex morphological structure governed by saliently complex rules of agglutination.

As set out above, an initial auditory-visual speech perception study of Turkish as an L2 should focus on its unique and intricate morphological structure on both word and sentence levels (Erdener, 2016). This can and probably must be looked at in conjunction with its orthography characterised by very transparent morpheme-grapheme correspondences. The ease of its orthography, Turkish can be predicted to be taught via both auditory-visual and orthographic means in a parallel fashion when designing instruction material. Such points that are briefly listed

out here are, in fact, akin to unzipped files containing a plethora of tasks for both psycholinguists and foreign language instruction experts.

REFERENCES

Biersack, S., Kempe, V., & Knapton, L. (2005). Fine-tuning speech registers: a comparison of the prosodic features of child-directed and foreigner- directed speech. *Proceedings of Interspeech-2005*, 2401–2404.

Burnham, D., Kitamura, C. & Vollmer-Conna, U. (2002). What's new, pussycat? On talking to babies and animals. *Science, 296*, 1435.

Cotton, J. C. (1935). Normal visual hearing, *Science, 82,* 592-593.

Erdener, D. & Burnham, D. (2005). The role of audiovisual speech and orthographic information in nonnative speech production. *Language Learning, 55*(2), 191-228.

Erdener, D. (2015). The McGurk illusion in Turkish. *Turkish Journal of Psychology*, *30*(76), 19-27.

Erdener, D. (2016). Basic to applied research: the benefits of audio-visual speech perception research in teaching foreign languages. *The Language Learning Journal*, *44*(1), 124-132. DOI: 10.1080/09571736.2012.724080

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

Hengirmen, M. (2001). *Turkish grammar for foreign students. Yabancılar için Türkçe dilbilgisi.* Engin: Ankara, Turkey.

Ingvalson, E., McClelland, J. L., & Holt, L. L. (2011). Predicting native English-like performance by native Japanese speakers. *Journal of Phonetics, 39*, 571-584.

Ketrez, N. (2012). *A student grammar of Turkish.* Cambridge University Press: Cambridge, UK.

Lees, N. & Burnham, D. (2005). Facilitating speech detection in style!: the effect of visual speaking style on the detection of speech in noise. *Proceedings of AVSP 2005 International Conference on Auditory-Visual Speech Processing*, 23-28.

McGurk, H. & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264*, 746-748.

Ortega-Llebaria, M., Faulkner, A. & Hazan, V. (2001). Auditory-visual L2 speech perception: Effects of visual cues and acoustic-phonetic context for Spanish learners of English. *Proceedings of AVSP 2001 International Conference on Auditory-Visual Speech Processing*, 149-154.

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, 1–2*, 75–87.

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

TÖMER (2012). *Yeni Hitit: Yabancılar için Türkçe ders kitabı 1.* Ankara Üniversitesi Basımevi: Ankara, Turkey.

Uther, M., Knoll, M.A. & Burnham, D. (2007). Do you speak E-N-G-L-I-S-H? A comparison of foreigner- and infant-directed speech. *Speech Communication, 49*, 2-7.

Wang, Y., Behne, D., & Jiang, H. (2009). Influence of native language phonetic system on audio-visual speech perception. *Journal of Phonetics, 37*, 344-356.