Cognitive Constraints on the Use of Visible Speech and Gestures

Laura A. Thompson

Box 3001/Dept. 3452/Psychology
New Mexico State University
Las Cruces, NM 88003

thompson@crl.nmsu.edu

Abstract

Gestures and visible speech cues are often available to listeners to aid their comprehension of the speaker's meaning. However, visual communication cues are not always beneficial over-and-above the audible speech cues. My goal is to outline several types of constraints which operate in the human cognitive processing system that bear on this question: When do visual language cues (visible speech and gestures) provide an aid to comprehension, and when do they not? Research on visual-spoken language comprehension carried out in my lab over recent years is described and recommendations will be made concerning the design of multi-modal interfaces.

Introduction

Due to the enormous variety of visual cues embodied in articulatory movements of the face (visible speech), in emotional facial expressions, in body language, and in the plethora of different classes of gestures, face-to-face communication involves a richness that extends far beyond spoken language. Much research in the area of auditory-visual speech perception has shown that people, instead of becoming overloaded with multiple language cues, make efficient use of them. In a recent word recognition experiment reported by Cohen and Massaro (1993) for example, participants' recognition of natural speech increased from 4% given just visible speech, to 55% given just audible speech, to 72% given both auditory and visible speech. Not only does visible speech improve comprehension, but it is very difficult to ignore after years of cross-modal pattern learning. Thompson and Lee (1996) presented auditory-visual speech syllables to individuals for classification in two conditions. In one condition, participants reported what they ***heard*** the speaker say. In the other condition, participants reported what they ***understood*** the speaker to have said, which cued them to report their "global impression" of the bimodal speech. Because the amount of influence of the visual source was the same in both conditions, this was evidence supporting the mandatory integration of the two sources of information during speech recognition.

Visible speech aids the comprehension of speech even when the speech signal is intelligible and the listener's hearing is normal, although it has been found to be especially influential when either the speech signal is degraded (e.g. noise, low bandpass filtering), or the listener's ability to interpret the signal is diminished (e.g. hearing impairment). In addition, research has shown that visible speech improves comprehension when the speaker has a heavy foreign accent (Reisberg, McLean, and Goldfield 1987).

Research has shown that people are flexible in their ability to process less-than-lifelike displays as though they were encoding real visible speech. For example, Massaro and Cohen (1990) and Thompson (1995) found that the degree of influence of synthesized visible speech on syllable identification responses is the same as that observed with videotaped images of real human speakers. Moreover, effects similar to processing the entire facial display can be seen when only the bottom half of the face is presented (e.g. Montgomery and Jackson 1983; Rosenblum and Saldana 1996). Surprisingly, the image does not even need to contain visual facial features, as the work of Rosenblum and Saldana (1996) shows, using moving point-light displays. The latter type of display offers the exciting possibility that point-light images can be modelled with just a few parameters on a computer, or transmitted over low bandwidth telephone lines, due to the minimal amount of information contained in this type of display.

In all the examples thus far, participants were tested using just single syllables, presented one at a time. As it turns out, when comprehension is tested on longer speech samples, the results do not always show a positive benefit of visible speech to comprehension. In the following, I describe different types of task situations that sometimes do, and sometimes do not, result in better comprehension with the inclusion of visual language information. These results show that the user's ability to profit from visible speech and gestures depends on the following cognitive constraints: working memory capacity, attentional focus, knowledge of the meaning of certain representational gestures, and age and individual variability.

Working Memory Capacity

We have conducted extensive investigations on memory for sentences containing visible speech, and also representational gestures. In one experiment, 9-year-old children were compared to adults (Thompson, Driscoll, and Markson in press). No differences were found between the two age groups in their ability to profit from visible speech. However, the aid to recall provided by gestures was substantially less than that provided by visible speech.

Interestingly, due partly to working memory capacity limitations, older adults' comprehension of language was not aided by the inclusion of representational gestures, while younger adults' comprehension was influenced by gestures (Thompson 1995; Thompson and Guzman forthcoming).

When auditory-visual speech is clearly presented without any extraneous demands on processing, older adults are particularly dependent on visible speech cues, even more so than young adults (Thompson 1995). However, when the attention demands of the listening task are considerable, older adults are unable to encode the visible speech cues and gestures (Thompson and Guzman forthcoming). Thus, visual information may not be encoded when working memory is taxed to its functional limits.

Attentional Focus

Another cognitive factor which could determine the extent of influence of visible speech is the listener's spatial attentional focus on the speaker's face. We have conducted a preliminary study to discover the spatial characteristics of attention distribution across the speaker's face (Thompson, Goodell, and Boring forthcoming). Videotaped reenactments of speeches by two famous public speakers were used as materials. The speeches were divided into segments. At the end of each segment, a pattern of dots appeared in one, two, or three of four positions on the face for either 16 or 33 milliseconds. The positions were in the center of the forehead directly above the eyes, on top of the left ear, on top of the right ear, and directly below the mouth. One of the participants' tasks was to record, onto caricatured faces on paper, the dots they had seen in the immediately preceding videotaped segment. To help ensure that participants concentrated on the content of the speaker's discourse, participants also had to complete a comprehension task related to the immediately preceding speech segment. The results were dramatic: compared to young adults, older adults' attentional focus was on the mouth, at the expense of attending to the other areas of the face, while younger adults did not show differential attentional focus to the various regions of the face.

Knowledge of Gesture Meaning

In a recent experiment, we found that representational gestures, gestures representing actions and attributes of objects or characters, contributed directly to the meaning of utterances (Thompson, Driscoll, and Markson in press). This is a controversial issue because the limited number of reported studies on the topic has yielded contradictory results. We compared memory for phrases (noun phrases versus predicate term phrases) containing a representational gesture to the same type of phrases which did not contain gestures. Adults' memory for the words in both noun and predicate term phrases was better when the words were accompanied by gestures. Nine-year-old children's memory was better when gestures appeared with predicate terms, but not when gestures accompanied nouns. We reasoned that the mapping between actions communicated by predicate term gestures and their verbs was more direct and easily interpretable compared to the mapping between nouns and gestures. Children might be especially susceptible to a processing advantage involving fewer steps of interpretation. However, as they gain world knowledge of the meanings communicated behind even the most obtuse of gestures, they stand a greater likelihood of incorporating these gestures into their understanding of the speaker's meaning.

Age and Individual Variability

Some of my research has shown a developmental trend toward greater influence of gesture and visible speech information across childhood (e.g., Thompson and Massaro 1994). As children get older, they seem better able to incorporate the multiple cues to meaning which are available for them to use.

We have found wide variability between adults in the amount of influence of visible speech in auditory-visual speech perception. In one experiment, participants watched a videotape of a speaker clearly articulating one of two syllables, /ba/ or /da/ (Thompson 1995). Their task in one condition was to discriminate the two visible speech tokens (to lipread). In the other condition, synthesized speech tokens ranging along a five-step speech continuum going from a clear /ba/ to a clear /da/ were dubbed onto the two visible speech tokens, and their task was to report what they "understood". In two adult age groups, lipreading ability in the visible-speech alone condition was significantly correlated with the extent of influence of visible speech in the auditory-visual condition. Further, two groups of participants emerged: those that were "visually-oriented" and those that were "auditorially-oriented" when the cues were ambiguous.

Implications

Most research conducted on human participants supports the general recommendation that computer systems which present speech to the user should conjointly present visible speech (Thompson and Ogden 1995). However, the narrow set of experimental conditions used to support this recommendation must be acknowledged. First, the task demands placed minimal cognitive processing requirements on the individuals being tested. Second, the participants in the experiments were usually young adults who are at their prime, cognitively speaking. Third, individual differences were ignored. Our research demonstrates the importance of taking into consideration many cognitive constraints which are revealed with lengthier speech samples embedded into more complicated task contexts. In so doing, it is apparent that gestures and visible speech can aid comprehension, but only when: (a) the users' working memory capacity is sufficiently strong to incorporate visual language, (b) the users' spatial attentional focus includes the mouth region of the face, and (c) the user has gained a certain amount of experience interpreting the meanings presented visually.

Being aware of cognitive constraints in the use of visual language can help facilitate intelligent design of multi-modal human-computer interaction. More specific recommendations will be m ade at the workshop.

References

Cohen, M. M., and Massaro, D. W. 1993. Modeling coarticulation in synthetic visual speech. In N. M. Thalmann and D. Thalmann (Eds.), ***Models and techniques in computer animation,*** (pp. 139-155). New York: Springer-Verlag.

Montgomery, A. A., and Jackson, P. L. 1983. Physical characteristics of the lips underlying vowel lipreading performance. ***Journal of the Acoustical Society of America 73: 2134-2144.***

Reisberg, D., McLean, J., and Goldfield, A. 1987. Easy to hear but hard to understand: A lipreading advantage with intact auditory stimuli. In B. Dodd and R. Campbell (Eds.), ***Hearing by eye: The psychology of lipreading.*** Hillsdale, NJ: Erlbaum.

Rosenblum, L. D., and Saldana, H. M. 1996. An audiovisual test of kinematic primitives for visual speech perception. ***Journal of Experimental Psychology: Human Perception and Performance*** 22: ***3 18-33*** 1.

Thompson, L. A. 1995. Encoding and memory for visible speech and gestures: A comparison between young and older adults. ***Psychology and Aging*** 10: 215-228.

Thompson, L. A., Driscoll, D., and Markson, L. in press. Memory for visual-spoken language in children and adults. ***Journal of Nonverbal Behavior.***

Thompson, L. A., Goodell, N., and Boring, R. ***Where is the listener s attentional focus on the speaker s face?*** Forthcoming.

Thompson, L. A., and Guzman, F. 1998. ***Some limits on encoding visible speech and gestures using a dichotic shadowing task.*** Forthcoming.

Thompson, L. A., and Lee, K. 1996. Information integration in cross-modal pattern recognition: An argument for acquired modularity. ***Acta Psychologica 92: 79- 104.***

Thompson, L. A., and Massaro, D. W. 1994. Children's integration of speech and pointing gestures in comprehension. ***Journal of Experimental Child Psychology 57: 327-354.***

Thompson, L. A., and Ogden, W. C. 1995. Visible speech improves human language understanding: Implications for speech processing systems. ***Artificial Intelligence Review 9: 347-358.***