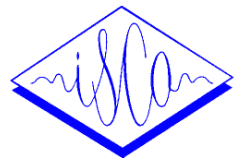

Auditory-Visual Speech Processing
(AVSP'98)
Terrigal – Sydney, Australia
December 4-6, 1998

REPEATING AND REMEMBERING FOREIGN LANGUAGE WORDS: DOES SEEING HELP?

Chris Davis [1] *& Jeesun Kim* [2]

[1]The University of Melbourne

Parkville, Vic, 3052, AUSTRALIA

[2]The University of New South Wales

Sydney, NSW, 2052, AUSTRALIA

ABSTRACT

Normal hearing people use lip-reading when listening conditions are not good [1]. Yet even when the listening environment is ideal, lip reading can help with a difficult signal such a listening to a foreign language [2]. These authors demonstrated that non-native French student's shadowing performance of French was improved by seeing the lips and mandible of the speaker. This observation suggests that learning the sounds of a foreign language may be aided by audio-visual presentation compared with audio alone. One recent example of such an application of the audio-visual approach has been in teaching a foreign alphabet by children [3]. The current experiment extended this approach by examining whether the mode of presentation affected the accuracy of repetitions of short phrases of a language participants had not heard before (Korean). Participants either heard a (five syllable) Korean phrase while watching the top part of face (no lips or jaw) or heard the phrase while watching the lips and jaw of the speaker. Three native speakers (blind to the presentation status of the participant) judged the accuracy of the participants' subsequent rendition of the phrase. The experiment also examined whether presentation mode affected performance on a subsequent old/new recognition task of the experimental phrases. The results are discussed in relation to the relative contribution of auditory and visual information in L2 acquisition in the immediate and longer term.

1. INTRODUCTION

There are a number of challenges that face a second language (L2) learner. These may range from the difficulties associated with trying to master high-level meta-linguistic skills such as rules of syntax or derivational morphology to the more basic ones associated with the reception and production of a new sounds.

Although it is not the intention of this article to enter the debate about how language learning can best be characterized, it is worth noting that low-level phonological information has an important part to play in spoken word recognition. Further, recent models of information processing provide intelligible accounts how the learning of phonological patterns may bootstrap higher-level linguistic representations [see for instance, 4].

However, as an adult, learning non-native phonemes is not that easy, as it has been amply demonstrated that when listeners are presented with speech stimuli from phonetic categories not used in their own language they typically show worse performance on discrimination tasks (sometimes at chance levels) than native speakers of the language from which the phonemes were chosen [e.g., 5].

What might assist an adult in acquiring the sounds of a second language? The idea that this article will pursue is that the optimal learning situation is one that provides the hearer-learner with a range of identification cues. That is, consider face to face communication, it provides multiple language recognition cues that the speaker/hearer potentially could make use of in understanding. For instance, the production of speech involves a series of correlated orofacial movements ranging from vocal-tract gestures that also alter jaw position and cheek shape, to movements of the tongue and lips [see 6].

Indeed, it has been argued that all available visual and auditory cues are employed in understanding speech [e.g., 2].

There are, in fact, numerous demonstrations that incorporating orofacial movement information benefits speech perception. For example, hearing people use lip-reading when listening conditions are not good [1]. Indeed, it has been shown that the accuracy of automatic machine speech recognition is improved even when only lip-mouth shape information is used in conjunction with a more standard audio only system [7].

Many of the studies that have shown that people do better on a range of auditory identification measures when they can see the speaker have used signals that are difficult to hear [e.g., 1, 10]. However, even when the listening environment is ideal, it turns out that lip reading appears to help with a difficult signal such as listening to a foreign language. For example, [2] have demonstrated that non-native French student's performance in shadowing French was considerably improved by seeing the lips and mandible of the speaker. This observation suggests that learning the sounds of a foreign language may be aided by audio-visual presentation compared with audio alone. In a more recent study, [8] have also shown that visible speech assists people to remember spoken words, with participants in their study recalling on average two more words from a 16 word sentence when they could see the speaker.

The following study will combine aspects of the [2] and [8] studies by examining both the production and the recognition of foreign sounds with and without the accompanying visual presentation. In conducting such an experiment a number of issues arise concerning the appropriate design: How best to test a participant's production of foreign sounds? The approach adopted in [2] required that participants shadow foreign sentences and their performance was rated by a native speaker. This task is appropriate with multi-word sentences using a language that participants are familiar with (e.g., those that participated in [2] had 4.4 years of tuition in the language). However, in the current experiment we are interested in whether the results of [2] extend to a language that is unknown to the participant. Further, it also seems to us that the shadowing response may be handled through echoic memory and so may not involve sufficient perceptual analysis to create a more permanent and flexible memory code. Thus, we decided to employ a task where a short phrase is presented three times with the participant required to repeat it after the end of the third presentation.

Another issue concerns how to test the participant's memory of the phrases. One of the few studies that has examined the effects of visible speech on memory [8] used a recall measure to assess memory for 16 word sentences in the participant's native language. However, in the current experiment, the stimuli (Korean phrases) would be very difficult for participants (who do not know this language) to recall. We therefore adopt the less demanding memory measure of a forced-choice recognition procedure.

One final issue concerns the nature of the visual information that is provided to the participants. Precisely what and how much of the face should we show? A recent study [9] has reported that performance was better on a lip-reading task when the speaker's jaw movements were also shown. It was therefore decided to show the entire bottom part of the speaker's face, indeed, we included all of the 11 EMG insertion sites that [6] used to demonstrate that the face indexes the basic spatio-temporal behavior of the vocal tract. Following [2], we employed a second visual condition designed to show that it is the orofacial gestures that improve task performance rather than the visual stimulus merely serving to focus the participant's attention on the task. To this end, the second visual condition showed the top part of the speaker's face (above the 11 sites of [6]).

2. METHOD

2.1. Participants

Ten graduate and undergraduate native speakers of English participated in the experiment.

2.2. Materials and design

The stimuli employed in the experiment consisted of 30 short phrases in Korean spoken by a female Korean native speaker. All the items were five syllables but varied between 2 and 3 morphemes. Each duration approximately 1 - 1.2 seconds. The female speaker was positioned at a fixed distance from the camera and recorded against a blank background. The speaker was instructed to keep whole head movements to a minimum during the recordings.

These utterances were recorded using a Sony video camera and converted to digital format by a Video Blaster FS200 video capture card at a sampling rate of 30 Hz. The resulting avi format files were imported into a Macromedia® director project and a short program was written in 'lingo' to present them. Only the top (eyes and above) or bottom (cheeks to jaw) of each video was presented. The frame of the presented video measures 110 cm x 45 cm with the image taking up on average 70% of this area.

The experiment consisted of two phases. In Phase 1 the participant was presented with 16 video and audio sequences (listening to the audio on a pair of headphones). Half the video sequences showed the top of part of the speaker's face (eyes and above) and half showed the bottom half. Two stimulus lists were constructed so that each video and audio sequence appeared in Phase 1 but no participant heard the same sequence twice.

In Phase 2, participants were presented with the 16 sequences that they had seen before plus 14 new sequences.

2.3. Procedure

Each participant was tested individually. The stimuli were presented on a 350cm video monitor by a Pentium MMX PC equipped with a 4 mg video card and 32 mg ram. Participants were given a standard set of instructions that explained that in Phase 1 they would be required to repeat a phrase spoken in Korean. It was explained that along with the spoken phrase they would also see a video that either showed the top or bottom part of the speaker's face. They were told about Phase 2 (the recognition test) only after they had finished phase 1. The auditory stimuli were presented to participants via a set of Sony DR-S7 headphone; they spoke their response into a Pro.2 DM 320 microphone.

Phase 1 began with a single practice item, showing the bottom part of the speaker's face, saying "an example" and they were told that this English example allowed them to get familiar with the experimental set up and so that the volume of the recording could be adjusted for comfort for each participant. After each presentation two screen button were presented that allowed the participant to either *repeat* the trial or get the *next* item. The participant was informed that they should present the stimulus 3 times after which they must repeat the phrase. Each participant received a different random order of stimuli and their renditions were recorded on JVC (KD 1136) cassette recorder.

Phase 2 commenced after a two minute break (in which the instructions were read). The participant was informed that they would be required to judge whether the spoken phrase was one that they had heard in phase 1. Participants signaled their judgment by depressing the left mouse screen button (labeled "old") that appeared after each video and auditory sequence was presented. The participant made a go-no go response, i.e., only pressing the button if they thought that the item was an old one. Response latency and error rate were collected but the participant did not receive this feedback.

2.4. Scoring of utterances

Judges were first familiarized with the phrases that would be presented. Scoring of the participant's repetition of the phrase was based on a 1 to 7 point scale of "goodness" of utterance. One was scored if the Korean judge could not recognize the utterance and 7 was awarded if the utterance was a clearly recognizable accurate rendition. After the judge made this rating they then indicated the number and position of the syllables that they thought were recognizable.

3. RESULTS

The mean ratings for the goodness of the utterance score for the two presentation conditions (top and bottom of face shown) are presented in Figure 1. As can be seen, the ratings of utterances where the participant could see the bottom part of the speaker's face was higher (i.e., a more accurate rendition) than when they could not. Two analyses of variances (one based on participants, the other on items) were conducted to determine whether this difference was significant, it was $F_1(1,8) = 13.84$, $p < 0.01$; $F_2(1,14) = 6.02$, $p < 0.05$.

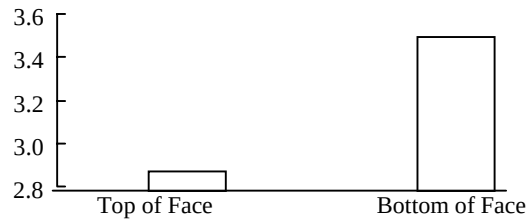

Figure 1: Mean goodness of utterance scores (1 [poor] - 7 [good]) for each of the two presentation conditions (top and bottom of face shown)

In addition to the ratings that judges made regarding the overall goodness of the participant's rendition, a determination of the number correctly pronounced syllables was made. The mean number of syllables that were rated as correct for the two presentation conditions are shown in Figure 2. As would be expected, given the results of the goodness of utterance ratings, participants got more syllables correct when they could see the lower part of the speaker's face, $F_1(1,8) = 17.63$, $p < 0.01$; $F_2(1,14) = 9.98$, $p < 0.05$.

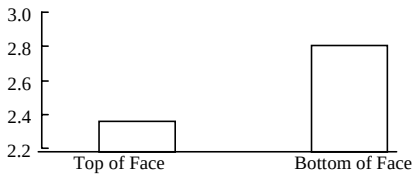

Figure 2: Mean number of syllables rated as correct (/5) for the two presentation conditions (top and bottom of face shown).

The mean number of correctly recognized spoken phrases for each of the two presentation conditions (top and bottom of face shown) are presented in Figure 3. Once again, participants did better when they could see the lower part of the speaker's face. That is, when people could see the lower part of the face they correctly recognized significantly more spoken phrases, $F_1(1,8) = 12.52$, $p < 0.01$; $F_2(1,14) = 4.71$, $p < 0.05$.

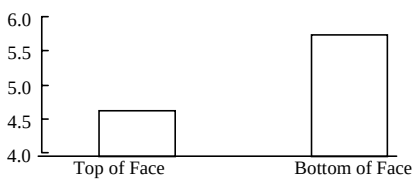

Figure 3: Mean numbers of correct "old" decisions (/8) as a function of type of presentation (top or bottom part of face shown)

The mean reaction times for correct "old' responses as a function of display condition are shown in Figure 4. Although participants were on average 100 ms faster in making their response when they could see the speaker's face compared to when they could not this difference was not significant, both F's <1. This lack of a significant effect was probably due to the substantial variance associated with the latency data, i.e., the standard deviation was more than half the mean for the 'top of face" condition and a third of the mean latency value for the "bottom of face" condition.

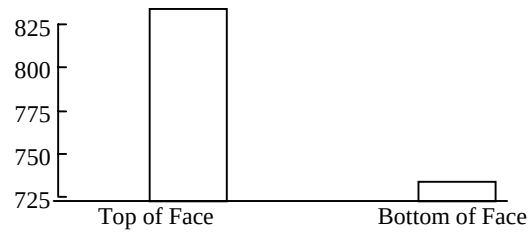

Figure 4: Mean latencies in milliseconds for correct old recognition responses as a function of the two presentation conditions (top and bottom of face shown)

4. DISCUSSION

The results have shown that the perception and repetition of unknown speech sounds was augmented by visual information of the tongue, teeth, lips, jaw and cheeks of the speaker. Further, it was shown that this visible information assisted the subsequent recognition of these sounds. These findings are consistent with the work of [2] and [8] that showed that the shadowing of foreign speech sounds and participant's memory of spoken sentences were improved with visible speech, extending them to a situation where the speech sounds to be learned are totally unfamiliar.

These findings suggest that the perception and memory of foreign speech sounds may be usefully enhanced by the addition of the speaker's face. The addition of this information not only provides additional sources of constraint about the identity of a particular sound, but also may possibly encourage the creation of a more procedural memory trace based upon an attempted mirroring of the speaker's articulatory gestures.

The current demonstration of the effectiveness of providing information about the movement of the speaker's face along with purely auditory information, suggests that language tuition could usefully utilize audio-visual techniques. Indeed, it may be that the development of a user interface such as the "talking head" [11], along with automated speech recognition and lip-tracking technology,

could provide an automated second language learning system where a visual tutor could demonstrate how an utterance is generated and monitor how the learner imitates this.

5. REFERENCES

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

2. Reisberg, D, McLean, J., & Goldfield, A. (1987). Easy to hear to understand: A lip-reading advantage with intact auditory stimuli. In B.Dodd and R. Cambell, (Eds.) Hearing by eye: The psychology of lip-reading. London: Lawrence Erlbaum, pp 97 -113.

3. Rahman M.D.S., & Uddin N.H.M.S .(1998). A Pc-Based Audio-Visual Alphabet Learning System. International Journal of Electrical Engineering Education. 35(1), 37-46.

4. Cairns, P., Shillcock, R., Chater , N., & Levy, J. P. (1995). Bottom-up connectionist modelling of speech. Chapter 15. In Connectionist models of memory and language. J.P Levy, D Bairaktaris, JA. Bullinaria & P. Cairns (Eds.). UCL Press: London.

5. Werker, J. .F. & Logan, J. (1985). Phonemic and phonetic factors in adult cross-language speech perception. Perception and Psychophysics, 37, 35-44.

6. Munhall, K. G. & Vatikiotis-Bateson, E. (1998) The moving face during speech communication. In R. Cambell, B. .Dodd, & D. Burnham (Eds.), Hearing by eye, Part 2: The psychology of speechreading and audiovisual speech. London: Taylor & Francis, Psychology Press.

7. Silsbee, P. L & Bovik, A. C. (1996). Computer lipreading for improved accuracy in automatic speech recognition, IEEE, Transactions on speech and audio processing, 4, 337-351.

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

9. Guiard-Marigny, T., Ostry, D .J., & Benoît, C. (1995). Speech intelligibility of synthetic lips and jaw. Proceedings of the 13th International Congress of Phonetic Sciences (3) (pp. 222 - 5). Stockholm, Sweden.

10. Sanders, D. & Goodrich, S. (1971). The relative contribution of visual and auditory attention. Journal of Speech and Hearing Research, 14, 154-159.

11. Cohen, M. M., & Massaro, D. W. (1994) Development and Experimentation with Synthetic Visible Speech Behavioral Research Methods and Instrumentation, 26, 260-265