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

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 this special volume of *Visible Language* visible speech in the form of the lip movements of the speaker influences what we hear the speaker to be saying. Viewing the speaker can enhance understanding, especially when the auditory signal is degraded by masking noise. Three decades ago, Sumby and Pollack (1954) demonstrated that perceiving the face of a speaker was equivalent to increasing the signal-to-noise ratio of the auditory signal by 20 dB. The visual influence is not limited to situations with degraded auditory inputs. As reported by McGurk and MacDonald (1976) the visual input from the speaker can change the perceptual experience of an auditory speech event. Using videotape, these investigations dubbed a labial speech sound/ba-ba/onto the visual articulation of a velar stop consonant/ga-ga/. Subjects viewing and listening to the dubbed videotape often heard /da-da/.

Massaro and Cohen (1983) extended the McGurk and MacDonald (1976) demonstration by independently varying auditory and visual information in a factorial design. Subjects identified as /ba/ or /da/ speech events consisting of high-quality synthetic syllables ranging from /ba/ to /da/ combined with a videotaped /ba/ or /da/ or no articulation. Although subjects were instructed specifically to report what they heard, viewing the visual articulation made a large contribution to identification. The results in figure 1 show effects of both visual and auditory information and an interaction between these variables. The contribution of one source is larger to the extent the other source of information is ambiguous. For example, the magnitude of the visual effect is smaller at the unambiguous ends of the auditory speech continuum than in the middle ambiguous region of the continuum. The tests of quantitative models provided evidence for the integration of continuous and independent, as opposed to discrete and nonindependent, sources of information.

The results in figure 1 are adequately described by a fuzzy logical model of perception (FLMP). According to the FLMP, recognition is carried out in three stages. The first

Observed (points) and predicted (lines) proportion of /da/ identifications as a function of the auditory and visual levels of the speech event (from Massaro and Cohen, 1983). The predictions are given by a fuzzy logical model of perceptual recognition.

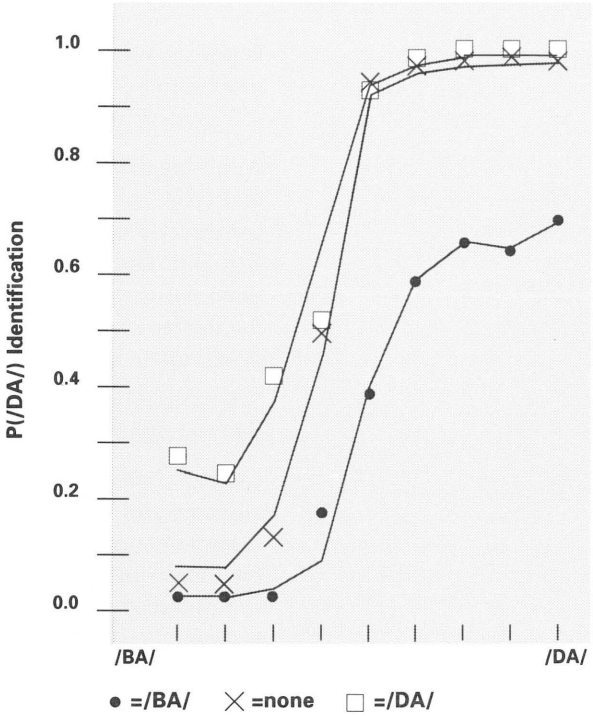


Figure 1 Auditory

stage is feature evaluation, during which the stimulus input is transduced by the sensory systems and various perceptual features are derived. The features are assumed to be continuous rather than discrete. The outcome of featural 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 featural information is compared with prototype definitions to determine to what degree each prototype is realized in the speech event. A prototype defines a segment of speech in terms of the conjunction of features that make it up. The third stage of recognition processing is pattern classification. During this stage the merit of each potential prototype is evaluated relative to the summed merits of all potential prototypes. The relative goodness of a prototype gives the proportion of times it would be selected as a response. An important property of the model is that one

feature has its greatest effect when the second is at its most ambiguous level. The most informative feature has the greatest impact on the judgments.

Applying the model to the present task using auditory and visual speech, both features are assumed to provide continuous and independent evidence for the alternatives /ba/ and /da/. Defining the onsets of the second (F2) and third (F3) formants as the important auditory feature and the degree of initial opening of the lips as the important visual feature, the prototypes are:

/da/ : Slightly falling F2–F3 & Open lips
/ba/ : Rising F2–F3 & Closed lips

Given a prototype's independent specifications for the auditory and visual features, the value of one feature cannot change the value of the other feature at the prototype matching stage. In addition, the negation of a feature is defined as the additive complement. That is, we can represent Rising F2–F3 as (1-Slightly falling F2–F3) and Closed Lips as (1-Open lips),

/da/ : Slightly falling F2–F3 & Open lips
/ba/ : (1–Slightly falling F2–F3) & (1–Open lips).

The integration of the features defining each prototype is evaluated according to the product of the truth values representing each feature. If a_i represents the degree to which the auditory stimulus A_i has Slightly falling F2–F3 and v_j represents the degree to which the visual stimulus V_j has Open lips, the outcome of the prototype matching would be:

/da/ : $a_i v_j$
/ba/ : $(1-a_i)(1-v_j)$.

If these two prototypes are the only valid response alternatives, the pattern classification operation would determine their relative merit leading to the prediction that

Equation 1

$$P(/da/ \mid A_i V_j) = \frac{a_i v_j}{a_i v_j + (1-a_i)(1-v_j)}$$

The predictions of the model require one parameter for each unique level of the auditory and visual features. Massaro and Cohen (1983) combined nine levels of the auditory stimulus with three levels of the visual giving a total of 27 experimental conditions (see figure 1). Given

nine levels of A_i and three levels of V_j, the predicitons of the model require 12 parameters (nine a_i values and three v_j values). The quantitative predictions of the FLMP were computed for the observed proportion of a /da/ response for each subject using the parameter estimation program STEPIT (Chandler, 1969). A model is represented to the analysis program STEPIT as a set of prediction equations and a set of unknown parameters. The goal of STEPIT is to find a set of parameter values that optimize the predictions of the observed data. Initially, all parameters are set to .5. By iteratively adjusting the parameters of the model, STEPIT minimizes the squared deviations between the 27 observed and 27 predicted points. As can be seen in the figure 1, 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 for a model assuming discrete rather than continuous features or a model with nonindependent features.

Auditory 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 featural information at the level of PAS. . .Supposedly, auditory feature selection can occur even in the absence of sound, as in pure lipreading.

An alternative account of bimodal speech perception is proposed by Crowder (1983) who modified his 1978 model to account for the contribution of visual information to speech perception. Auditory information is assumed to be transduced and the output of auditory feature detectors are stored in a preperceptual acoustic storage (PAS). The primary evidence for PAS has been a suffix effect, which occurs when an auditory speech stimulus follows an auditory memory list and interferes with recall of the last item on the list. A pure tone suffix or a nonauditory but meaningful suffix does not produce similar interference. These results seem to provide evidence for an auditory representation that has specific sensory channel characteristics. Since publication of Crowder's (1978) model, however Spoehr and Corin (1978), Campbell and Dodd (1980), and Greene and Crowder (1984) have shown that watching someone else articulate the suffix or mouthing the suffix silently yourself also produces a suffix effect. This result appeared to Crowder (1983) to be troublesome for a purely auditory entry into PAS. To modify the PAS model, Crowder (1983) and also Morton, Marcus, and Ottley (1981) assume that visual-speech (lipread) information is translated into the same type of representation as the auditory speech at an early stage of analysis.

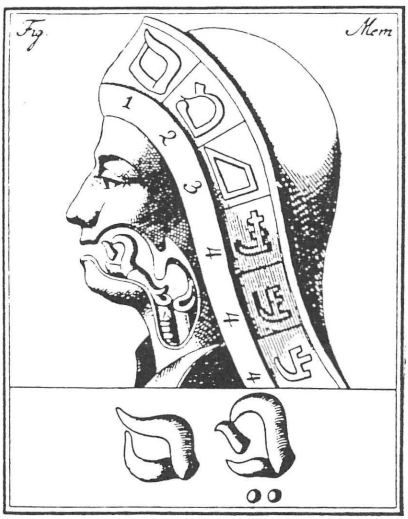
In Crowder's revised model, both the visual and auditory consequences of speech provide featural information at the level of PAS; that is, both auditory and visual speech can place auditory features in PAS. Supposedly, auditory feature selection can occur even in the absence of sound, as in pure lipreading. The putative link between speech perception and speech production rationalizes the revised PAS model. This model might predict no 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 could still have an influence in identification, however, even though it doesn't influence auditory experience. This effect would be post-perceptual 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 following a brief 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, would bias the auditory perception of a spoken syllable towards /ba/. To test for this possibility, the present experiment directly compared the contribution of lipread to written information in speech perception. Subjects were asked to watch a monitor and to listen to a speech sound. They were told to report whether they heard the sound /ba/ or /da/. The speech sound was chosen from nine synthetic speech sounds along a /ba/ to /da/ continuum. Simultaneous with the speech sound, a visual event could also be presented. In the lipreading condition, the person on the TV 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 asterisks on the monitor were sometimes changed to the letters BA or DA during the audible presentation of the syllable. On other trials, no change in the asterisks was made. In both conditions, subjects identified whether or not a visual event

occurred, in addition to identifying the speech syllable that was heard. This dual task provided a check on whether the subject was actually looking at the visual event when it occurred.

There is historical precedence that is of interest. In 1667, Baron Franciscus Mercurius ab Helmont proposed that the letter symbols of the Hebrew alphabet were not arbitrary but actually represented the tongue positions of the corresponding speech segments.

Figure 2



Hebrew letter M as a tongue position according to Helmont. The lower panel gives the Hebrew (to be read from right to left) pronunciation of the letter /Mm/.

Figure 2 gives one of Helmont's illustrations for M, the 13th letter of the Hebrew alphabet. The letter is pronounced /mɛm/ 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. Not unlike some extant ideas, Helmont's position was not airtight; it would have been enjoyable to watch him justify the small appendage 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 vocal action in 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 speech-production connotations. This

. . . The symbols adopted and still used by the International Phonetic Association to represent speech sounds have no speech-production connotations. . . a unique speech gesture is not necessary to produce a given sound catagory. . . The evidence encourages asking whether an orthographic stimulus could influence speech perception in the same manner as a visible spoken articulation. Evaluating the contribution of written information to speech perception also invites a test between the FLMP and Crowder's revised PAS model. . . If identification is truely based on what the subject heard, then the written information should have no effect.

might be due, in part, to the fact that a unique speech gesture is not necessary to produce a given sound category.

There is some basis for expecting that printed language might influence the perception of spoken language. Ehri (1984) makes a strong case for the influence of orthography on a child's spoken language processing. As an example, a prereader has difficulty recognizing spoken function words (such as might, could, or from) as single words. A novice reader, on the other hand, performs the same task quite easily. Learning to read also enables children to segment spoken words into their constituent phonemes more easily. Written language also influences the processing of spoken language for literate adults. In one task, subjects are asked to indicate as quickly as possible whether or not two spoken words rhyme (Seidenberg & Tanenhaus, 1979). They are faster in detecting that two orthographically-similar words rhyme compared to two dissimilar words. As as example, subjects respond yes more quickly to the spoken words name-blame than to the words name-claim. Other encouraging evidence comes from Campbell who used written pseudohomophones (wunn, tooe, threa) in the suffix memory task. Recency effects were observed and this advantage for the last few items in the list was eliminated by an auditory suffix (the spoken word go). Ehri (1984) reviews other positive evidence for the influence of spelling on the perceptual processing of spoken language. For our purposes, the evidence encourages asking whether an orthographic stimulus could influence speech perception in the same manner as a visible spoken articulation.

Evaluating the contribution of written information to speech perception also invites a test between the FLMP and Crowder's revised PAS model. The latter would seem to predict both quantitatively and qualitatively different results for the lipreading and reading conditions. The model allows for a large contribution of the lipread information, as has been observed in previous studies (e.g., figure 1). However, only the direct correlates of speech should influence the /ba/ or /da/ identification reponses. If identification is truly based on what the subject heard, then the written information should have no effect.

It is possible that the written information will influence identification even though it does not influence what the subject hears. The visual event might bias the subjects to report that event 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 lipreading and reading conditions, but for different reasons. For effects at the perceptual level, the integration should follow that described by the FLMP. A post-perceptual bias should produce a different pattern of results. If the bias occurs after auditory perception, we might expect the probability of a /da/ identification, P(/da/), to be described by

Equation 2 $P(/da/ | A_i V_j) = p[P_h(/da/ | A_i V_j)] + (1\text{-}p)[P_s(/da/ | A_i V_j)]$

Given a stimulus event with auditory level i and visual level j, the probability of identifying that event as /da/ is equal to hearing it as /da/and responding on the basis of what was heard $(p[P_h(/da/ | A_i V_j])$ and seeing it as /da/ and responding on the basis of what was seen $((1\text{-}p)[P_s(/da/ | A_i V_j)])$. That is, the subject is assumed to respond on the basis of what was heard on proportion p of the trials and on the basis of what was seen on proportion $(1\text{-}p)$ of the trials. We might expect p to be much larger than $(1\text{-}p)$ since subjects are instructed to respond on the basis of what they heard.

In contrast to the qualitative differences between the lipreading and reading conditions predicted by Crowder's model, the FLMP predicts no qualitative differences between the two conditions. The FLMP is aimed at describing the perceptual recognition of well-learned patterns, regardless of the particular nature of the patterns involved. In addition to speech perception, the FLMP has been successfully applied to letter and word recognition (Massaro, 1979; Oden, 1977), object recognition (Oden, 1981), and sentence interpretation (Oden, 1977). It should be noted that the FLMP only predicts the nature of the processes involved in perceptual recognition it does not predict the feature values of various aspects of the stimulus environment. Although the FLMP makes no formal prediction about the relative magnitude of lipread and written influences in speech perception, a larger effect of the lipread information seems most likely.

Our experience of speech events usually consists of the joint occurrence of lipread and sound information whereas it seldom consists of the pairing of written and sound information (except in reading to our children, but we tend not to listen anyhow). In fact, our experience also consists of situations in which the written and spoken messages are completely uncorrelated as, for example, in reading subtitles while watching and listening to a foreign film.

Both the FLMP and the PAS model can predict 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 audible and visual information. This integration should be identical for lipread and written information for the FLMP, but should differ for Crowder's PAS model. Crowder's model might be granted the possibility of predicting the integration of lipread information and auditory information in the same manner as the FLMP (equation 1). The integration of written information and auditory information, however, should follow the quantitatively different form given by equation 2.

Both the FLMP and PAS model can predict a larger influence of lipreading 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 audible and visual information.

Method

Subjects

Seventeen adult subjects were recruited from the University Community. Three subjects were eliminated for failing to follow instructions and two because of an error in recording the results, giving a total of twelve subjects contributing to the results.

Stimuli

For the lipreading condition, the speech events were recorded on a videotape. The author was seated in front of a wood panel background, illuminated with ordinary fluorescent fixtures in the ceiling. The speaker's head was centered in the video field and filled about 2/3 of the frame in both the horizontal and vertical directions. On each trial the speaker said either /ba/ or /da/ or nothing as cued by a video terminal under computer control.

The original audio track was replaced with synthetic speech. The speaker's /ba/'s or /da/'s were analyzed using linear prediction to derive a set of parameters for driving a software formant serial resonator speech synthesizer (Klatt, 1980). By altering the parametric information regarding the first 80 msec of the CV, a set of nine 400 msec CVs covering the range from /ba/ to /da/ was created. Figure 3 gives sound spectrograms for 5 of the synthetic syllables along the continuum. During the first 80 msec F1 went from 300 Hz to 700 Hz following a negatively accelerated path. The F2 followed a negatively accelerated path to 1199 Hz from one of nine values equally spaced between 1000 and 2000 Hz from most /ba/-like to most/da/-like, respectively. The F3 followed a linear transition to 2729 Hz from one of nine values equally spaced between 2200 and 3200 Hz. All other characteristics of synthetic CVs were identical for the nine test stimuli. Additional details of the video recording and the speech synthesis are given in Massaro and Cohen (1983).

An experimental tape was made by copying the original tape and replacing the original sound track with the synthetic speech. The presentation of the synthetic speech was synchronized with the original audio track on the videotape and gave the strong illusion that the synthetic speech was coming from the mouth of the speaker. To accomplish this synchronization, the audio signal was monitored by a schmidt trigger circuit. When the original audio channel on the videotape exceeded a preset threshold, one of the 400 msec CV syllables was played.

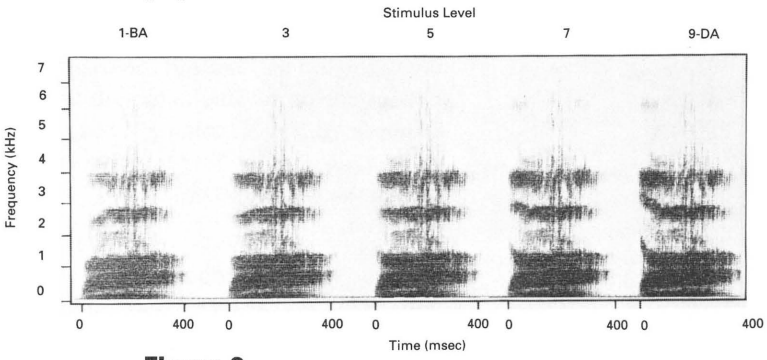


Figure 3

Spectrograms for five of the syllables along the /ba/ to /da/

On each trial of the lipreading condition, one of the nine auditory stimuli on the continuum from /ba/ to /da/ was paired with one of the two possible visual articulations, /ba/ or /da/, or with no articulation. The stimuli were presented in 11 blocks of the 27 possible combinations, sampled randomly without replacement. A practice block of 10 trials preceded the 297 experimental trials. The subjects had about three seconds to make their response before the next trial.

The reading condition was designed to duplicate the lipreading condition except for the nature of the visual information. Subjects viewed a TV monitor and fixated on a row of two asterisks centered on the monitor. On two-thirds of the trials, the asterisks could be replaced by the letters strings BA or DA during the 400 msec presentation of the speech sound. On the other one-third of the trials, the asterisks remained in view during the presentation of the speech sound. The sequence, number, and timing of speech and visual events were identical to those in the lipreading condition. In both the lipreading and reading conditions, subjects listened to the speech stimuli over headphones (Koss Pro 4AA) at a comfortable listening intensity (71 dB-A).

On each trial, subjects were instructed to hit one of four buttons, indicating the outcome of two events: first, whether they heard the sound /ba/ or /da/ and second, whether or not there was a change in the visual domain. A visual change represented the speaker moving his lips to say /ba/ or /da/ in the lipreading conditions and the occurrence of the letter strings BA or DA in the reading condition. The buttons were arranged in a two-by-two configuration with the ba and da alternatives corresponding to the top and bottom rows, and the yes and no alternatives corresponding to the left and right columns. For example, hitting the top right button indicated that the subject heard a /ba/ and that there was no visual change during the speech sound.

With an open-ended set of response alternatives in the task, subjects have reported a variety of percepts: /tha/, /va/, /bda/, and /ga/ (Massaro and Cohen, 1983). We limited the choices to two alternatives 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 provides an assessment of perception in the same manner as the open-ended-alternative task. There is strong evidence that subjects have continuous information indicating the degree of support for each alternative and choose an alternative from the permissible set of alternatives based on Luce's (1959) choice rule (Massaro, 1987). Given this evidence, two choice alternatives in the present task provide an appropriate measure of the influence of visual information on speech perception.

All subjects were tested in both the 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 594 experimental trials, giving a total of up to 11 observations for each subject at each of the 54 experimental conditions.

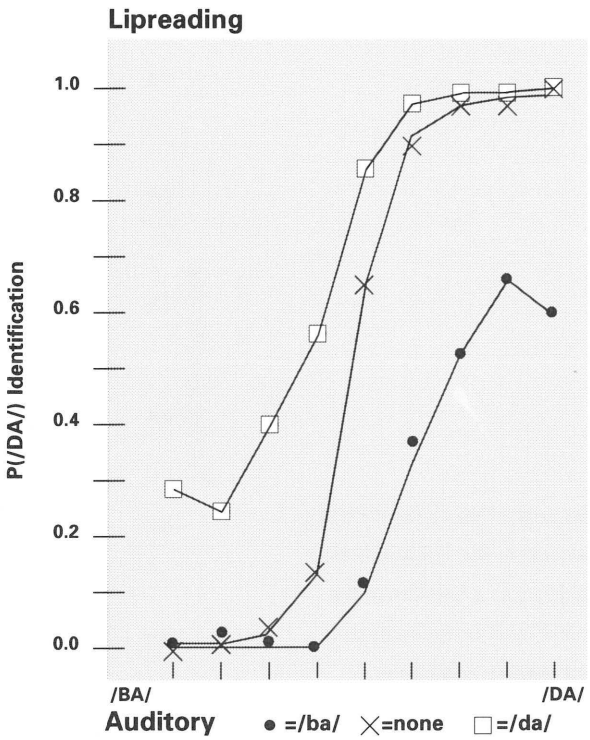
Results

One important requirement in the present test 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 96% and 97% correct in the lipreading and reading conditions, respectively. In both conditions, subjects were about 2% or 3% more accurate in determining 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 the reading conditions, it is meaningful to analyze the identification results. The proportion of /da/ identifications was computed for each subject at each of the 27 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 give the average results for the lipreading and reading conditions, respectively. The proportion of /da/ responses as a function of the nine levels along the auditory speech continuum is shown with the visual "ba", "da", or "none" as the curve parameter. The average proportion of/da/ responses increased significantly as the level of the auditory syllable went from the most /ba/-like to the most /da/-like level, $F(8,80) = 311$, $p<.001$. There was also a large effect on the proportion of /da/ responses as a function of the visual stimulus, with fewer /da/ responses for visual "ba" than for a visual "da", $F(2, 20) = 26$, $p< 0.001$. The interaction of these two variables was also significant, $F(16,160) = 11.2$, $p<.001$, since the effect of the visual variable is smaller at the less ambiguous regions of the auditory dimension.

The result of central interest is the difference between the lipreading and reading conditions given in the two panels in figure 4. What is most apparent is the much larger effect of the visual information in the lipreading relative to the reading condition. The visual variable was about 9 times more effective in the lipreading than in the reading condition. Figure 5 gives a graphical reprsentation of the visual effect for each subject in the lipreading and reading conditions. Every subject showed a larger effect of lipreading relative to reading. Only two subjects showed lipreading effects of about the same size as the reading effect. The lipreading/reading comparison inter-acted with the auditory variable, $F(8,80) = 2.68$, $p<.025$, the visual variable, $F(2,20) = 4.11$, $p<.05$ and the auditory/visual interaction, $F(16,160) = 5.5$, $p<.001$. Although the magnitude of the visual variable was much less in the reading condition, it was still statistically significant, $F(2,22) = 14.3$, $p<.001$, as was the interaction between the auditory and visual variables, $F(16, 176) = 2.73$, $p<.005$. Thus, although the magnitude of the visual variable differed greatly between the lipreading and reading conditions, the form of its interaction with the auditory variable was very similar in the two conditions. The visual influence was always largest at the most ambiguous levels of the auditory variable.



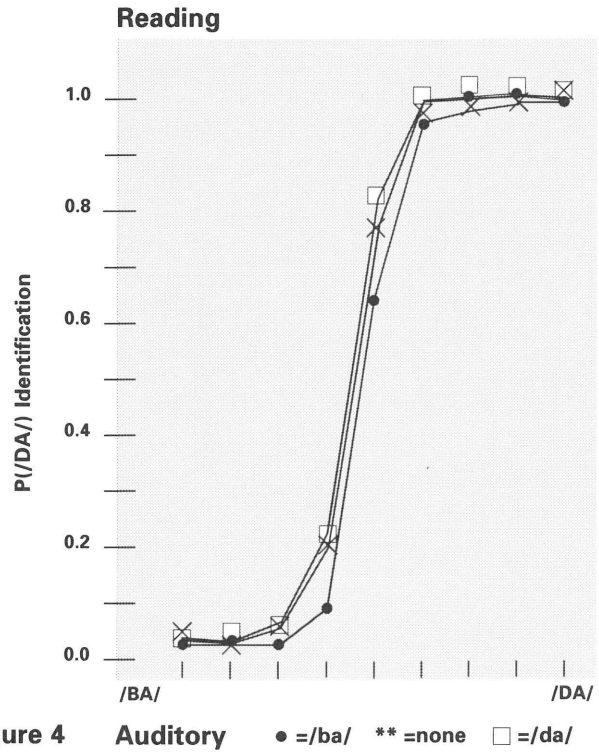


Figure 4 Observed (points) and predicted lines proportion of /da/ identifications as a function of the auditory and visual levels of the speech event. The top panel gives the results for the lipreading condition and the bottom panel for the reading condition. The predictions are for the fuzzy logical model of preception.

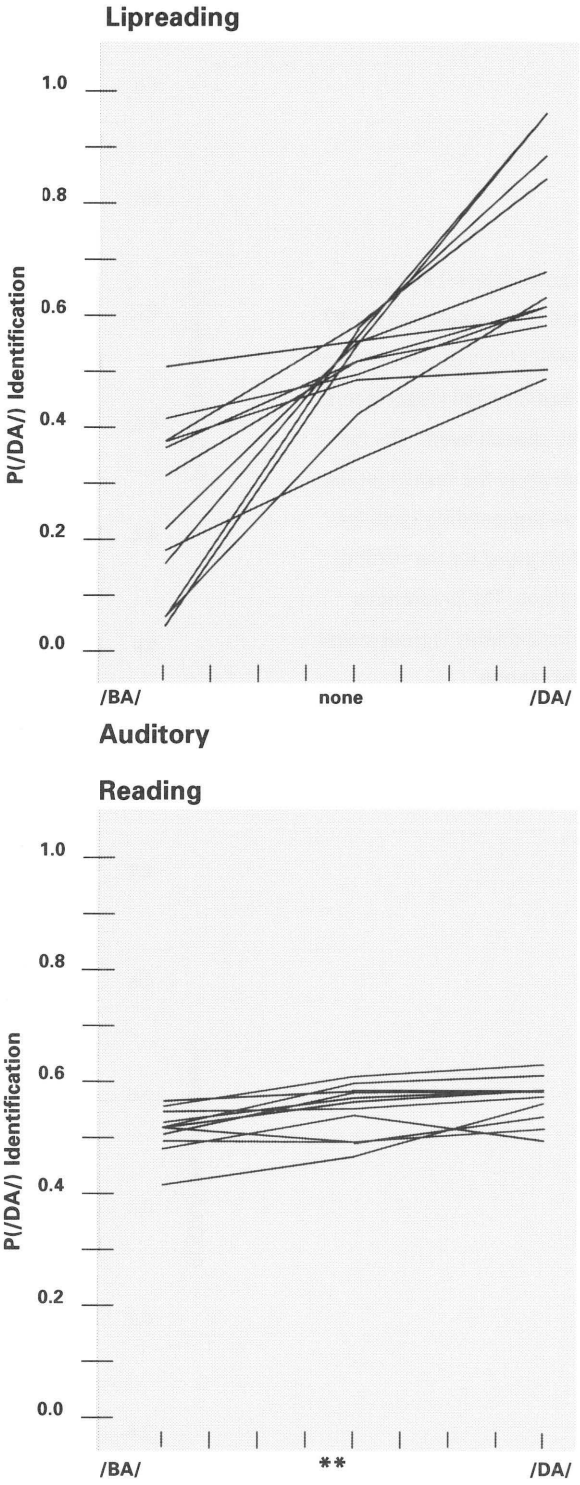


Figure 5 The proportion of /da/ identifications for the 12 individual subjects as a function of the visual level in the lipreading and reading conditions.

The FLMP 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 event. Given that the same auditory information was used in both the lipreading and reading conditions, the visual V_j parameters should differ for the lipreading and reading conditions, but the auditory parameters should not. Given this formalization, only 9 auditory and 6 visual parameters are necessary to predict the results of 2x3x9=54 independent experimental conditions. The model was fit to the average proportion of /da/ responses for 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 (RMSD) between predicted and observed values averaged 0.037 across the fits of the 12 subjects.

. . .Although the magnitude of the visual variable differed greatly between the lipreading and reading conditions, the form of its interaction with the auditory variable was very similar in the two conditions. The visual influence was always largest at the most ambiguous levels of the auditory variable.

Another evaluation of the goodness of fit is to assess to what extent the fit can be improved by removing certain 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 24. This new model was fit to the results, but improved the description of performance only slightly, giving an average RMSD of 0.030, 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 that required identical visual parameters for these two conditions was tested. The model assuming 9 auditory and 3 visual parameters gave an average RMSD of 0.147. A fourth model estimating 18 auditory parameters and 3 visual parameters gave an average RMSD of 0.060.

The parameter values of the FLMP also provide an index of the influence of lipreading and reading. The parameter value gives the degree to which /da/ is supported by the level of the independent variable. The parameter values for the three visual levels were .030, .600, and .940 for /ba/, none, and /da/ under the lipreading condition. The corresponding parameter values for BA, **, and DA were .511, .705, and .765 under the reading condition. The magnitude of the visual effect is measured by the differences in the parameter values across the three visual

conditions. The magnitude of the effect of the visual variable depends on whether the response probabilities or the parameter 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 over a magnitude of 9 in the identification judgments and less than a magnitude of 4 in parameter values.

The post-perceptual guessing model was also fit to the results. This model was first fit to the results of both the lipreading and reading conditions. The model required 9 parameters for P_h(/da/), and 3 parameters for P_s(/da/) for the lipreading and 3 for the reading conditions. One additional parameter was estimated for p. As can be seen in figure 6, this model gave a poor description of the results with an RMSD of .162.

To test the revised PAS model, the post-perceptual guessing model given by equation 2 with 13 parameters was fit to just the reading condition. This model gave a poor description with an RMSD of .054, compared to the RMSD of .029 for the FLMP with 12 parameters fit to the same results.

Discussion

The results of the present study are difficult to evaluate primarily because of the finding of a small reading effect. Without a doubt, lipreading a face has a substantial influence on auditory speech recognition. Reading print, on the other hand, had a comparitively smaller effect. The size of the effect of reading compared to lipreading was not as critical in distinguishing among the different theories as was how the visual information interacted with the auditory information. The FLMP gave a good description of both the lipreading and reading conditions. Furthermore, the guessing model gave a poor description of the reading conditions which weakens the argument of a post-perceptual influence of reading. These outcomes are contrary to what would be expected from the revised PAS model.

Future research should be aimed at inducing a larger effect of reading to allow a better test for the contrasting

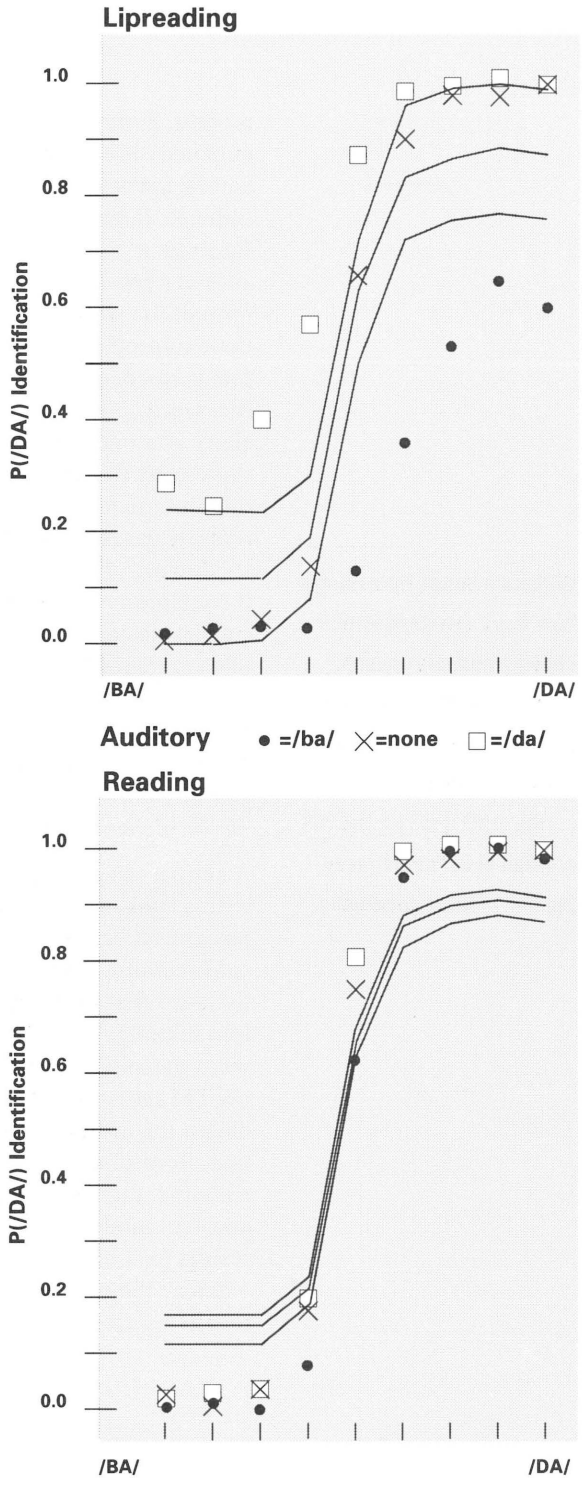

Lipreading
1.0
0.8
0.6
0.4
0.2
0.0
P(/DA/) Identification
/BA/
/DA/
Auditory
● =/ba/
×=none
□=/da/
Reading
1.0
0.8
0.6
0.4
0.2
0.0
P(/DA/) Identification
/BA/
/DA/
Auditory
● =/ba/
** =none
□=/da/

Figure 6

models. Perhaps some other form of presentation would enhance the contribution of a written input. For example, a word rather than a meaningless syllable might induce a larger effect of reading. Based on previous findings, a printed multisyllabic word should influence auditory perception of the latter syllables of the spoken form of the word. Marslen-Wilson and Welsh (1978) had their subjects shadow (repeat back) a spoken message that contained mispronunciations of some of the words (the word confusion might be pronounced as gunfusion). Of interest was the extent that subjects would be swayed by the linguistic context to miss these errors in the pronunciation. If subjects fail to notice the mispronunciations, they should not include them in their shadowing of the message; that is, they should restore the mispronounced words to their correct form. In fact, subjects restored many of the mispronunciations and were more likely to restore mispronunciations in the third syllable than in the first syllable of a three-syllable word. A reasonable explanation is that recognition of the word occurred before the third syllable was heard and this information influenced how the latter part of the word was heard.

Without a doubt, lipreading a face has a substantial influence on auditory speech recognition. Reading print, on the other hand, had a comparitively smaller effect . . . the FLMP gave a good description of both the lipreading and reading conditions.

A similar result might occur if a printed word is paired with a spoken word. Because the printed word might be recognized before hearing the third syllable of the spoken word, a positive result would still not necessarily mean that print influenced auditory speech perception directly. The effect could have been mediated by word 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 outcome with respect to word meaning, this task would still test between the FLMP and PAS model. The FLMP predicts qualitatively similar results for lipreading, reading, and word meaning whereas the PAS model predicts qualitatively different results for reading and meaning compared to lipreading.

Acknowledgement

The writing of this paper and the research reported in the paper were supported, in part, by NINCDS Grant 20314 from the Public Health Service and Grant BNS-83-15192 from the National Science Foundation.

About the authors

Dominic W. Massaro has been professor of psychology at the University of California since 1980. He has received research fellowship awards from both the National Institutes of Mental Health and the Guggenheim Foundation. His research interests are human information processing, speech perception, and reading. Among his publications are Experimental Psychology and Information Processing (Rand McNally, 1975), Understanding Language (Academic Press, 1975), and Letter and Word Perception (North Holland, 1980), and Speech Perception by Ear and Eye: A Paradigm for Psychological Inquiry (Erlbaum, 1987).

Michael M. Cohen is a research associate at the University of California, Santa Cruz, since receiving his doctorate in 1984. His research interests include the development of synthetic auditoryu and visible speech and the testing of mathematical 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.

References

Bell, A. M. 1867. *Visible speech: The science of universal alphabetics.* London: Simpkin, Marshall & Co.

Campbell, R. 1987. Remembering with impurity when pre-categorical acoustic storage is not acoustic, what is it? In D. A. Allport, D. MacKay, W. Prinz & E. Scheerer (Eds.) *Language perception and production: Common mechanisms in listening, speaking, reading and writing.* Academic Press, N.Y.: 132–150.

Campbell, R., & Dodd, B. 1980. Hearing by eye. *Quarterly Journal of Experimental Psychology,* 32, 85–99.

Chandler, J. P. 1969. Subroutine STEPIT – Finds local minima of a smooth function of several parameters. *Behavioral Science,* 14, 81–82.

Crowder, R. G. 1978. Mechanisms of backward masking in the stimulus suffix effect. *Psychological Review,* 85, 502–524.

Crowder, R. G. 1983. The purity of auditory memory. *Philosophical Transactions of the Royal Society, Section B.* 302, 251–265.

Ehri, L. C. 1984. How orthography alters spoken language competencies in children learning to read and spell. In J. Downing & R. Valtin (Eds.), *Language awareness and learning to read,* (pp. 119–147). N.Y.: Springer Verlag.

Greene, R. L. & Crowder R. G. 1984. Modality and suffix effects in the absence of auditory stimulation. *Journal of Verbal Learning and Verbal Behavior,* 23, 371–382.

Helmont, B. F. M. ab. 1667. *Alphabeti vere naturalis Hebraici Brevissima Delineatio.*

Klatt, D. H. 1980. Software for a cascade/parallel formant synthesizer. *Journal of the Acoustical Society of America,* 67, 971–995.

Luce, R. D. 1959. *Individual choice behavior.* N.Y.: Wiley.

Marslen-Wilson, W., & Welsh, A. 1978. Processing interactions and lexical access during word recognition in continuous speech. *Cognitive Psychology,* 10, 29–63.

Massaro, D. W. 1979. Reading and listening (Tutorial paper). In P. A. Kolers, M. Wrolstad, & H. Bouma (Eds.) *Processing of Visible Language: Vol. 1,* (pp. 331–354). N.Y.: Plenum.

Massaro, D. W. 1987. *Speech Perception by Ear and Eye: A Paradigm for Psychological Inquiry.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Massaro, D. W., & Cohen, M. M. 1983. Evaluation and integration of visual and auditory information in speech perception. *Journal of Experimental Psychology: Human Perception and Performance,* 9, 753–771.

McGurk, H. 1981. Listening with eye and ear (paper discussion). In T. Myers, J. Laver, & J. Anderson (Eds.), *The cognitive representation of speech.* Amsterdam: North-Holland.

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

Morton, J., Marcus, S. M., & Ottley, P. 1981. The acoustic correlates of "speechlike": A use of the suffix effect. *Journal of Experimental Psychology: General,* 110, 568–593.

Oden, G. C. 1977. Integration of fuzzy logical information. *Journal of Experimental Psychology: Human Perception and Performance,* 3, 565–575.

Oden, G. C. 1979. A fuzzy logical model of letter identification. *Journal of Experimental Psychology: Human Perception and Performance,* 5, 336–352.

Oden, G. C. 1981. A fuzzy propositional model of concept structure and use: A case study in object identification. In G. W. Lasker (Ed.), *Applied Systems Research and Cybernetics.* Elmsford, NY: Pergamon Press.

Seidenberg, M. S., & Tanenhaus, M. K. 1979. Orthographic effects on rhyme monitoring. *Journal of Experimental Psychology: Human Learning and Memory ,* 5, 546–554.

Spoehr, K.T., & Corin, W. J. 1978. The stimulus suffix effect as a memory coding phenomenon. *Memory & Cognition,* 6, 583–589.

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