Effects of Familiarity with Faces and Voices on Second-language Speech Processing: Components of Memory Traces

Debra M. Hardison

Department of Linguistics and Languages
Michigan State University, East Lansing, Michigan, USA
hardiso2@msu.edu

Abstract

Familiarity with a talker's voice and face was found to facilitate processing of second-language speech. This advantage is accentuated when visual cues are limited to either the mouth and jaw area, or eyes and upper cheek areas of a talker's face. Findings are compatible with a multiple-trace model of bimodal speech processing.

Index Terms: auditory-visual, memory traces, talker familiarity, language learners

1. Introduction

The effect of visual cues on speech perception by native listeners has a rich history beginning with the early McGurk effect experiments [1]. More recently, studies exploring the contribution of visual input for learners of English as a second language (L2) have shown that they can also experience the McGurk effect subject to influence from the first language (L1), L2 experience, information value of cues, and assumption of perceptual unity [2]. Visual cues such as a talker's lip movements can enhance the accuracy of segmental perception and word identification processes by nonnative as well as native listeners of English, the results of which reflect the considerable variability that exists across the auditory and visual dimensions of the speech event [3,4,5]. The characterization of speech perception as a context- and talker-dependent process has been supported by numerous studies demonstrating that listeners preserve details of a talker's voice in memory to facilitate subsequent processing of that individual's speech [6,7]. The current study explored the effects of familiarity with areas of a talker's face as well as the voice on spoken language processing by Korean learners of American English.

Auditory-visual (AV) perceptual training of Japanese and Korean learners of American English has shown superior effects compared to auditory-only (A-only) in the development of perceptual accuracy for sounds such as /r/, /l/, /f/ (vs. /p/), and /θ/ (vs. /s/), with generalization to novel stimuli and voices, transfer to production improvement [3][1], and to earlier word identification across styles of speech and talkers [4,5]. The temporal precedence of articulatory gestures over the associated acoustic signal gives visual cues a priming role in AV speech processing [8]. van Wassenhove et al. suggest that visual speech speeds up the cortical processing of auditory signals [9]. This priming role can be enhanced through training to facilitate L2 word identification [4]. Findings from these studies are compatible with episodic models of learning in which all attended details of a bimodal perceptual event are preserved in traces in long-term memory following multiple-trace memory theory [3,10].

Although a talker's lip movements are generally considered to be the primary source of linguistic information on the face, other areas may also contribute to the perceptual process. Eye-tracking data obtained from one Japanese and one English-speaking observer showed that they spent 45-70% of each presentation of stimuli in their respective L1s gazing at the talker's eyes [11]. The most common gaze sequence patterns were eye-to-eye and eye-mouth-eye. Vatikiotis-Bateson et al. suggested that phonetically relevant information may be distributed on the face beyond the area around the mouth as a result of changes in the orofacial muscles that accompany movement of the lips and jaw to produce speech sounds. They speculated that observers detect "well-learned, phonetically correlated events" (p. 938).

Eye-tracking data from a speechreading task involving native speakers of English showed the observers' eye gaze was mostly directed toward the talker's mouth [12]. Gazes in the middle region of the face (i.e., the cheeks and nose) accounted for 37% of the data – gazes that would have been assigned to one of the eye regions in the study by Vatikiotis-Bateson et al. Lansing and McConkie suggested that gaze directed toward the nose as a central feature might represent visual attention to the face as a whole. Perhaps listener-observers shift their gaze and attention strategically to other areas of the face to meet task demands.

Given that studies using McGurk effect stimuli, an approach which tags the input source as being from the auditory or visual modality, have shown significant effects for nonnative listeners of English, it is evident that they attend to the mouth area of a talker's face, even in the absence of training or substantial time in the L2 environment [2]. The current study was designed to investigate whether nonnative listeners attend to other areas of a talker's face preserving details of these areas in memory traces to facilitate subsequent processing of that person's speech in a word identification task in noise. If so, an AV presentation involving only a partial visual stimulus (e.g., the eye & upper cheek areas) of a familiar talker should serve as an effective memory probe to produce greater identification accuracy compared to a comparable stimulus from an unfamiliar talker, or an A-only stimulus from either a familiar or unfamiliar talker.

2. Experimental design

The experiment had two components: the study or familiarization phase (experimental groups only) and a test session (experimental and control groups). In the study phase, subjects viewed videotaped presentations of female native speakers of American English producing a series of words. They saw the entire face of the talkers. This was repeated for 10 days. In the test session, they were divided into four groups, each presented with a different stimulus condition in noise (-5 S/N ratio): AV-face (entire face visible), AV-mouth & jaw (see Appendix), AV-eyes & upper cheek areas, A-only (voice only, black screen). Some talkers in the test session were familiar to the subjects from the study phase and others were new.

A control group for whom all talkers were unfamiliar was assigned to each stimulus condition and participated only in the test session. Control groups were used to address the issue of the information value of the stimulus conditions independent of any possible carryover effects the experimental groups might have experienced from the input flooding over the 10-day study phase, which is not typical of the natural language environment for these subjects.

The within-subjects variable was talker (familiar, unfamiliar), and between-subjects variable was stimulus condition (AV-face, AV-mouth & jaw, AV-eyes & upper cheeks, A-only).

3. Method

3.1 Subjects

A total of 64 native speakers of Korean participated in this study: 32 were randomly assigned to the experimental groups, which participated in both the study phase and test session; 32 were assigned to the control groups, which participated only in the test session. Subjects were graduate students at a large Midwestern university in the U.S. who were not receiving English language instruction. Their aural/oral skills in English were assessed at the high intermediate range by the author. Subjects also reported their listening and speaking abilities in English were lower than their grammar and reading skills. They had been in the U.S. for 1-3 months at the time of the study. They reported no vision or hearing deficits.

3.2 Materials

There were two sets of stimuli: one for the study/familiarization phase and one for the test session. This approach required subjects in the experimental groups to generalize knowledge of the voices and articulations from the study phase to novel words presented in the test session, and avoided repetition effects, which were not a focus of the study. All stimuli were bisyllabic content words from a database used in previous studies, and were from low-density neighborhoods.[2] They had been rated as highly familiar (rating of 6 or 7 on a 7-pt. scale) by nonnative speakers at a lower level of English proficiency than those in the current study. All stimuli had initial-syllable stress. They were balanced across five visual categories based on the initial consonant: bilabial (/p, b, m/), labiodental (/f, v/), /r/, /l/, and nonlabial (e.g., /s, t, k/). Results of a previous gating study had revealed significantly earlier identification of words beginning with /r/ and /l/ compared to other visual categories in AV vs. A-only presentation for Korean learners of English [4]. In balancing the stimuli, consideration was also given to the amount of lip rounding in the production of vowels.

Stimuli were produced by a total of 12 different talkers, all female native speakers of American English from the upper Midwest of the U.S. Of the 12 talkers who were recorded, subjects saw 8 in the study/familiarization phase. Of those 8 talkers, 4 returned later in the test session where they served the role of familiar talkers, and 4 did not return. Subjects were not told in advance how many talkers they would see or given details about the subsequent test session. Talkers were selected in order to provide a range of articulatory styles for both the familiar and unfamiliar talkers.

In the study phase, a total of 280 words were presented (7 words x 5 visual categories x 8 talkers). Recordings were done in a sound-attentuated room using a Sony HI-8 videocamera, providing a full-sized image of the head with a fully visible jaw drop, and an Electrovoice lavaliere microphone. Talkers were instructed to look directly at the camera to produce each word. No specific articulation instructions were given. The objective was natural speech. The recordings were dubbed onto Betacam SP tape and then downloaded for editing using AVID Meda composer (MC8000) version 5.51 for MacIntosh and were digitized at a sampling rate of 44.1 kHz. All stimuli were tested in advance for auditory intelligibility by a native speaker group.

To create the stimuli for the two conditions in the test session that showed only parts of the face (i.e., AV-eyes & upper cheeks, AV-mouth & jaw), the picture-in-a-picture tool was used. Noise (-5 S/N ratio) was added to the test session stimuli in all conditions to reduce ceiling effects in the identification task. The S/N ratio was determined through pilot testing. Edited stimuli were recorded to broadcast-quality tape for presentation.

In the test session, there was a total of 200 words (5 words x 5 visual categories x 8 talkers). Stimuli and talkers were the same across conditions in the test session to control for articulatory differences and permit comparison of results across conditions.

3.3 Procedure

In the study/familiarization phase, 32 subjects (experimental group) were seated comfortably in front of a large TV monitor and presented with all stimuli audiovisually, showing the full face of the talkers. Subjects were asked to become familiar with the talkers they would see but were not given any further details in order to discourage them from focusing attention on particular areas of the face. The task was described to them as being comparable to trying to remember people you are introduced to so you can recognize them if you meet them later. Subjects viewed the study-phase talkers over a period of 10 days. Each day, they saw a different randomized order of talkers.

In the test session, eight subjects were randomly assigned to each of four stimulus conditions: AV-face, AV-eyes & upper cheeks, AV-mouth & jaw, and A-only. These stimuli were presented in noise at –5 S/N ratio. Subjects had not been told of the different conditions or the noise background in advance. They were given response sheets on which to write the word they thought the talker was producing.

Eight subjects who had not participated in the study phase were assigned as controls to each of the four stimulus conditions in the test session and were given the same task. Test session stimuli were randomized within each condition.

4. Results

The number of words correctly identified in the test session was tabulated per stimulus condition and talker type (familiar or unfamiliar).

Data for the experimental group, shown in Figure 1, were submitted to a mixed design ANOVA. Results revealed a significant main effect of talker, $F(1,28) = 30.175$, $p<.001$. Identification accuracy was greater overall for words produced by familiar talkers. This finding is consistent with the literature on auditory perception [6]. There was also a significant main effect of stimulus condition, $F(3,28) = 273.194$, $p<.001$. The order of conditions from highest to lowest in accuracy was: AV-face, AV-mouth, AV-eyes, and A-only. Tukey's post hoc tests indicated that all pairwise comparisons among the stimulus conditions were significant except the comparison of the AV-face and AV-mouth conditions. This suggested that seeing the mouth area of a talker was statistically as good as seeing the whole face in terms of word identification; however, this analysis was collapsed across the data from both familiar and unfamiliar talkers. All AV conditions, including AV-eyes & upper cheeks, produced significantly better word identification accuracy than A-only.

Importantly, there was a significant Talker x Stimulus Condition interaction, $F(3,28) = 27.479$, $p<.001$. The data suggested that the advantage of a familiar talker was particularly evident for the conditions in which a partial visual stimulus was available (i.e., the mouth & jaw, and the eye & upper cheeks).

Therefore, a separate one-way ANOVA was conducted on the identification accuracy scores for words produced only by unfamiliar talkers across the four conditions in the test session. Results revealed a significant effect of stimulus condition, $F(3,28) = 376.424$, $p<.001$. Tukey's post hoc tests indicated a significant difference between all the conditions except the comparison between AV-eyes and A-only. Comparison of the above analyses indicated that the addition of visual cues involving the eye and upper cheeks of a talker's face produced significantly greater identification accuracy than the A-only presentation, but only when these cues belonged to a familiar talker. In addition, the equivalence of the AV-face and AV-mouth conditions suggested by the earlier analysis was further clarified as being contingent upon talker familiarity.

Results of the analysis of the control group data paralleled the above findings for the unfamiliar talker data produced by the experimental group. Post hoc results confirmed that if the talkers were unfamiliar to the subjects, there was no significant difference in word identification accuracy between the A-only and AV-eyes conditions. This analysis also indicated that the study phase itself had not biased the experimental groups' responses.

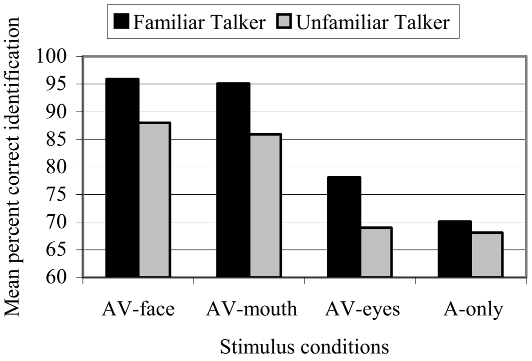

Figure 1. *Experimental groups mean percent correct word identification accuracy: Stimulus Condition x Talker Type*

5. Discussion

This study investigated the effects of familiarity with areas of a talker's face as well as voice in spoken word identification by nonnative speakers of English. The results are compatible with previous studies showing the contribution of visual cues to spoken language processing by Korean and Japanese learners of English. Taken together, all of these findings support the hypothesis that listeners preserve details of a talker's face and voice in long-term memory, and that this information facilitates subsequent processing of that talker's speech.

Following multiple-trace memory theory, a retrieval cue or probe – an active representation of an experience in primary memory – contacts in parallel all stored traces in secondary or long-term memory, simultaneously activating each one and all of its properties, with the contribution of each weighted according to its similarity to the features of the probe or cue [3,10]. The probe is said to return an echo to primary memory. The echo can enhance a probe's representation by filling in missing details, thereby facilitating the association of a subsequent probe with previously stored information. In the case of the findings of the present study, a visual cue composed of a familiar but only partial stimulus, such as a talker's mouth or eyes, can be enhanced by echoes from stored representations of the talker's entire face in memory to facilitate a task such as bimodal word identification. Subjects who participated in the study phase would have been able to store such representations from the 10 days of AV presentation of the talkers' entire faces and voices producing associations of the visual and auditory components of the speech events.

In the present study, findings did not support the notion that the eye and upper cheek areas of the face contribute significant linguistic information independent of talker familiarity. Examination of the AV-eyes stimuli in the present study with the audio turned off revealed that a couple of the talkers showed some visible movement in the muscles from time to time around the eyes; however, in the majority of cases, even the author who was very familiar with these talkers could not always determine when they were speaking or what they were saying based solely on these visual stimuli. It is likely that there is considerable variability across talkers in the information value of facial areas beyond the mouth just as there is in the discernibility of their articulatory gestures and the intelligibility of their voices. It remains a question as to how much exposure to a face observers need in order to learn visual-phonetic associations.

6. Notes

1 There is also variability in the findings of studies comparing AV and A-only training of L2 learners. Hazan et al. [13] did not find an advantage for Spanish speakers trying to identify British English /p, b, v/ embedded in nonsense words. However, comparability across studies is substantially reduced by differences in subject and stimulus factors, and numerous methodological differences.

2 The data for neighborhood density were obtained from the Speech Research Laboratory, Department of Psychological and Brain Sciences, Indiana University, Bloomington, IN 47405, USA.

7. References

[1] McGurk, H., and MacDonald J., "Hearing lips and seeing voices", Nature, 264:746-748, 1976.

[2] Hardison, D. M., "Bimodal speech perception by native and nonnative speakers of English: Factors influencing the McGurk effect", Lang. Learn., 49:213-283, 1999.

[3] Hardison, D. M., "Acquisition of second-language speech: Effects of visual cues, context, and talker variability", Appl. Psycholing., 24:495-522, 2003.

[4] Hardison, D. M., "Second-language spoken word identification: Effects of perceptual training, visual cues, and phonetic environment", Appl. Psycholing., 26:579-596, 2005.

[5] Hardison, D. M., "Variability in bimodal spoken language processing by native and nonnative speakers of English: A closer look at effects of speech style", Speech Comm., 46:73-93, 2005.

[6] Nygaard, L. C., Sommers, M. S., and Pisoni, D. B., "Speech perception as a talker-contingent process", Psycholog. Science, 5:42-46, 1994.

[7] Johnson, K., and Mullennix, J. W., Eds., Talker Variability in Speech Processing, Academic Press, San Diego, 1997.

[8] Munhall, K. G., and Tohkura, Y., "Audiovisual gating and the time course of speech perception", J. Acoust. Soc. Amer., 104:530-539, 1998.

[9] van Wassenhove, V., Grant, K. W., and Poeppel, D., "Visual speech speeds up the neural processing of auditory speech", PNAS, vol. 102, no. 4, 2005, pp. 1181-1186.

[10] Hintzman, D. L., " 'Schema abstraction' in a multiple-trace memory model", Psycholog. Review, 93:411-428, 1986.

[11] Vatikiotis-Bateson, E., Eigsti, I-M., Yano, S., and Munhall, K. G., "Eye movement of perceivers during audiovisual speech perception", Percept. & Psychophys., 60:926-940, 1998.

[12] Lansing, C. R., and McConkie, G. W., "Attention to facial regions in segmental and prosodic visual speech perception tasks", J. Sp. Lang. Hrng. Res., 24:526-539, 1999.

[13] Hazan, V., Sennema, A., and Faulkner, A., "Audiovisual perception in L2 learners", Proc. ICSLP 2002, 1685-1688.

8. Appendix

The figure below is a cropped image of one of the frames from the AV-mouth & jaw stimulus condition in the present study. The talker is about to say the word *lighter*.