Visible Language in Speech Perception: Lipreading and Reading

Dominic W. Massaro, Michael M. Cohen, and Laura A. Thompson

Program in Experimental Psychology, University of California, Santa Cruz
Santa Cruz, CA 95064

Visible Language XXII, 1
Dominic W. Massaro, Michael M. Cohen, and Laura A. Thompson, pp. 8–31
Visible Language, Rhode Island School of Design
Providence, RI 02903

Watching a speaker in face-to-face communication can influence what the perceiver hears the speaker saying. Faced with this influence of visible language on the perception of audible language, an interesting question is whether written language would also influence audible speech perception. To test this possibility, subjects identified spoken syllables either while viewing the speaker's face or while reading a written syllable. In both conditions, subjects identified what they heard the speaker saying. Replicating previous studies, lipreading had a large influence on the identification. In contrast, reading a written syllable had a much smaller, but statistically significant effect. A fuzzy logical model of perception accounted for both the lipreading and reading contributions to speech perception. A model assuming that the reading contribution was due to a post-perceptual bias gave a poor description of the results. Although lipreading appears to be much more influential than reading, it remains a possibility that written language can contribute to our auditory experience of speech.
Speech Perception

Although speech perception is usually thought of as an auditory process, it appears to be visual as well. As exemplified by the speed of the voice and the visual aspect of the speaker's movements, vision and speech are co-occurring events in the real world. Visual information can enhance understanding, especially when the auditory signal is degraded by masking noise. This idea was originally proposed by Puthoff (1964) then extended by Massaro and Cohen (1983) to the idea of incorporating the visual aspect of speech into the auditory recognition process. The visual influence is not limited to situations where degraded auditory inputs. As shown by McCorkle and MacFarland (1983) the visual input from the speaker can change the perception of an auditory speech event. Using videotape, these investigations demonstrate that speech and visual cues are integrated into the visual information, and that visual cues can affect the perception of speech.

Massaro and Cohen (1983) extend the McCorkle and MacFarland (1983) demonstrations by independently varying auditory and visual information in a factorial design. Subjects heard a list of 'real' speech events consisting of high-quality synthetic syllables ranging from /ba/ to /da/ and contrasted with a videotaped /ba/ or /da/ at no articulation. Although subjects were instructed specifically to repeat what they heard, viewing the visual articulation mode is a large contribution to identification. The results in Figure 1 show how often visual and auditory information are separate or co-occurring in the visual information. The contribution of one source is larger than the other source in a situation where there is no articulation. For example, the magnitude of the visual effect is smaller at the ambiguous end of the auditory continuum than in the neutral region of the continuum. The tests of quantification models provide evidence for the integration of visual and auditory information in the perception process.

The results in Figure 1 are uniquely described by a fuzzy logical model of perception (FLEM). According to the FLEM, perception is carried out in three steps. The first step is the recognition of the speech event, the second step is the interpretation of the speech event, and the third step is the decision-making process. The FLEM model is based on the notion that there are multiple interpretations of speech events, and that the decision-making process is based on the integration of all possible interpretations.

Figure 1 Auditory

Stage 1: Feature Evaluation, during which the stimulus is categorized by the sensory system. The features are assumed to be continuous rather than discrete. The outcome of feature evaluation represents the degree to which each relevant feature is present in the speech stimulus. The degree of presence of a feature is represented as a truth value between 0 and 1. The second stage of recognition is prototype matching, which involves the integration of the features. During this stage the feature information is compared with prototype definitions to determine if the current feature matches the prototype. The third stage of recognition is pattern identification. During this stage the recognition of a prototype is evaluated relative to the summary statistics of all potential prototypes.
Auxiliary information is assumed to be transduced and the output of auditory feature detectors are stored in a perceptual acoustic storage (PAS). In Crowder's revised model, both the visual and auditory consequences of speech provide functional information at the level of PAS. Supposedly, auditory feature selection can occur even in the absence of sound, as in pure lipreading.

Equation 4

The prediction of the matching parameter for each unique pair of auditory and visual features, Meier and Liberman (1982), defined the levels of the auditory stimuli with three levels of the visual stimuli. In the experimental condition of the figure, the stimulus predictions for three auditory and three visual levels of A and three levels of V, the predictions of the auditory require 12 parameters (nine q values and three r values). The quantal predictions of the TNC were compared for the observed proportion of 0.61 to the model using the parameter estimation program STORM (Schandler, 1968). A model is represented as a set of prediction equations and a set of unknowns. The goal of STORM is to find a set of parameters that optimizes the predictions of the observed data. Initially, all parameters are set to 0. By iteratively adjusting the parameters of the model, STORM minimizes the squared deviations between the 27 observed and 27 predicted points. As can be seen in the figure, the predictions of the model give a good description of the results. In addition, the description of each subject's performance was significantly better than in a model assuming discrete rather than continuous features of a model with nonindependent features.

An alternative account of bilingual speech perception is proposed by Crowder (1983) who modified his 1978 model to account for the integration of visual information in speech perception. Auxiliary information is assumed to be transduced and the output of auditory feature detectors are stored in a perceptual acoustic storage (PAS). The primary evidence for PAS has been a suffix effect, which occurs when an auditory speech stimulus following an auditory memory list and includes all of the last item on the list. A pure language suffix is a meaningful but meaningless suffix does not produce similar interference. These results suggest evidence for an auditory representation that has specific sensory characteristics. Since publication of Crowder's (1983) model, however, Spurlock and Liberman (1984), Campbell and Liberman (1982), and Creamer and Crowder (1984) have shown that matching sound also articulates the suffix in something the suffix simply cannot do: it produces a suffix effect. This result appears to suggest that Crowder's (1983) model is incorrect for a purely auditory memory list. To modify the PAS model, Crowder (1983) and also Morton, Marcus, and Othway (1983) assume that visual-speech (lipread) information is translated into the same type of representation as auditory speech at an earlier stage of analysis.
In breiter's revised model, both the visual and auditory consequences of speech provide footfall information at the level of the FAS model, but auditory and visual speech do not share auditory features in FAS. Supposedly, auditory features are selected and presented to the visual system in pure lipreading. The putative link between speech perception and speech production underlies a revised FAS model. This model might predict the effect of written information. Written information should not influence the selection of auditory features and, therefore, should not contribute to the auditory experience. Written information can still have an influence in identification, however, even though it doesn’t influence auditory experience. This effect would be processing and should differ qualitatively from the effect of lipreading. Post-perceptual refers to a response or decision bias in which the judgment might be influenced by the written information, but after auditory perception is complete. A post-perceptual model is developed allowing for a discussion of how writing might influence speech perception.

Given the impact of visible speech in the form of a speaker's articulations, it appeared possible that visible language in the form of writing might also influence how speech is heard. In this case, seeing a written segment, such as "BA" suggests the auditory perception of a spoken syllable towards "BA". To test this possibility, the present experiment directly compared the contribution of lipreading to written information given speech perception. Subjects were asked to watch a monitor and to listen to a spoken sound. They were told to repeat whether they heard the sound "BA" or "DA". The speech sound was created from nine synthetic speech sounds along a /BA/ to /DA/ continuum. Simultaneously with the speech sound, a visual event could also be presented. In the lipreading condition, the visual event was a TV monitor. In the auditory condition, the monitor was sometimes seen articulating the syllable "BA" or the syllable "DA". On some trials, no articulation was produced. In the reading condition, the two sounds on the monitor were sometimes changed to "BA" and "DA" during the audible presentation of the syllable. On other trials, no change in the sounds was made. In both conditions, subjects identified whether or not a word was heard.

Therefore, in addition to identifying the speech syllable "BA" or "DA", this dual task provided a check on whether the subject was actually looking at the visual event when it occurred.

There is historical precedence for this approach. In 1669, Baron Franciscus Mersurcas ab Helmont proposed that the letters of the Hebrew alphabet were not arbitrary but actually represented the tongue positions of the corresponding speech segments.

Figure 2

Figure 2 gives an example of Helmont's illustrations for "M", the 21st letter of the Hebrew alphabet. The letter is pronounced /M/ as indicated in Hebrew writing (right to left) in the bottom panel of the figure. The headband consists of other forms for the letter "M" as found on ancient coins, for example. No other extant ideas. Helmont's position was not unique; it would have been easy to find him justify the small appendages at the tip of the tongue. Actually, it would not be unreasonable to interpret this element as corresponding to the teeth and alveolar ridge. Helmont's study was followed by a series of studies culminating in Alexander Melville Bell's (1867) visible speech symbols. These symbols illustrated the visual nature of producing the sounds. It is interesting, however, that the symbols adopted and still used by the International Phonetic Association to represent all speech sounds have no direct production or articulation. This
It is possible that the written information will influence identification even though it cannot influence what the subject hears. The visual input might bias the subjects to report that events more often, even though the visual event did not influence what was heard. It is possible to observe an influence of the visual information in both the lip-reading and reading conditions, but for different reasons. For the latter, at the perceptual level, the integration should follow the description of the Lip-reading. A postperceptual bias should produce a different pattern of results. If the information after auditory perception, we might expect the probability of a total identification, $P(d/e)$, to be determined by:

$$P(d/e) = P(d/e|A, V) + P(d/e|A, V, p) + P(d/e|A, V, p)$$

Given a stimulus event with auditory level $A$ and visual level $V$, the probability of identifying the event as hearing $d$ is equal to hearing it as $d$ and responding on the basis of what was heard ($p_{d/e|A, V}$) and writing it as $d$ and responding on the basis of what was written ($p_{d/e|A, V, p}$). That is, the subject is assumed to respond on the basis of what was heard in proportion $p$ of the trials and on the basis of what was written in proportion $(1-p)$ of the trials. We might expect $p$ to be much larger than $(1-p)$, since subjects are instructed to respond on the basis of what they heard.

In contrast to the qualitative differences between the Lip-reading and reading conditions described by Coover's model, the Lip-reading predicts qualitative differences between the two conditions. The FLIP is aimed at describing the perceptual recognition of well formed patterns, regardless of the particular nature of the patterns involved. In order to test the hypothesis, the Lip-reading and reading conditions were compared.

Evoking the contribution of written information to the lip-reading perception also touches a test between the Lip-reading and Coover's revised FLIP model. The latter was chosen to predict both quantitatively and qualitatively different results. In the lip-reading and reading conditions, the study allows for a long-term fraction of the lip-reading information, since it has been observed in previous studies (e.g., figure 1). However, only the direct content of speech appears to influence the subject's identification responses. Identification is truly based on what the subject heard, and not on any written information. If written information is introduced, the subject should have more difficulty with the task.
Our experience of speech usually involves a joint occurrence of auditory and visual information. While it is clear that both types of information are important, they are not always equally salient. In fact, our perception of speech often relies on the interplay between auditory and visual information. For example, when watching TV or movies, we often rely on subtitles to fill in gaps in the auditory input.

Both the PAPA and PAS models provide a larger influence of lipreading than reading in the identification task. The critical difference between the predictions of the two models is not in terms of the magnitude of the visible effect, but in terms of the integration of auditory and visual information. This integration should be identical for lipread and written speech for the PAS model, and differ for Crowder's PAS model. Crowder's model captures the possibility of processing the integration of visual and auditory information in the same way as the PAPA (equation 1). The integration of visual and auditory information, however, would follow qualitatively different forms given by equation 2.

Method

Subjects

Seven adult subjects were recruited from the University of California community. These subjects were eliminated for failure to follow instructions and for reasons of error in recording the recall, giving a total of twelve subjects contributing to the study.

Stimuli

For the lipreading condition, the speech sounds were rendered on a videocassette. The talker was seated in front of a visual panel that contained only the upper body, eliminating potential distractions. For the auditory condition, the talker was recorded on a microphone and played back. The microphone was positioned in the center of the visual scene, allowing for movement in both the horizontal and vertical directions. On each trial the talker said either a word or nothing, as controlled by a computer-controlled switch.

Figure 3

The figure shows the recall of the talker's speech by the subjects in the lipreading condition. The recall is plotted as a function of the number of words spoken by the talker. The figure demonstrates the superiority of the lipreading condition over the auditory condition.
On each trial of the lipreading condition, one of the nine auditory stimuli on the continuum from neutral to loud was paired with one of the two possible visual articulations, 

On each trial, subjects were instructed to listen to the auditory stimulus and indicate whether they heard the sound "ba" or "da" on the second, whether or not there was a change in the visual domain. A visual change represented the speaker moving his lips in any way except in the lipreading condition and the movement of the latter string BA in DA in the reading condition. The tokens were arranged in a two-by-two configuration with the ba and da stimuli corresponding to the top and bottom rows, and the yes and no items corresponding to the left and right columns. For example, hitting the top right button indicated that the subject heard "ba" and that there was no visual change during the speech sound.

With an open-ended set of response alternatives in the book, subjects have reported a variety of perceptions: "neutral," "neither," and "yes" (Massaro and Orgler, 1983). We limited the choices to two options for practical reasons because subjects also had to report whether there was a change in the visual domain. What is important is that the two-alternative task provided an assessment of perception in the same manner as the open-ended alternative task. There is strong evidence that subjects base their confidence in their judgments on the degree of support for each alternative and choose an alternative from the available set of alternatives based on Luce's (1959) choice rules (Massaro, 1987). Given this evidence, two choices alternatives in the present task provide an appropriate measure of the influence of visual information on speech perception.

All subjects were tested in both lipreading and reading conditions in two consecutive sessions on a given day. The order of the two conditions was counterbalanced across subjects with six of the subjects receiving the lipreading condition first and six receiving the reading condition first. Each subject was tested for 592 experiments, trials, giving a total of up to 11 observations for each subject at each of the 84 experimental conditions.

Results
One important requirement in the present task is that the subjects looked at the visual event during the speech sound. To encourage the subjects to monitor the visual information and to evaluate whether they were looking at it, they were required to indicate whether or not a visual event occurred during the speech sound. Subjects were extremely accurate in this task, averaging 99% and 97% correct in the lipreading and reading conditions, respectively. In both conditions, subjects were about 2% or 3% more accurate in describing the presence, rather than the absence, of a change in the visual event.

Given that the subjects were looking at the visual event in both the lipreading and reading conditions, it is meaningful to analyze the identification results. The proportion of ideal identifications was computed for each subject at each of the 57 stimulus conditions for both the lipreading and reading conditions. A preliminary analysis revealed no effect on the order of presentation of the lipreading and reading conditions and this variable is ignored in the analysis presented here.
The left and right panels of figure 4 show the average proportion of the Upreading and Reading conditions, respectively. The proportion of the responses as a function of the auditory and visual conditions is shown in the upper left panel, and the auditory and visual conditions are shown in the lower right panel. The frequency of the responses as a function of the auditory and visual conditions is shown in the middle panels. The frequency of the responses as a function of the auditory and visual conditions is shown in the bottom panel. The frequency of the responses as a function of the auditory and visual conditions is shown in the right panel.
The RTM can be formalized to predict the results of the two visual conditions in the present study. A unique parameter is needed for each unique level of the speech error. Given that the visual and auditory information were used in both the lipreading and reading conditions, the visual parameters should differ for the lipreading and reading conditions, but the auditory parameters should not. Given this formalization, only 9 auditory and 8 visual parameters are necessary to predict the results of 2 x 8 x 54 = 1080 independent experimental conditions. The model was fit to the average proportion of trials recognized in each of the 12 subjects in the experiment. The average predictions shown in Figure 4 illustrate that the model gave a very good description of the results. The root mean squared deviation (RMSE) between predicted and observed values averaged 0.087 across the 12 subjects.

Another evaluation of the goodness of fit is to assess in what extent the fit can be improved by relaxing specific constraints. One constraint is that identical auditory parameters are assumed for the lipreading and reading conditions. Eliminating this constraint requires 9 additional parameters, for a total of 34. This new model was fit to the results, but improved the description of performance only slightly, giving an average RMSE of 0.080, and gave an equally good fit for the lipreading and reading conditions, respectively. To illustrate the large differences due to the visual information in the lipreading and reading condition, a third model was required identical visual parameters for these two conditions was used. This model assuming 9 auditory and 3 visual parameters gave an average RMSE of 0.147. A fourth model estimating 18 auditory parameters and 3 visual parameters gave an average RMSE of 0.060.

The parameter values of the FIM provide an index of the influence of lipreading and reading. The parameter value gives the degree to which ideaphon is replaced by the level of the independent variable. The parameter values for the three visual levels were .962, .962, and .962 for /ba/, /na/, and /da/ under the lipreading condition. The corresponding parameter values for /ba/, /na/, and /da/ under the reading condition. The magnitude of the visual effect is measured by the differences.

Figure 5

<table>
<thead>
<tr>
<th>Auditory</th>
<th>Reading</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.962</td>
<td>0.962</td>
</tr>
<tr>
<td>0.962</td>
<td>0.962</td>
</tr>
<tr>
<td>0.962</td>
<td>0.962</td>
</tr>
</tbody>
</table>
conditions. The magnitude of the effect of the visual variable depends on whether the respective probabilities of the parameters estimates are used. The difference between the lipreading and reading conditions appears to be much larger when the identification probabilities are compared relative to when the parameter estimates are compared. The differences are shown in magnitude of 8 in the identification judgments and less than a magnitude of 2 in parameter values.

The model for the experimental data was fit to the results. The model was then fit to the results of both the lipreading and reading conditions. The model required 8 parameters for \( P_{1} \), \( P_{2} \), \( P_{3} \), \( P_{4} \), \( P_{5} \), \( P_{6} \), \( P_{7} \), and \( P_{8} \) parameters for \( P_{1} \), \( P_{2} \), \( P_{3} \), \( P_{4} \), \( P_{5} \), \( P_{6} \), \( P_{7} \), and \( P_{8} \) for the lipreading and reading condition, respectively. One additional parameter was estimated for \( p \). As can be seen in Figure 5, the model gives a poor description of the results, with an R² value of 0.6.

To test the revised FAS model, the model for the experimental data was fit to the results. The model was then fit to the results of both the lipreading and reading conditions. The model required 8 parameters for \( P_{1} \), \( P_{2} \), \( P_{3} \), \( P_{4} \), \( P_{5} \), \( P_{6} \), \( P_{7} \), and \( P_{8} \) parameters for \( P_{1} \), \( P_{2} \), \( P_{3} \), \( P_{4} \), \( P_{5} \), \( P_{6} \), \( P_{7} \), and \( P_{8} \) for the lipreading and reading condition, respectively. One additional parameter was estimated for \( p \). As can be seen in Figure 5, the model gives a poor description of the results, with an R² value of 0.6.

**Discussion**

The results of the present study are difficult to explain. Significant changes in the magnitude of the finding for visual reading effect. Without doubt, lipreading data has a substantial influence on auditory speech recognition. On the other hand, data are inconsistent with the visual information. The results gave a good description of both the lipreading and reading conditions. However, the model gives a poor description of the reading conditions which were the main focus of the present visual information. These comments are contrary to what would be expected from the revised FAS model.

Future research should be aimed at inducing a larger effect of reading to allow a better test for the continuing...
Acknowledgement
The writing of this paper and the research reported in this paper were supported, in part, by NINCDS Grant 20314 from the Public Health Service and Grant 85-33-15292 from the National Science Foundation.

About the authors
Dominick W. Massaro has been professor of psychology at the University of California since 1993. He has received a research fellowship from both the National Institutes of Mental Health and the Organization Foundation. His research interests are human information processing, speech perception, and reading. Among his publications are Experimental Psychology and Information Processing (Cand McNally, 1978), Unearthing Language (Academic Press, 1973), and Later and Word Perception (North Holland, 1980), and Speech Perception by For and by a: A Paradigm for Psychological Inquiry (Tellemann, 1978).

Michael M. Cohen is a research associate at the University of California, Santa Cruz, since receiving his doctorate in 1993. His research interests include the development of synthetic auditory and visual speech and the testing of theoretical models.

Laura A. Thompson is a recent doctorate from the University of California, Santa Cruz. Her research interests include cognitive development and information processing. She is currently a research associate at the Max-Planck Institute for Human Development and Education in Berlin, West Germany.

about a doubt. Lipreading a has a substantial influence auditory speech perception. Reading print, in other hand, had a comparatively smaller effect. The FLAP gave a good description of both the lipreading and reading conditions.

models. Perhaps some other form of presentation could enhance the identification of a written input. For example, a word that is a meaningful syllable might influence a larger effect of reading. Based on previous findings, a printed word should influence auditory perception of the latter syllabic of the spoken form of the word. Mary 10 Wilson and Welsh (1978) and Easterbrooks et al. (1978) have found that a spoken message that contained mispronunciations of some of the word's final consonants might be pronounced as a syllable. All subjects were then instructed to repeat the last syllable of the words. The subjects were asked to notice the mispronunciations, they should not count them in their shadowing of the words. That is, they should select the mispronounced words as their correct form. In fact, subjects noticed many of the mispronunciations and were not able to repeat the mispronunciations in the third syllable but in the first syllable of the three-syllable words. A reasonable explanation is that recognition of the word occurred before the third syllable was heard, and this information influenced the later part of the word was heard.

A similar result might occur if a printed word is paired with a spoken word because the printed word might be recognized before hearing a three-syllable form of the spoken word. A similar result would still not necessarily mean that print influenced auditory speech perception directly. The effect could have been elicited by sound meaning. Printed and spoken nonwords could be used to assess whether word meaning is necessary to obtain the influence of print on auditory speech perception. Regardless of the meaning with respect to word meaning, this task would still test between the FLAP and the model. The FLAP predicts qualitatively similar results for lipreading, reading, and word meaning whereas the model predicts qualitatively different results for reading and meaning compared to lipreading.
References


