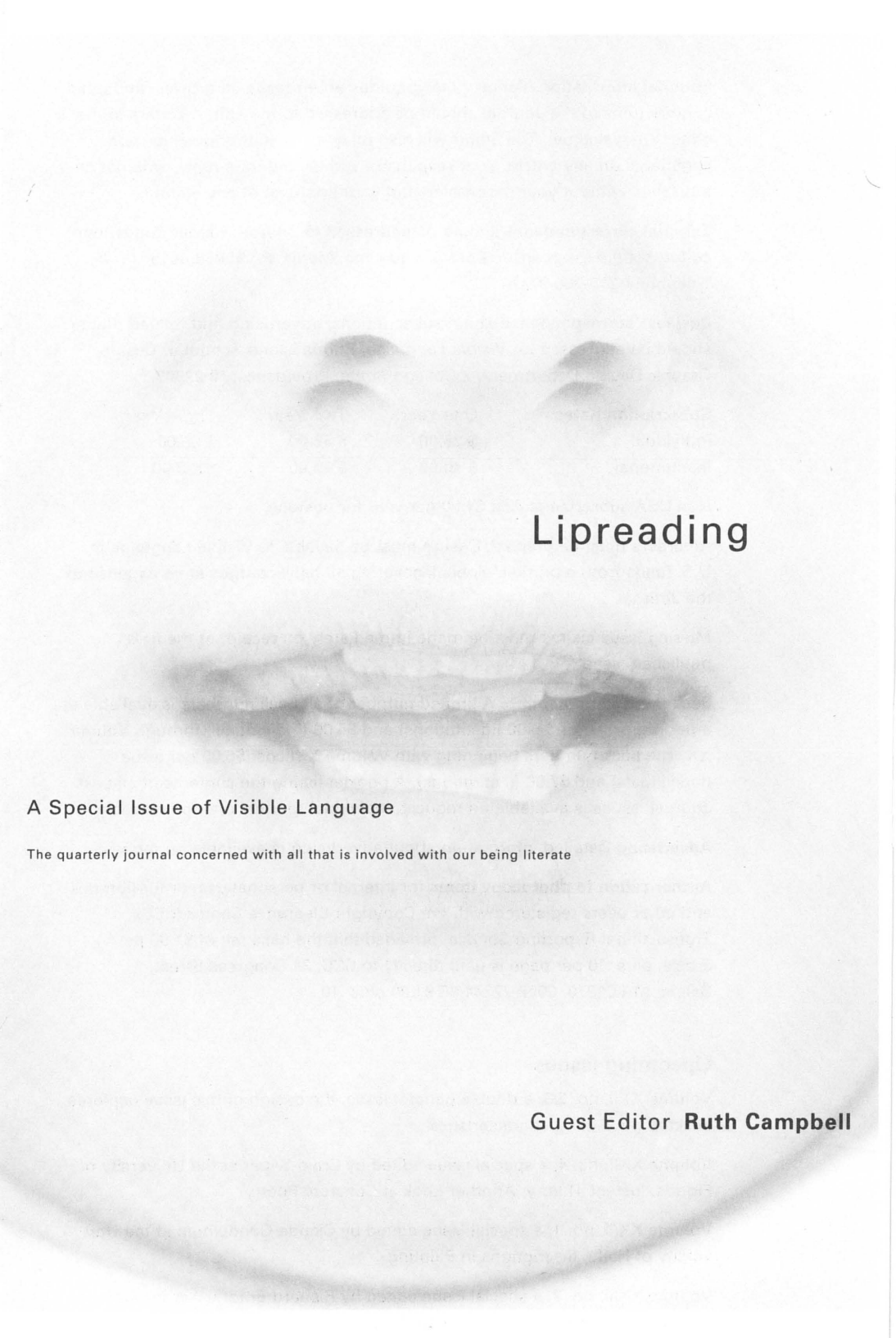


Lipreading

A Special Issue of Visible Language

The quarterly journal concerned with all that is involved with our being literate

Guest Editor **Ruth Campbell**

Editorial Information Manuscripts, inquiries about research articles, and other contributions to the Journal should be addressed to the editor. Letters to the editor are welcome. The Editor will also relay to the author questions or comments on any article. Your response - and the author's reply - will not be published without your permission and your approval of any editing.

Editorial correspondence should be addressed to Sharon Helmer Poggenpohl, editor, *Visible Language*, 6 Cold Spring Lane, Media, PA 19063-4510 U S A. Telephone 215-565-9747.

Business correspondence about subscriptions, advertising and related matter should be addressed to: *Visible Language*, Rhode Island School of Design, Graphic Design Department, 2 College Street, Providence, RI 02903

Subscription Rates	One Year	Two Year	Three Year
Individual	$ 25.00	$ 45.00	$ 65.00
Institutional	$ 40.00	$ 75.00	$110.00

Non USA subscribers: Add $4.00 per year for postage.

All orders must be prepaid. Checks must be payable to Visible Language in U.S. funds from a prime U.S. bank covering all bank charges at no expense to the Journal.

Missing issue claims must be made immediately on receipt of the next published issue.

Back Copies and Reprints A limited number of all back numbers is available at a per issue cost of $ 6.00 (institutions) and $5.00 (individuals) through Volume XX. The back numbers beginning with Volume XXI cost $6.00 per issue (individuals) and $7.00 (institutions). A booklet listing the contents of all past Journal issues is available on request. Individual reprints are not available.

Advertising Detailed information about advertising is available on request.

Authorization to photocopy items for internal or personal use, or for libraries and other users registered with the Copyright Clearance Center (CCC) Transactional Reporting Service, provided that the base fee of $1.00 per article, plus .10 per page is paid directly to CCC, 21 Congress Street, Salem, MA 01970. 0022-22244/86 $1.00 plus .10.

Upcoming issues

Volume XXII, no. 2/3, a double general issue, the design of this issue explores desktop publishing characteristics.

Volume XXII, no. 4, a special issue edited by Craig Saper at the University of Florida, Instant Theory, Another Look at Concrete Poetry.

Volume XXIII, no. 1, a special issue edited by Claude Gandelman at the University of Haifa, Inscriptions in Painting.

Volume XXIII, no. 2, a special issue edited by Richard Bradford at the University of Ulster, Printed Poetry and Its Criticism.

Table of Contents

Volume XXII Number 1 Winter 1988

Advisory Board

Introduction

Ruth Campbell

Department of Experimental Psychology, University of Oxford, South Parks Road, Oxford, OXI 3 UD, U.K.

Visible Language XXII, 1
Ruth Campbell, pp. 4-7

The British scientific journal, *New Scientist*, recently ran an interesting correspondence. A reader reported that he heard better when wearing his spectacles. This elicited various lively comments; the sound conduction properties of spectacle frames, the 'cupping' of the pinna of the ear by the ear-pieces, the sense of ease when one sees better were all adduced as potential sources of this advantage. But the original correspondent suspected what many of the contributors to this volume believe, that seeing the mouth movements of the speaker can be as useful an aid to hearing as sound amplification.

Indeed, various studies have been reported from time to time that show spectacled listeners are at an advantage when trying to follow speech with their spectacles on. And not only spectacle wearers; Reisberg, McLean & Goldfield (1987) found that anyone watching the speaker can show a gain in comprehension of difficult to understand, clearly heard speech. And we have known for a long time that in noisy environments people usefully rely on their eyes in trying to understand what is said.

When I met Merald Wrolstad in London in the summer of 1984, I asked him whether *Visible Language*, as it was concerned with language processes and with visual perception, had ever considered investigating lipreading. He said not, but that he would be happy to consider such investigations. Taking him at his (seen and heard) word, this collection is the response to his generous invitation.

I am sad that he has not had the opportunity to tell me what he thinks of it; his long-distance support and encouragement sustained the preparation of this issue, which is really a sampler of some current work concerned with the fact that speaking is seen as well as heard.

Why has lipreading been ignored as an interesting phenomenon? It is not necessary to hear the speaker to understand what is said, though it can sometimes be sufficient. Why investigate such tangential, secondary aspects of speech? There are many reasons. My own motive for studying lipreading is that as a unique, natural, aspect of speech perception that has no auditory/acoustic properties it can enlighten us about the extent and manner of the reliance of speech perception on hearing. A similar motive underlies Massaro's paper. The studies by Dodd and her colleagues lead in a different and interesting direction; they ask how do the immediate memory effects of lipreading differ from those of hearing and of reading? This question has been asked using a different paradigm than that reported here, and with different results (Campbell & Dodd, 1980). These new studies lead us into important and interesting areas.

For developmental psychologists, concerned to uncover the relationship between infant and adult cognitive processes, it can be very illuminating to discover what infants do when they look at speakers. Vinter reminds us of the astonishing imitative powers of the newborn child and asks what might be the relationship between the change in the child's imitation of mouth movements and the development of its cognitive powers, including speech perception and production?

Graves & Potter provide a fascinating example of how cerebral asymmetry affects seen speech. This new demonstration that most of us speak more clearly out of the right side of our mouths prompts the speculation 'why don't we notice it more?' Storey & Roberts show us how we might start to be able to conduct useful investigations of all the phenomena reported here using synthesized stimuli — and without bringing out the big computational guns. This modest and effective putting together of face-image and speech-sound is a first step in a very

useful area of investigation. It is likely to be more helpful to many hard of hearing people than sound amplification alone.

In fact, all the studies reported here will have some relevance to language and deafness, for, by describing and exploring the extent and manner of lipreading function in hearing people, we may gain a more informed view of the potential of lipreading to help those with impaired or abolished hearing. But these studies range further than this. They show how speech is visible in natural, 'unlearned' and effective ways. Visible Language existed before the birth of reading and writing and may, yet, outlive it.

References

Campbell, R. & Dodd, B. 1980. Hearing by eye. *Quarterly Journal of Experimental Psychology,* 32, 85–89.

Reisberg, D., McLean, J. & Goldfield, A. Easy to hear, but hard to understand: a lipreading advantage with intact auditory stimuli. In Dodd, B. & Campbell, R. (Eds.) 1987. *Hearing by Eye: The Psychology of Lipreading.* London: Lawrence Erlbaum Associates, 97–114.

Visible Language in Speech Perception

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.

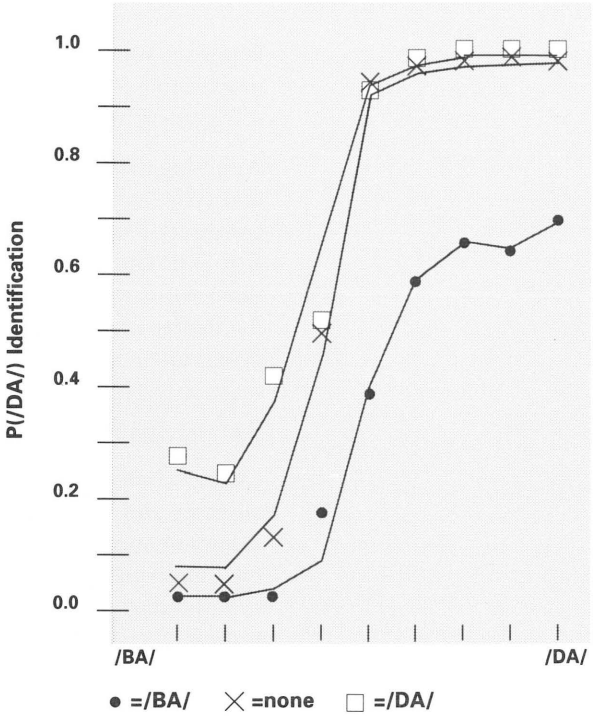


Figure 1 Auditory

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.

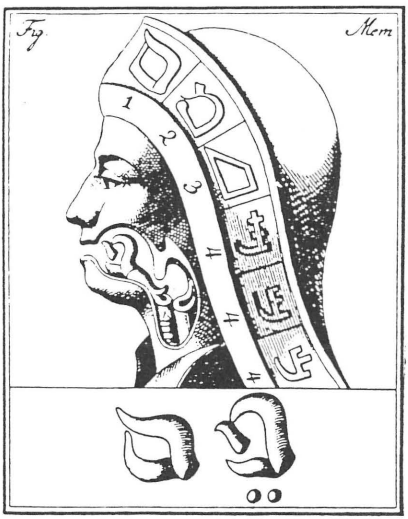
In Crowder's revised model, both the visual and auditory consequences of speech provide featural information at the level of PAS; that is, both auditory and visual speech can place auditory features in PAS. Supposedly, auditory feature selection can occur even in the absence of sound, as in pure lipreading. 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.

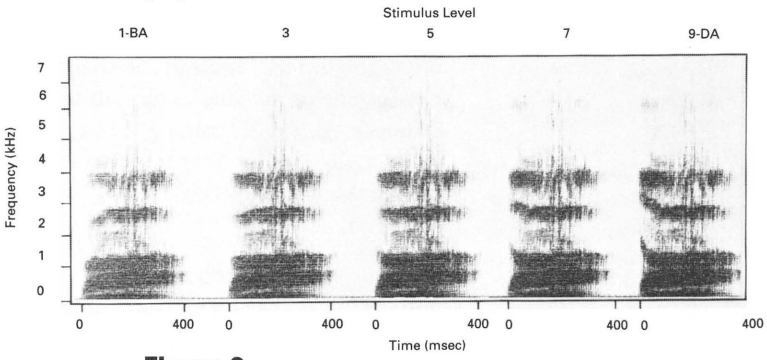


Figure 3

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.

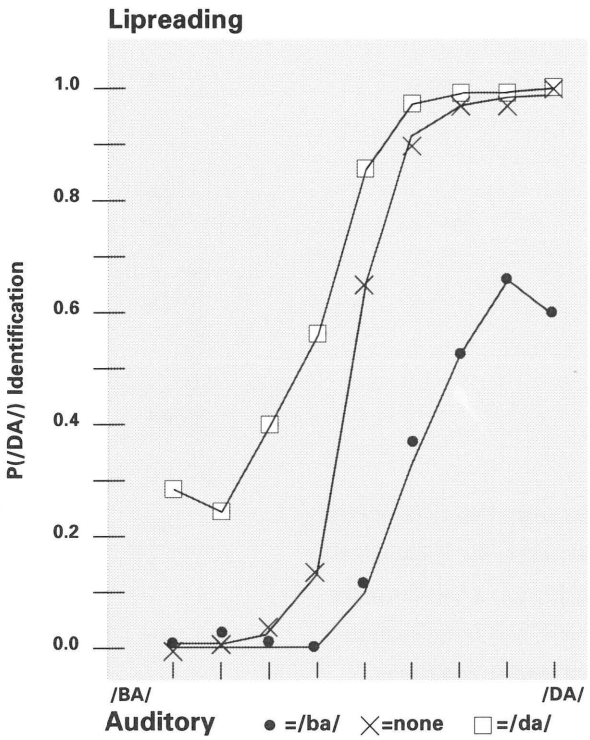
Results

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.



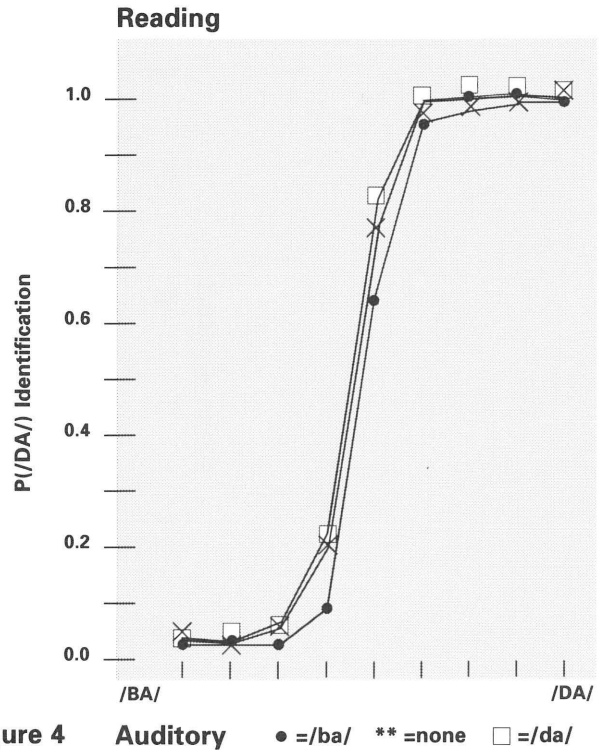


Figure 4 Observed (points) and predicted lines proportion of /da/ identifications as a function of the auditory and visual levels of the speech event. The top panel gives the results for the lipreading condition and the bottom panel for the reading condition. The predictions are for the fuzzy logical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

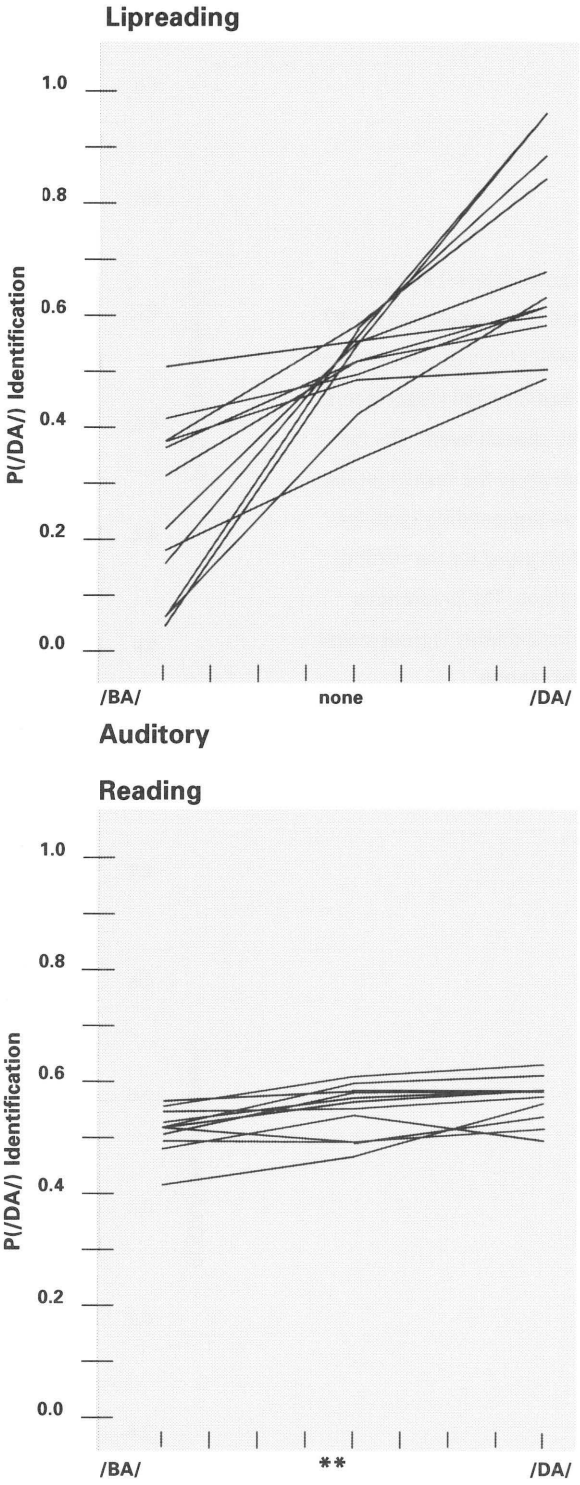


Figure 5 The proportion of /da/ identifications for the 12 individual subjects as a function of the visual level in the lipreading and reading conditions.

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

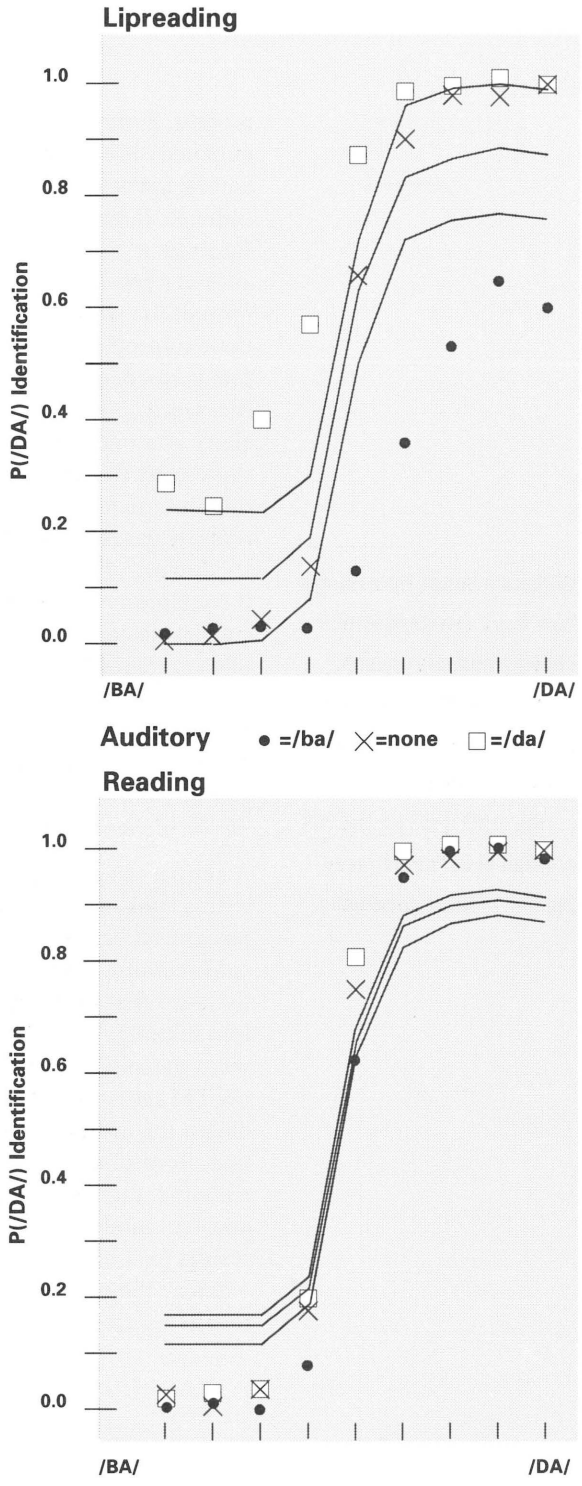

Lipreading
1.0
0.8
0.6
0.4
0.2
0.0
P(/DA/) Identification
/BA/
/DA/
Auditory
● =/ba/
×=none
□=/da/
Reading
1.0
0.8
0.6
0.4
0.2
0.0
P(/DA/) Identification
/BA/
/DA/
Auditory
● =/ba/
** =none
□=/da/

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.

Tracing Lip Movements

Tracing Lip Movements: Making Speech Visible

Ruth Campbell

Department of Experimental Psychology, University of Oxford, South Parks Road, Oxford, OXI 3UD, U.K.

Visible Language XXII, 1
Ruth Campbell, pp. 32-57

Lipreading cannot deliver the phonetic structure of a spoken language very effectively; for no phoneme can be unambiguously identified from lip-pattern alone. Nevertheless, under some circumstances, speech that is not heard, but just seen by lipmovements on a speaker's face, can be understood and recalled verbatim. Moreover, under some conditions, heard speech that is different to that which is seen to be spoken, seems to 'fuse' to produce a different speech percept (The McGurk Effect).

These paradoxical aspects of lipreading and the constraints on the conditions under which lipreading can be helpful or can 'fuse' with heard speech are hard to accomodate within some theories of auditory speech perception. An interactive activation account is offered in which lipreading is considered to provide a phonetic feature - that of seen mouth opening and closing - to the speech analysis system. While such a feature appears to be necessary to account for these effects, it is not yet clear whether such a single seen phonetic feature may be sufficient for effective integration of seen and heard speech in all circumstances.

When people speak, their lips move. Is this natural visible consequence of articulation important in the normal perception of speech, or is it a useless, even disturbing epiphenomenon; best ignored unless we are hard of hearing?

It is clear that lipreading can
A. usefully complement heard speech that has been degraded by hearing loss or noise. It seems to be useful because it can sometimes offer phonetic cues that can be lost in noise. For example, /pa/ and /ka/ differ phonetically in where the plosive part of the sound is made. The first is made by puffing the closed lips open, the other can only be made with the lips apart, the plosion being effected by the back of the tongue hitting the velum. In terms of phonetic gestures needed to make these sounds, this distinction, that of *place* of articulation, is the major way in which they differ, and it is a distinction that has low acoustic energy characteristics and so can be easily lost when listening to speech in noise or with impaired hearing — but of course, it is a distinction that is easily seen.
B. add a comprehension component to *clearly* heard speech where material is difficult to follow because it is conceptually complex or pragmatically unclear (Reisberg et al. 1987).
C. support full comprehension without any heard input at all; for instance where the spoken material is limited by context to a few possibilities that are reasonably visually distinctive (Gailey, 1987). Silently spoken digits are a good example (Cambell & Dodd, 1980).
D. interact with heard speech to give illusory speech perceptions. So, when /pa/ is heard, while /ka/ is seen to be spoken, /ta/ is often reported as the heard sound (McGurk & MacDonald, 1976).

We know, further, that efforts to describe what makes a good lipreader have been unsuccessful in pinpointing any *particular* attribute. On the whole, what makes for good lipreading is what makes for good understanding of heard speech; an awareness of speech structures and their possibilities — and a flexible and powerful intelligence that can back such awareness (Gailey, 1987; Jeffers, 1967; Kitson, 1915).

We also know that infants, from their first days, are biologically predisposed to be sensitive to face movements and imitate them readily (Meltzoff & Moore, 1982). This sensitivity transmutes into more complex perceptual patterns which could come to be intertwined with speech perception (Dodd, 1987; Mills, 1987; Vinter, this volume).

Speech is probably best acquired and utilised *bimodally*. What one sees of the speaker seems to form a natural, sometimes a necessary complement to what is heard.The acoustic quality of heard speech in everyday settings often leaves much to be desired, yet we are able to understand such material with little effort.

Not only are we able to understand spoken speech distorted through various transmission devices (telephones, PA systems, wind, whispers) but usually the auditory attributes of natural speech bear no immediate invariant relationship to the natural 'units of speech' — phonemes, words. Instead, the articulatory and auditory context, as well as the lexical context and linguistic and semantic knowledge of the listener *set* the perception of the parts of an utterance in complex and interactive fashions. It is in this swirl of natural speech sound identification that lipreading can start to be understood and needs to be explained.

Theoretical Approaches to Lipreading

There are a number of plausible theoretical ways to accomodate lipreading within the context of natural speech perception.

The Motor Theory of Speech Perception

The simplest is perhaps the motor theory of speech perception. According to this proposal (Liberman & Mattingley, 1985), the invariance of heard speech perception in highly variable auditory contexts resides in the speech *gesture* — in the invariance of the highest form of speech articulation. The identification of a particular phoneme is a function of the ability to produce the phonetic structure that characterises it. In order to explain how a speech might be perceived as a speech sound, such theorists tend to make use of J. J. Gibson's theories of 'direct perception' (1950; 1966). The natural

invariance of the speech gesture reflects an emergent sensitivity to such meaningful units in the heard/spoken environment; a sensitivity that is genetically given. We hear speech sounds for what they are, whoever says them, however they are said, through a perceptual system which allows the abstraction of such identities as 'higher order perceptual invariants'. It was by such higher order invariance that Gibson claimed that we perceived visual aspects of the environment such as distance and shape. In this context it is natural that lipreading should have a place. If the invariant unit of speech perception is the gesture, not the sound made, then 'naturally', the perceptual analysis of the gesture will be as effective when it is seen as when it is heard. But, almost by definition, the mechanisms of direct perception are hard to instantiate. It becomes difficult to assess the explanatory, rather than descriptive, power of this aspect of motor theory.

Nevertheless, the motor theory of speech perception can direct us to many relatively unexplored aspects of speech perception. What would it predict, for example, to be the *lipreading* skills of the person struck mute by a brain lesion affecting the central cortical sites of speech production? Such central motor disorders do not always dissociate cleanly from perceptual ones; impairment of speech production is often intimately related to speech comprehension impairment. We are working on the neuropsychology of lipreading at present, in our laboratory.

Integration Theories of Speech Perception

Massaro's paper (this volume) suggests a different approach to speech perception, one which he has vigorously pursued for several years. Massaro's achievement is to indicate the *flexibility* of speech integration processes. His goal is to effect a powerful, predictive description of the metric whereby phonetic processing occurs. His solution is the specification of a fuzzy logical integration model which allows sensory information to be handled in a probabilistic and context–sensitive way. This approach invests computational power in the properties of the sensory systems themselves. Lipreading adds yet another component to the speech-integrative system.

One potential problem with this approach is that it may fail to discriminate effectively between natural and unnatural percepts. For example, the methodology that this approach uses requires the perceiver to make sense of a range of unnatural inputs. The extent to which he can do this will certainly tell us about how such decisions can be made — but runs the risk of equating such decisions with natural speech perception. Natural speech perception may well use processes that can be demonstrated by presenting unnatural patterns of stimulation; but the relationship is not transparent.

In fact, one critical test of an adequate theory of lipreading in speech perception must be whether it can distinguish, as normal hearers do, between 'natural' and 'unnatural' (automatic and voluntary, perhaps?) combinations of input. One must also ask, can the effects of reading written material and reading mouth movements be distinguished easily within the theory? If they cannot, then the theory will lack power at one of the most important points at which it should exert it. When one reads the syllable /ga/ while hearing /ba/ one does not experience an illusory, heard /da/; but when the visual stimulus is lipread, one can (McGurk & McDonald, 1976). An integration model (taken at face value) might suggest it might. This coarse example suggests that the highly flexible integration theory exemplified by Massaro's work may, nevertheless, not capture a crucial aspect of lipread speech, but may serve instead as an 'overmodel'; a description of the combinational rules involved in each and every sensory identification process.

Special Purpose Mechanisms: Can seeing the speaker add to hearing?

The particular problem posed by the McGurk illusion, where discrepant heard and seen speech can sometimes give rise to perceptual fusions and blends, has given rise to some thoughtful and provocative speculations by Summerfield (1987).

Rather than develop either a powerful, all-round model of heard and seen speech perception, Summerfield takes a more modest tack and asks what particular 'add-on' components might make an auditory theory of speech

perception lipread? A number of possibilities are considered; none of them completely viable, but all are systematically explored.

The McGurk illusion is the strongest evidence so far that speech that is seen and heard can generate a clear speech percept, different from that predicted from vision alone or from audition alone. The immediacy of the percept suggests that the integration occurs at the most basic level of speech recognition; in the identification of the spoken phoneme. (It is worth noting that this statement remains to be experimentally tested.) Since it is primarily the phonetic feature of *place* of articulation that can be seen on the lips, perhaps vision 'dominates' hearing for the extraction of such phonetic information; the integrated percept reflecting a best fit between the visually and the acoustically specified /ga/ and /ba/. This cannot be the whole story. Seen /ga/ and heard /ba/ fuse to generate / da/ or /ta/. But when seen /ba/ and heard /ga/ are synchronised, the illusory percept varies; /bda/ or /bga/ etc. can be reported (see Massaro, here). So if place of articulation is specified by the lips it is specified in a highly conditional manner.

Indeed, the 'ifs and buts' needed to describe the vagaries of illusory blend percepts lead Summerfield to consider other integration possibilities. Perhaps lexical identification, rather than phoneme identification, is the source of effective lipread-auditory processing? Why propose a word recognition model that sidesteps the phoneme? Because natural speaking causes troublesome problems of phoneme identification, due to coarticulation. Coarticulation describes the necessary way in which fluent speech causes the production of a specific speech sound to vary as a function of what is said before and after it. The movement and inertial constraints on the mouth, vocal cords and tongue are such that the natural production of the sounds of a word cannot be achieved in a simple, serial manner, "like pearls on a string". Consider speaking the two words, "bath" and "both". The only perceived distinction is in the interconsonantal vowel; but this changes the articulatory/acoustic specification of the consonants fore and aft of it. Yet these differences are not *heard* in running speech; the speech identification system has

The movement and inertial constraints on the mouth, vocal chords and tongue are such that the natural production of the sounds of a word cannot be achieved in a simple, serial manner, "like pearls on a string". Consider speaking the two words "bath" and "both". The only perceived distinction is in the interconsonant vowel; but this changes the articulatory/acoustic specification of the consonants fore and aft of it.

managed to absorb and take account of such coarticulatory effects. If you watch yourself in the mirror saying "both" and "bath" you will see that the lip positions of the consonants, too, are quite different in the two words. Coarticulation is as much of a problem for seen as for heard speech. Klatt's (1979) model of lexical identification on the basis of 'delayed commitment' to a particular interpretation of the auditory speech signal, can, Summerfield shows, be extended to include lipreading. Lexical identification, on this model, is achieved independently of phoneme identification– for phonemes are hard to specify*a priori.* Instead, the model allows concatenation of phonetic features into units that will vary with the specificity of the word to be recognized; that is, there can be a variety of 'units' including direct sound spectral specification, that can characterise a particular lexical item. Lipreading could work in such a 'direct access' model through lipreading-specified lexical representations, corresponding to the aural ones. But the problem with this approach is that it fails effectively to account for the gain that lipreading can give to auditory speech understanding, for the word specified by the lips will have similar coarticulation characteristics and problems as the word specified by ear. While this might mean that the invariant speech percept has its source at a level beyond the phoneme, it makes it hard to see just how lipreading can improve listening performance.

How can the lipread phoneme uttered in running speech achieve phonemic invariance? How do we know that /t/ was spoken when 'tooth' or 'teeth' is the word uttered? A glance in the mirror should confirm that the lip patterns of the two 't' sounds are utterly unlike each other because of the articulatory pattern of the aftercoming vowel. Here Summerfield suggests an articulatory theory may be useful. One of the skills of the listener-speaker may be the knowledge of vocal tract configurations, for this is reflected in the speaker's own ability to produce invariant speech sounds under different articulatory conditions. From this 'deep' dynamic knowledge will come awareness of the seen as well as the heard kinematic reflections of this set of movements. Very similar theories have been expounded to explain the developing child's perceptual sensitivity to the dynamics of hand and arm actions in

One of the skills of the listener- speaker may be the knowledge of vocal tract configurations, for this is reflected in the speaker's ability to produce invariant speech sounds under different articulatory conditions.

holding, catching, reaching, lifting (e. g. Mounoud & Hauert, 1982). An explication of this 'deep motor' approach to the recognition of auditory-visual vowels can be found in Summerfield and McGrath (1984).

Summerfield, then, suggests a range of 'add-on' components to account for lipreading in normal speech perception. These components are 'customized' to fit the particular model of auditory speech perception. The implication of this work is that each and every extension of each model may play a part in auditory-visual speech perception. However, as Summerfield's chapter title makes clear—these are *preliminaries* to a theory of audio-visual speech perception. Could a more integrated theory of vision in speech perception be attempted?

Lipreading in an interactive theory of speech perception: Can TRACE lipread?

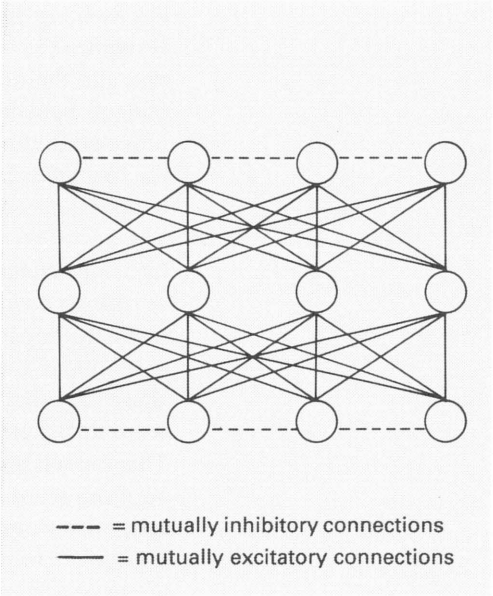
Interactive Activation Models

Twenty years of research on the recognition of written words have made clear that although the identification of letters and letter features is important, words can often be identified when the letters in them cannot. Text comprehension can set word identification, too. Letters in words are better identified than letters in non-word letter strings. Thus not only are bottom-up and top-down processes important in word recognition but these processes interact systematically. McClelland and Rumelhart (1981) and Rumelhart & McClelland, (1982) proposed an explicit interactive activation model to account for these phenomena. This model was the mother of a series of recent explanatory and predictive devices with similar formal properties. These have been used to explain, among other things, speech production (Stemberger, 1985; Dell, 1985) and auditory speech perception (McClelland and Elman's TRACE model of Speech Perception, 1986).

What are the formal properties of these models? Horizontal levels of representation are posited at increasingly molar levels from the 'purely sensory' to the cognitive. Thus for auditory speech perception a *phonetic*, a *phonemic* and a *lexical* level of representation are proposed. Connectivity between and across item representations is organised as follows: lateral inhibition is the dominant

A fully connected network with excitatory and inhibitory connections

Figure 1



connection type between represented items within each horizontal layer, that is, mutually inconsistent units inhibit each other. Across levels, however, the pattern of interaction is different. Here, mutually consistent units can excite each other (the original model of visual word recognition includes both excitatory and inhibitory cross-unit connections, but such inhibitory cross-level connections tend to make the model too inflexible in dealing with poorly and partly specified information). Finally, patterns of excitation and inhibition across and within layers work in cascade; there is temporal recruitment of these processes. Thus a particular stimulus array generates a temporary and highly dynamic pattern of excitation across the network. This then settles to a stable distributed pattern of activity. It is this stable pattern of excitation — rather than the firing of a particular 'node' in the array — that is the concomitant of categorisation or identification of a particular stimulus.

Such highly interative, distributed models, might, at first sight, appear to be *too* interactive to generate anything but noise, but, as recent theoretical and simulation developments show, they can be both precise and practical in implementation. They behave like human beings (see McClelland and Rumelhart, 1986).

For auditory speech perception an important component is required in addition to the three levels of representation and their interconnections. This is the TRACE component. Spoken words take time to say and this is reflected in their characteristic recognition time. One of the first speech recognition models to take seriously the temporal characteristic of spoken language was Marslen-Wilson & Tyler's COHORT model which showed how decisions concerning word identification start to operate as soon as a word starts to be spoken, rapidly and automatically constraining the identification of a possible word depending on the size of the cohort of words that share the same phonemic features up to a critical, unique word decision point (see Marslen-Wilson & Tyler, 1981). This model, however, cannot backtrack; it is a feature of auditory word recognition that we are easily able to recognise words whose initial phonemes may have been misspoken or misheard. Then, if co-articulation effects are to be accomodated in a sensory, rather than a motor theory, it is necessary for the speech recognition process to take account of aftercoming speech context, as well as prior context, in order to achieve invariant phoneme perception. An auditory theory of speech perception should incorporate a delay window to account for such right-context as well as mispronunciation effects. The TRACE serves this purpose. TRACE is not an acronym; it is a literal description of the sustained state of activation of the system; the TRACE, that is the pattern of activation corresponding to a not-yet resolved stimulus pattern, persists until categorical recognition of the speech sound is achieved. In this crucial sense, this model is not an 'on-line' model of speech recognition, in the way that the COHORT model is; rather it suggests, along with others (e.g. Crowder, 1983), that the distinction between perception and immediate memory in speech sound processing is a blurred one.

To summarize the general history of this type of model: a visual word recognition model that has different levels of representation all of which are fully interactive each with the other, has been extended into the time dimension, both in its representational and activation components, in order to process heard speech. *Representations* need to be temporally organised (for example, in the specification of

the phonemes required to distinguish 'god' and 'dog'), but also the *state of activation* of each unit at each level persists as long as and until a final, categorical, decision can be made.

Lipreading by TRACE?

Now, how could a TRACE model accomodate lipreading? Let us remind ourselves what such a model should achieve. It should

1. explain how lipreading can aid noisy speech perception; it should also suggest how clear speech can be helped by lipreading and how silent lipreading can be achieved.

2. predict the specific patterns of interaction of the blend and fusion illusions. In particular, it must distinguish between the effect of a seen /ba/ and a heard /ga/ (/bga/, /bda/, etc. . .) and the opposite conjunction (which leads to a fusion like /da/).

Additionally, if TRACE works for lipreading, we might expect some similarity between heard and lipread material in terms of a close relationship between perception and immediate memory processes for lipreading and for hearing.

Can TRACE do this? Clearly in such fully interactive models it would be possible to introduce one or several lipreading features. So let us start with the most basic proposal of all; that the only visual (lipread) feature that the model permits is that of mouth opening and closure. Where should this feature be introduced? Let us, again for simplicity, introduce it at the bottom; as an additional *phonetic feature*. Note that because TRACE is a theory in which identification is a function of distributed processes at different levels of representation, this will mean that at all other levels of representation than the phonetic feature level, the lipread information will have some effect; it will leave a trace.

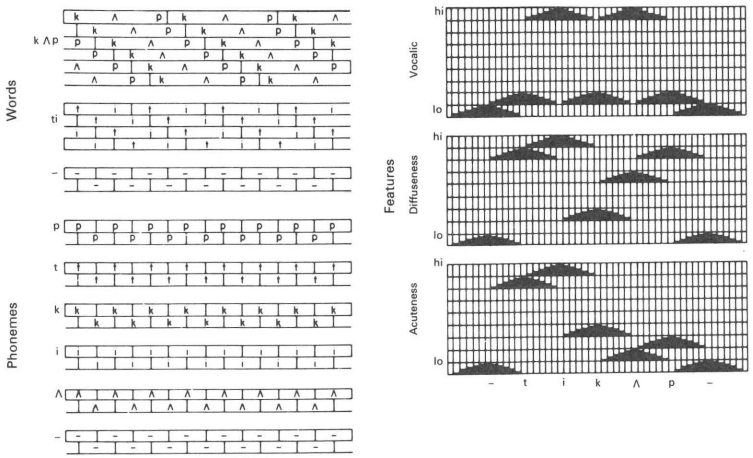
General Effects of Introducing a Visible Phonetic Feature: Mouth Closure

Simply by virtue of the addition of a seen component at the feature level, the state of stable activation corresponding to identification of a word will be more efficiently achieved when one can see as well as hear the

speaker; an additional feature, relevant to speech perception, is added to the system. In this way, the gain of lipreading a clearly heard speaker can be indicated. Where speech is noisy, the phonetic feature specification of the *acoustic* phonetic features will lead to unstable, persistent activation of many possible categorical speech sounds; under these conditions a clearly seen face will improve recognition for those sounds that can be distinguished visually.

How would this work in detail?

Figure 2



A subset of the units in TRACE II. Each triangle represents a different unit. The labels indicate the item for which the unit stands, and the horizontal edges of the rectangle indicate the portion of the Trace spanned by each unit. The input feature specifications for the phrase "tea cup", preceded and followed by silence, are indicated for the three illustrated dimensions by the blackening of the corresponding feature units.

Figure 2, from McClelland & Elman, shows how the auditory input "teacup" corresponds with three levels of unit representation in TRACE. Only the pattern of excitation at the phonetic level is indicated for clarity. Furthermore, only three acoustic feature dimensions are indicated for simplicity. The full simulation model used seven acoustic dimensions including voicing, consonantality, power and burst, as well as the three indicated here.

Consider the phonemes /k/ and /p/ in the figure. In terms of the three acoustic phonetic feature dimensions shown they are similar, differing only in a small change in the level of activation for diffuseness and acuteness. What would the pattern of activation be like for a seen and heard 'noisy' "cup"? Remember the principles of

Remember the principles of activation are these; within a level, units that are inconsistent have mutually inhibitory connections. Across levels, units that are mutually consistent have mutually excitatory connections.

activation are these; within a level, units that are inconsistent have mutually inhibitory connections. Across levels, units that are mutually consistent have mutually excitatory connections. In noise, the acoustic feature dimensions for identifying /k/ and /p/ would not be inconsistent for some would be too poorly specified, and there could not be mutual inhibition between these feature patterns on this basis. So if the recognition of the spoken word were dependent solely on acoustic inputs, the failure of inhibition at the phonetic feature level would mean that a large range of possible phonemes and phoneme combinations would all be relatively excited, and the identification decision on the word would be delayed, or even wrong. Typically, this characterises the identification of heard words in noise.

But the visual feature of mouth-opening is consistent with a range of phonemes (e.g., k, g, t, d), and inconsistent with others (b, th, m, p). Thus, at the phonetic feature level the *seen* sounds /p/ and /k/ will exert maximal mutual interference. So 'pup', 'puck', or 'cuck', cannot be activated by lipread "cup". In this way, *some* distinctive feature information can travel up the system when lipreading noisy speech. This has been refined by lateral inhibitory mechanisms at the phonetic level *without* necessarily being sufficient to enable particular phoneme identification to take place. If, at the lexical level, the general pattern of activation of phonetic features and of possible phonemes is more consistent with 'cup' than with any other word, the principle of mutually consistent units causing excitation across different levels enables the phonemic and phonetic aspects of that word to become further activated. Thus, top-down information affects the categorical perception of lower level units. Lexical superiority effects should arise in lipreading noisy speech, and common sense suggests that they do. But it is important to remember that TRACE allows a relatively stable pattern of activation to be established sufficiently to identify a word without necessarily being sufficient to identify the constituent phonemes in the word. Klatt's (1979) solution to the problems of coarticulation (and Summerfield's extention of this direct lexical access theory to the domain of lipreading) can be reinterpreted. Interactive activation theory allows words

to be identified without full identification of their constituents. Because lipreading can add a featural component to the phonetic activation pattern, it will aid hearing under many conditions, but particularly when auditorily-based phonetic feature components become less discriminating, through noise or disease.

Improving Lipreading Comprehension

If there is only one seen phonetic feature but several acoustic ones (which may or may not be orthogonal to each other), then the conditions under which lipreading could affect heard speech will be more constrained than the conditions under which hearing could affect lipreading. And they are (Easton & Basala, 1982). It is conceivable that all or any acoustic inputs, if congruent with defined speech percepts, will help the lipreader. One example of this might be the effect of lipreading with an auditory pulse train (Rosen, Fourcin & Moore, 1979). When an acoustic signal, corresponding to the activity of a speaker's vocal cords, is heard in synchrony with the speaker's lip movements, there is a marked improvement in lipreading comprehension for connected discourse over silent lipreading alone. Under these conditions only two potential phonetic feature dimensions are available to the perceiver; mouth-opening and voicing, yet, not only are these feature dimensions useable, they often seem to generate the illusion of hearing noisy speech, which persists until one loses sight of the speaker (the auditory pulse train alone cannot support speech comprehension).

Interactive activation theory allows words to be identified without full identification of their constituents.

The relatively greater effect of heard inputs on speech identification than of seen (lipread) ones can, in an interactive activation model, help to explain why the best lipreaders seem to be people who are generally good at language skills. They could use information at all levels of representation more efficiently to improve the value of lipreading. There may still be individual differences, of course, in the precise level of efficient unit representation (see Gailey, 1987).

It should also be clear how it is possible to lipread rather well when context constrains the number of possible words that could be spoken; in other words, how top-

down processing improves lipreading. Under these conditions only a subset of potential words are possible targets for identification and hence activation. Thus digits can be easily lipread from silent speech, as can footballer's (English) expletives (as seen on TV). Such phenomena suggest that a lipread phonetic feature detector can sometimes provide sufficient information for effective speech recognition by sight. Whether born-deaf and post-lingually deafened people show similar activation patterns for words that both groups lipread well is a subject for experimental research.

Audio-Visual Fusion and Blend Illusions

The simple principle of mutual inhibition between inconsistent feature units can also help to explain the contingencies of the audio-visual (McGurk) illusions. An open mouth is consistent with a range of consonantal speech sounds including /k/, /g/, /t/, /d/. It is also consistent with all vowels. A closed mouth is not consistent with any of these but with the consonants /p/, /m/, /b/. Both lip positions are consistent with no sound at all — the sound of silence cannot be *unambiguously* seen.

When /pa/ is heard and /ka/ is seen, the 'classical' fusion reported is that of a heard "ta" (McGurk & MacDonald, 1976). Those authors suggested that place of articulation of the illusory consonant was calculated by a compromise between the seen (front of mouth) and the heard (back of mouth) place of articulation of the tongue to a more intermediate one. /ta/ is an alveolar sound, made by placing the tongue just behind the dental ridge, at the front of the hard palate — though there are allophones of this sound with more variable places of articulation. But it may not be correct to suggest that /ta/ and /ka/ are systematically distinguished from each other solely by position of the tongue with respect to the front-to-back of mouth dimension (jaw drop may be a feature). In any case, in terms of an interactive model, the first stage of the processing of the dubbed seen and heard syllables is that of phonetic feature unit activation.

Figure 3

Phoneme values for /p/, /t/ and /k/ as used in TRACE II (McClelland & Elman, 1986) with mouth-opening added.

Feature	Phoneme		
	p	t	k
power	4	4	4
vocalic	1	1	1
diffuse	7	7	2
acute	2	7	3
consonantal	8	8	8
voice	1	1	1
burst	8	6	4
seen feature			
mouth open	1	8	8

What are the acoustic activation patterns of the consonants, 'p', 'k' and 't' in terms of such phonetic feature activation? Figure 3 shows the full values used in McClelland & Elman's TRACE simulation of auditory speech discrimination (McClelland & Elman, 1986, p. 15). Now we can ask how, when 'pa' is heard and 'ka' is lipread, does 'ta' seem to have been spoken? How, in other words, can the phonetic activation pattern needed for a /t/ decision be mimicked by a combination of /p/ and /k/ features?As we saw in the exampleon the opposite page, the activation patterns for acoustic phonetic features are quite similar. The differences (the inconsistent and therefore inhibitory features) are, first, acuteness; 'ta' is inconsistent with 'pa' and 'ka' and second, diffuseness; 'ka' is inconsistent with 'pa' and 'ta'. Note also, that on one dimension, burst, 'ta' is intermediate in value between 'pa' and 'ka'. In terms of the seen feature of mouth opening, 'pa' would be inconsistent with 'ka' and 'ta'. Presumed values are shown in the table.

Superficially, the inconsistency on two acoustic dimensions might suggest that these syllables are unlikely to be confused with each other. But the acoustic correlate of acuteness is high frequency spectral energy; that is, the energy that is most likely to be lost with small amounts of noise (put another way, "pa", "ka", and "ta" are all likely to be confused with each other when even small amounts of white noise are added to the acoustic

signal). So, with a little noise, acuteness is unlikely to be a reliable dimension for identifying these phonemes.

Now if 'pa' is heard, while 'ka' is seen, 'ta' could be the predicted percept if:
1. a small amount of white noise accompanies the acoustic signal, rendering the potentially distinctive feature of acuteness nondiscriminating.
2. the *visible* feature of 'open-mouth' is activated. This will *inhibit* /p/ activation by the principle of mutual inhibition between mutually inconsistent units.
3. the remaining feature that distinguishes /k/ and /t/ — that of burst quality — reflects relatively more inhibition of the low value /k/ by high value /p/ than of medium value /t/ by /p/.

This point is worth stressing; the apparent 'compromise' decision on place of articulation results from relatively greater inhibition between the more extreme feature values and less between each of these extreme values and a middle one. Lateral inhibition at the feature level achieves apparent compromise at the phoneme level. It may be worth noting that the same percept — 'ta' from 'pa' and 'ka' has been reported when 'pa' and 'ka' are heard in each ear. Presumably the 'open mouth' feature is not the only one that can produce a 'ta' percept. The *converse* pattern of input; seen 'pa' synchronised with heard 'ka' cannot give rise to the same, illusory 'ta'. Closed mouth detection generates inhibition between 'pa' and 'ka' or 'ta'. The perceptual system resolves this, usually, by ascribing the discrepant inputs to different times slots; 'pka', 'pta', being common reports of this stimulus configuration. Because vision and hearing are mutually inhibitory for this stimulus configuration at the feature level, such blends, (i.e., more than one perceived sound, rather than the unitary fusion illusion, where only one sound, usually different from the one that was acoustically present, is reported) are the *only* permissible percepts for a lipreading TRACE model.

Place of articulation is not determined more by vision than audition, but vision contributes to the pattern of interactive activation at the phonetic level. This allows place of articulation to emerge as a seen feature in a systematic and predictable way.

Place of articulation is not determined more by vision than audition, but vision contributes to the pattern of interactive activation at the phonetic level. This allows place of articulation to emerge as a seen feature in a systematic and predictable way.

Rate of Articulation

Seeing the mouth open and close can inform the speech processing system about another phonetic dimension. The identification of voiceless plosives, like /pa/, is contingent on the perceived auditory rate of speech (Summerfield, 1981). Green & Miller (1986) have shown that the perception of speech rate does not have to be auditory to affect the categorisation of a heard speech sound as /pi:/ or /bi:/. The rate of seen lip-movement can shift categorical perception of the voiceless plosive in a similar manner to the heard rate of speech. This highly context-contingent phonetic effect, can, of course, be accomodated by TRACE as comfortably as other effects of coarticulation, through the delayed commitment principle that TRACE embodies.

Place of articulation is not determined more by vision than audition, but vision contributes to the pattern of interactive activation at the phonetic level. This allows place of articulation to emerge as a seen feature in a systematic and predictable way.

Immediate Memory Processes: TRACE, PAS and Other Phenomena

If TRACE is a good model in which to accomodate lipreading in speech perception, it should indicate ways in which lipread and heard material might show similar immediate memory characteristics. In particular, if it is a powerful model, it might distinguish between such effects and those for written material. What distinguishes the immediate memory processing of written and heard material? While several lines of investigation are being pursued (see, for example, Dodd, Oerlemans & Robinson, this volume), one route has been extensively travelled. This is the investigation of immediate list recall, which shows a very robust effect of auditory recency. The last item of a heard list is easier to remember than earlier list items; this recency effect is less marked, or absent altogether, when the list items are read, rather than heard. Auditory recency can be eliminated by a heard speech sound after the end of the list. This makes little difference to written list recall. Because this suffix effect is not dependent on lexical status of list and suffix, it can be considered precategorical. The combination of auditory recency and suffix effects led Crowder & Morton (1969) to suggest that auditory lists leave a record of their *acoustic* properties in the cognitive system — a precategorical acoustic store (PAS). If this were a sensory-acoustic trace then lipreading should not produce recency and suffix

effects similar to those for hearing spoken lists. But it does (Campbell & Dodd, 1980, 1982, 1984; Greene & Crowder, 1984). Moreover lipread lists are affected by heard suffixes, and heard lists by lipread suffixes; and these effects are highly specific to those input combinations (Gathercole, 1987; Campbell, 1987).

TRACE suggests that, because of deferred commitment to phonetic decisions, just such a record can persist, which will be instigated by phonetic rather than acoustic activation. If the lipread feature of mouth-opening has access to the TRACE system then similarity and interactivity between heard and lipread recency and suffix effects are to be predicted. They will occur because, in recall, TRACE activitation allows a 'second look' at the stimulus array, and the most recently presented material will be relatively more accessible in this phonetically active state. Since reading does not activate such a feature-based, persistent trace, written lists do not *usually* show recency (but see Campbell, 1987 and Massaro, this volume).

As an aside, it may be worth noting that the principle of deferred commitment embodied in TRACE, might lead to recency effects not only for heard lists, but also for other material that demands that categorisation wait on serial presentation of featural information. Campbell, Dodd & Brasher (1983) showed that recall of serially presented lists of unnameable arrow shapes showed recency that depended on the order of display of the discriminating arrow fleche compared with the non-discriminating arrow shaft for each item as it was shown. When the fleche was presented before the shaft, (no deferred commitment to a categorical decision was needed to identify and recall this type of item), no recency was observed. When the shaft was presented before the discriminating fleche, recency occurred.

The deferred commitment principle embodied in TRACE would suggest, moreover, that auditory/lipread recency and suffix effects do not reflect fixed-size storage capacities for phonetic material but, rather, that recency and suffix effects will vary with the extent to which deferred commitment characterises the operating system. The recognition of spoken words, because it is so automatised in skilled hearers, may appear inflexible, giving the

impression that fixed storage size is a characteristic of the immediate memory system that gives rise to recency/suffix effects.

Is Mouth Opening Enough?

The simple detection of mouth opening and closing was proposed as a sufficient feature for a TRACE model that lipreads. But what is the evidence that mouth closure, rather than some other aspect of lip-movement, is the critical feature? Could mouth closure be one of several features that the lips can offer to speech recognition? Summerfield (1979) examined a range of visual manipulations of seen mouth movement to see whether they could be distinguished in clarifying noisy heard speech. The control condition, which provided a significant lipreading advantage over unseen speech, was a synchronised videotape of the lower (full) face. Significantly useful, also, were disembodied lips—that is the lips painted with ultraviolet reflecting paint, videotaped under ultraviolet light. The tongue and teeth are *not* visible in this condition. There was a significant difference between lipreading gain for the control and the disembodied lips condition; *something* is missing from the disembodied lips that can help in the visual clarification of heard speech. However, two conditions gave *no* lipreading advantage. These were the movement of the illuminated four corners of the mouth and the presentation of a visible annulus whose inner diameter varied with vocalisation of the stimulus. Both these conditions display a visual stimulus that corresponds, formally, with properties of the heard speech. But neither of these displays indicated lip closure.

So it seems that lip-closure is a necessary feature of effective lipreading. But is it sufficient? Are more features needed? Kuhl & Meltzoff (1984) examined 19 week-old infants' sensitivity to face-voice synchrony. The infants were shown sequences of the vowel sounds 'ah' or 'ee' being spoken. These were synchronised to the same or the other vowel being produced on the auditory channel. The infants looked significantly longer at the correctly synchronised than the wrongly dubbed face and voice. The investigators confirmed that the identity of the spoken vowel was important, rather than just the temporal

It would seem that lip-shape, as well as lip-opening can be important in extracting speech from faces.

synchronisation of seen and heard sounds, by replacing the vowel with a pure tone signal that maintained the duration of the spoken vowel, its amplitude envelope over time and its synchronisation to the visual signal. Now the infants showed no preference for one face over the other, but responded in the arbitrary way that they looked at wrongly dubbed faces.

Since only spoken vowel shape was manipulated in this experiment, it would seem that *lip-shape*, as well as *lip-opening* can be important in extracting speech from faces. A further piece of evidence suggests that, indeed, lip-shape can be a useful and important feature of phonetic perception. Summerfield & McGrath (1984) examined the integration of discrepant seen and heard vowel sounds. If lip shape were relatively uninformative about vowel identity—if it simply signalled a possible vowel,—then one would not expect that auditory- visual fusions or blends should occur when viewing one lip shape and hearing another. Yet such blends do occur when viewing one lip-shape and hearing another, though not always in the clearcut, categorical fashion that they do for some stop consonants. Storey & Roberts' simulation (this volume) assumes lip-rounding has a role to play, too.

Should lip-shape be considered as a phonetic feature orthogonal to lip-closure in a speech recognition system? Or are lip-closure and lip-rounding nonindependent aspects of the same, essential feature? It is possible to think of vowel movement from 'ee' through 'ah' to 'oo' as following the movement from lip closure to lip-opening. But this question is worth careful experimental investigation. So too are questions concerning the necessity of tongue and teeth visibility. In Big Nambas, it seems, the bilabial – /pa/ is distinguished from a very similar sound produced by smacking the tongue, voice-lessly, against the outer part of the upper lip; a highly visible, but hardly hearable distinction (Ladefoged, 1985).

Here, then, are fruitful areas for further research. An interactive theory of the role of lipreading in auditory speech perception can focus and drive the search for an understanding of the processes involved in speech perception, whether it is seen or heard. TRACE theory seems,

at present, to be the most powerful and accomodating of all theories of speech perception. With a visual (lipreading) component it can be usefully extended and can provide an answer to some of the puzzles that face us when we listen with our eyes

About the author

Ruth Campbell is a University Lecturer in Experimental Psychology at the University of Oxford, England. Her research has been on the neuropsychology of lipreading and, with Barbara Dodd, on immediate memory for lip-read lists. While these seem typically obscure and trivial subjects for psychological research they have turned out to be useful in indicating just where the conceptual line needs to be drawn when considering what underlies the perception of auditory and of visual material. Her present research interests also include deafness and cognition as well as aspects of the processing of facial information (other than reading speech from them) and also reading and writing. Her own hearing problems might have contributed to these research interests, though she is a bad lipreader!

References

Campbell, R. 1987. Common processes in intermediate memory. In Allport, D. A., MacKay, D., Prinz, W. & Scheerer, E. (Eds.) *Language Perception and Production: Common Mechanisms in Listening, Speaking, Reading and Writing.* NY: Academic Press, 131–149.

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

Campbell, R. & Dodd, B. 1982. Some suffix effects on lipread lists. *Canadian Journal of Psychology,* 36, 509–515.

Campbell, R. & Dodd, B. 1984. Aspects of hearing by eye. In Bouma, H. & Bouwhuis, D. G. (Eds.) *Attention & Performance,* 10, L.E.A., Hillsdale, 300–311.

Campbell, R., Dodd, B. & Brasher, J. 1983. The sources of visible recency; movement and language in immediate serial recall. *Quarterly Journal of Experimental Psychology,* 35A, 571–587.

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

Crowder, R. G. & Morton, J. 1969. Precategorical Acoustic Storage (PAS). *Perception & Psychophysics,* 5, 365–373.

Dell, G. 1985. Positive feedback in hierarchical connexionist models: applications to language production. *Cognitive Science,* 9, 3–23.

Dodd, B. 1987. The acquisition of lipreading skills by normally hearing children. In B. Dodd & R. Campbell (Eds.) *Hearing by Eye: The Psychology of Lipreading.* London: Lawrence Erlbaum Associates

Dodd, B. & Campbell, R. 1984. Non-modality specific speech coding. *Australian Journal of Psychology,* 36, 171–184.

Easton, R. D. & Basala, M. 1982. Perceptual dominance during lipreading. *Perception & Psychophysics,* 32, 562–570.

Gailey, L. 1987. Psychological Parameters of Lipreading skill. In B. Dodd & R. Campbell (Eds.) *Hearing by Eye: The Psychology of Lipreading.* London: Lawrence Erlbaum Associates. 115–137.

Gathercole, S. 1987. Lipreading: Implications for short-term memory. In Dodd, B. & Campbell, R. (Eds.) *Hearing by Eye: The Psychology of Lipreading.* London: Lawrence Erlbaum Associates. 227–242.

Gibson, J. J. 1950. *The Perception of the Visual World.* Boston: Houghton.

Gibson, J. J. 1966. *The Senses Considered as Perceptual Systems.* Boston: Houghton.

Green, K. P. & Miller, J. L. 1985. On the role of visual rate information in phonetic perception. *Perception & Psychophysics,* 38, 269–276.

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

Jeffers, J. 1967. The process of speech reading. *Conference on Oral Education for the Deaf,* 1530–1561.

Kitson, H. O. 1915. Psychological Tests for Lipreading Ability, *Volta Review.*

Klatt, D. 1979. Speech perception: a model of acoustic-phonetic analysis and lexical access. *Journal of Phonetics,* 7, 279–302.

Kuhl, P. K. & Meltzoff, A. N. 1984. The intermodal representation of speech in infants. *Infant Behavior & Development,* 7, 361–381.

Ladefoged, P. 1985. *Unpublished lecture to the Laboratory of Phonetics,* University of Oxford.

Liberman, A. & Mattingley, I. 1985. The motor theory of speech perception revisited. *Cognition,* 21, 1–33.

Marslen-Wilson, W. & Tyler, L. 1981. Central Processes in speech understanding. *Philosophical Transactions of the Royal Society, London, B,* 295, 317–332.

McClelland, J. L. 1979. On the time relations of mental processes; an examination of processes in cascade . *Psychological Review,* 86, 287–330.

McClelland, J. L. & Elman, J. L. 1986. The TRACE model of speech perception. *Cognitive Psychology,* 18, 1–86.

McClelland, J. L. & Rumelhart, D. E. 1981. An interactive activation model of context effects in letter perception. *Psychological Review,* 88, 375–407.

McClelland, J. L. & Rumelhart, D. E. (Eds.) 1986. *Parallel Distributed Processing.* Cambridge, Mass: MIT Press.

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

Meltzoff, A. & Moore, K. M. 1982. The origins of imitation in infancy: paradigm, phenomena & theories. In L. P. Lipsett & C. K. Rovee- Collier (Eds.) *Advances in Infancy Research.* Norwood, Ablex, 263–299.

Mounoud, P. & Hauert, C. A. 1982. Development of sensorimotor organisation in young children. In G. Forman (Ed.) *Action and Thought: from sensori-motor schemes to symbol operations.* New York: Academic Press.

Reisberg, D., McLean, J. & Goldfield, A. 1987. Easy to hear but hard to understand: a lipreading advantage with intact auditory stimuli. In B. Dodd & R. Campbell (Eds.) *Hearing by Eye: The Psychology of Lipreading.* London: Lawrence Erlbaum Associates, 97–114.

Rosen, S. M., Fourcin, A. J. & Moore, B. C. J. 1979. Voice pitch as an aid to lipreading. *Nature,* 291, 174–177.

Rumelhart, D. E. & McClelland, J. L. 1982. An interactive activation model of the effect of context on perception (part 2), *Psychological Review,* 89, 60–94.

Stemberger, J. P. 1985. An interactive activation model of language production. In A. W. Ellis (Ed.) *Progress in the Psychology of Language, 2.* London: Lawrence Erlbaum Associates.

Summerfield, Q. 1979. Use of visual information for phonetic perception. *Phonetica,* 36, 314–331.

Summerfield, Q. 1981. Articulatory rate and perceptual constancy in phonetic perception. Journal of Experimental Psychology: *Human Perception & Performance,* 7, 1074–1095.

Summerfield, Q. 1987. Some preliminaries to a comprehensive account of audio-visual speech perception. In B. Dodd & R. Campbell (Eds.) *Hearing by Eye: The Psychology of Lipreading.* London: Lawrence Erlbaum Associates, 3–52.

Summerfield, Q. & McGrath, M. 1984. Detection and resolution of audio-visual incompatibility in the perception of vowels. *Quarterly Journal of Experimental Psychology,* 36A, 51–74.

Cross-Modal Effects in Repetition Priming

Cross-Modal Effects in Repetition Priming: *A Comparison of Lipread Graphic and Heard Stimuli*

Barbara Dodd, Michael Oerlemans and Ray Robinson

Speech, Hearing, and Language Research Centre
Macquarie University
North Ryde, N.S.W. 2113
Australia.

Visible Language XXII, 1
Barbara Dodd, Michael Oerlemans and Ray Robinson, pp. 58-77

A series of experiments investigated the processing of lipread information, as compared to that of heard and read stimuli, using the repetition priming paradigm. Experiment 1 showed that lipread priming facilitated the semantic categorization of lipread words to the same extent as that found for auditory prime, auditory test, and graphic prime, graphic test conditions. Experiments 2, 3 and 4 measured the effects of cross-modal priming. Lipreading primed both auditory and graphic processing, and is primed by both. While auditory priming did not speed the processing of graphic stimuli, graphic priming facilitated the semantic categorization of heard words. A tentative explanation of the findings is offered: lipreading provides incomplete information about words, and thus there is a need to access stored linguistic knowledge to 'fill in' missing features, allowing identification of the stimulus.

Lipread stimuli are of great interest because they share some common characteristics with both auditory and graphic stimuli. Like orthographic stimuli, lipread stimuli are perceived visually. However, they also differ in one important respect. Orthographic stimuli are static, and can be perceived as a gestalt at the moment of display. Lipread stimuli are dynamic, consisting of sets of transient features.

Speech perception is often assumed to be specific to the auditory modality. However, interest in the role of lipreading as a complementary source of information about spoken language has recently generated a great deal of research (Dodd and Campbell, 1984). Lipread stimuli are of interest because they share some common characteristics with both auditory and graphic stimuli. Like orthographic stimuli, lipread stimuli are perceived visually. However, they also differ in one important respect. Orthographic stimuli are static, and can be perceived as a gestalt at the moment of display. Lipread stimuli are dynamic, consisting of sets of transient features. The relationship between lip movements and heard speech is unique. The two sources of information are congruent, and provide complementary information. No other everyday activity provides such complex bimodal stimuli as face-to-face communication. Both are dynamic, and phonological. However, they are perceived through different modalities. Lipreading differs from both reading and hearing in that lipread stimuli are partial and therefore difficult to identify, whereas print and words heard in a quiet environment are easy to discriminate.

Comparison of the three types of stimuli, lipread, heard and graphic, have provided some interesting findings concerning the organization of short term memory. Until recently, it was assumed that verbal short term memory was organized along modality-specific lines. That is, heard stimuli are processed separately and differently from seen stimuli. Evidence supporting this hypothesis came from experiments showing limited cross-modal interference in short term memory tasks. For example, when a list of digits are recalled in serial order, there is enhanced accuracy of recall for the last items of the list if it has been heard, as compared to that found for read lists (Morton and Holloway, 1970). Providing further evidence, Wood (1974) found that when two successive stimuli from the two senses have to be compared, the stimuli are matched in the modality code of the second stimulus. For example, when a seen letter E was followed by a heard C, interference was more likely to occur than when a heard E was followed by a seen C. In the first case there is phonological similarity, in the second there is no visual (graphic) similarity. The results were

interpreted as an indication that stimuli perceived by different modalities are recoded into the modality-specific code of the second stimulus for comparison.

One common feature of these experiments is that the visual stimuli were presented graphically. However, subsequent experiments showed that cross-modal interference effects do exist if the visual stimuli are lipread rather than text read. There is enhanced end of list recall of lipread lists, and cross-modal suffix effects (Campbell and Dodd, 1980; Gardiner, Gathercole and Gregg, 1981). Subjects have more difficulty remembering if they have lipread or heard a word, than if they have heard or read a word (Dodd and Campbell, 1984). These results have led to the conclusion that lipread and heard speech share a degree of common processing (Summerfield, 1984; Campbell, 1987).

The investigation of more central processing of lipread information has been limited by the inherent difficulty of lipreading as a task. Normally hearing subjects can only identify 25% of a silently presented word list correctly.

However, the investigation of more central processing of lipread information has been limited by the inherent difficulty of lipreading as a task. Normally hearing subjects can only identify 25% of a silently presented word list correctly (Dodd, 1977), although when a closed set is presented (e.g. numbers, color names) subjects make few errors. The number of paradigms able to be adapted for the presentation of lipread stimuli is therefore restricted. One paradigm that has been extensively used to clarify the nature of language processing is repetition priming (e.g. Kirsner and Dunn, 1985; Monsell, 1985). Subjects are asked to perform a verbal task, e.g. lexical decision. Some of the stimuli occur twice. Reaction times show that subjects can perform the task more quickly the second time they process a word. The degree of advantage varies according to a number of factors, e.g. whether the two stimuli have been presented to the same sensory modality. This phenomenon has been used to explore aspects of the mental lexicon, such as the modality specificity of codes (Clarke and Morton, 1983), levels of representation (Meyer and Schvanevelt, 1971), and bilingualism (Christoffanini, Kirsner and Milech, 1986). Many short term memory tasks, like serial ordered recall, do not necessitate the involvement of the mental lexicon. Nonsense syllables or graphic shapes can also be effectively used as stimuli. However, word repetition priming tasks can require the recognition of meaningful

stimuli, and may therefore tap a more central level of representation. So far there has been no report of how lipread stimuli are processed in repetition priming tasks.

The experiments reported here are a preliminary exploration of the repetition priming phenomenon using lipread stimuli. Lipread stimuli may be processed similarly to heard stimuli. That is, there may be effective priming between heard and lipread stimuli but little or none between lipread and heard versus orthographic stimuli. If this were so, it would be in agreement with the pattern of results found using short term memory paradigms, and would provide evidence that lipread and heard speech share a common processing code when accessing the mental lexicon. Another possibility is that there may be absolute or relative same-modality priming. That is, visually perceived stimuli (lipread and text read) may prime each other to a greater extent than they prime, or a primed by, auditorally perceived stimuli. This hypothesis fits with previous findings (e.g. Clarke and Morton, 1983) showing significantly less cross-modal facilitation than within-modality facilitation. A third hypothesis arises from the fact that lipread stimuli are difficult to identify because they provide only partial information. The lipreader has to 'fill in' missing information, e.g. voicing. The recognition of lipread words is therefore likely to involve a greater use of stored knowledge about words than either hearing or reading. The increased "top-down" processing involved in perceiving lipread words might result in lipread stimuli being primed equally by both heard and text read words. If this were so, the pattern of priming should be unidirectional. That is, lipreading should be primed by everything, but prime nothing (except itself). These three hypotheses are not necessarily mutually exclusive, since an advantage in processing time may result from the operation of a number of factors.

The lipreader has to "fill in" missing information, e.g. voicing. The recognition of lipread words is therefore likely to involve a greater use of stored knowledge about words than either hearing or reading.

Experiment 1

While it has been demonstrated that within-modality repetition priming effects can be obtained for both auditory and visual (graphic) stimuli, there has been no investigation of whether prior lipreading experience can prime lipread performance. Before the cross-modal

priming effects between lipread, auditory and graphic stimuli can be assessed, it is necessary to establish whether lipread information is subject to within stimulus-type priming effects, and if so, the extent of the effect compared to that found for auditory and graphic stimuli.

Method

Subjects. Twenty unpaid volunteers, 6 female and 14 male, acted as subjects. They were students and staff in a university department. All have Australian-English as their native language, and were aged between 18 and 45 years of age. None had any detected hearing loss, or uncorrected visual impairment.

Procedure. Subjects participated in one experimental session, lasting less than half an hour, in a soundproof room. They sat facing a VDU, and were given a small hand-held panel on which there were two buttons. One button was labeled "A" for animal, the other button, "P" for plant. Subjects were told that they would see/hear/lipread words, and that they were to decide whether each stimulus word was an animal or a plant, and to press the appropriate button as quickly as possible. The need for accuracy was stressed.

Each subject performed three conditions: auditory priming task, auditory test; graphic (written words) priming task, graphic test; and lipread priming task, lipread test. The task was identical in both priming and test phases of the experiment, i.e. categorization of words as either animal or plant. The test phase followed the priming phase after an interval of approximately two minutes. There were 10 items in the priming list, and 20 (10 primed and 10 unprimed) in the test list. The order of presentation of the three conditions was randomized across subjects.

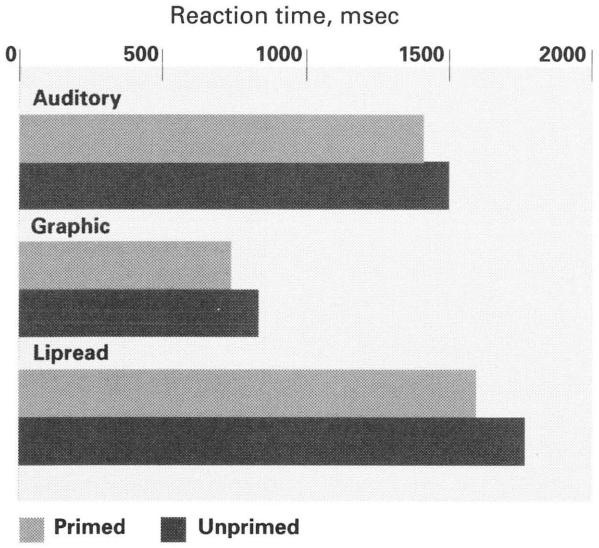
Stimuli. There were three lists of twenty words, matched for frequency of occurrence (Thorndike and Lorge, 1944) and syllable length. There were ten animal and ten plant words in each list. Words for the lipreading conditions were carefully chosen in terms of their "lipreadability". This was achieved by limiting the words to a closed set (Australian native animals and common plants). The final list of words to be lipread was determined by asking 20 students to lipread a long list of words, and

selecting only those words that were lipread accurately by at least 16 of the students in a noisy, distracting environment (the enrollment hall). Only one experimental subject showed significant lack of accuracy when lipreading stimuli, and was replaced. The other two word lists, for graphic and auditory conditions, were alternated. That is, while each subject had the same stimuli in the lipread condition, half the subjects heard one of the other lists, which other subjects read. Each list of twenty words was further divided into two lists of ten containing five animals and five plants; List A and List B. Half the subjects received List A in the priming task, and half List B.

In both priming and test phases the stimuli were presented at a rate of one every 5 seconds. In the lipreading conditions, subjects watched the VDU showing a life-size head in color, which remained on the screen for the entire stimulus presentation, but in the interstimulus interval the presenter did not look at the camera. The presentation was silent, since no audio track had been recorded. In the graphic condition (driven by a S100 microcomputer), subjects saw words in upper case (height:15mm), in bold, enclosed in a box in the center of the screen. The word was visible for 0.5 of a second, approximately the length of time taken to say each word. In the auditory condition subjects heard the stimuli through headphones.

Measurement. The dependent variable was reaction time. When subject pressed a button to categorize the stimuli as animals or plants, they stopped a timer that had been activated as each stimulus was presented. In the auditory and lipread conditions this was done by placing a tone (not heard by the subjects) on a second channel of the stimulus tape that coincided with the onset of the stimulus. In the graphic condition, presentation of the stimulus item activated the timer. At the end of each test phase the printer provided a sheet stating subject name, order of condition presentation, prime list (A or B), condition tested, and each stimulus word of the test list, with subjects' categorization choice, and reaction time.

Figure 1 Within-Modal Priming



Results

Data for each subject were analyzed to provide the mean reaction time for the ten primed and ten unprimed words in each condition (figure 1). A two-factor analysis of Variance (condition: lipread, auditory or graphic; priming: primed or unprimed) was used to analyze the data. The conditions term was significant ($F=178.1$, d.f. 2,34; $p < 0.001$). Post hoc Newman Keuls tests indicated that reaction times were shorter in the graphic condition than in the auditory condition; reaction times in the lipread condition were longest. The Priming term was also highly significant ($F = 37.5$, d.f. 1, 17; $p < 0.001$). Priming resulted in consistently faster reaction times. The interaction term was not significant, indicating that the extent of the priming effect was the same for all three conditions. Accuracy was high in all conditions. The mean number of categorization errors for the lipread condition was 3.7, while it was 0.65 for graphic condition and 1.1 for the auditory condition.

Discussion

Experiment 1 showed that lipread stimuli could be successfully used in the repetition priming paradigm. Although reaction times in the lipreading condition were

significantly longer than those for the graphic and the auditory condition, the extent of priming advantage did not differ across conditions. One contributing factor to the short reaction times for graphic stimuli is that they can be perceived as a whole at the moment of display, whereas identification of both auditory and lipread stimuli awaits the completion of the stimulus presentation. The finding that lipread stimuli took longer to process than auditory stimuli may result from lipread stimuli providing incomplete information. That is, subjects needed to "fill in" missing features, e.g. voicing, from stored representations of words. This additional processing would result in longer reaction times.

Experiment 1 provided evidence that it is possible to use lipread stimuli in cross-modal repetition priming experimental designs. Experiment 2 tested the effect of auditory priming on the semantic categorization of lipread and read words.

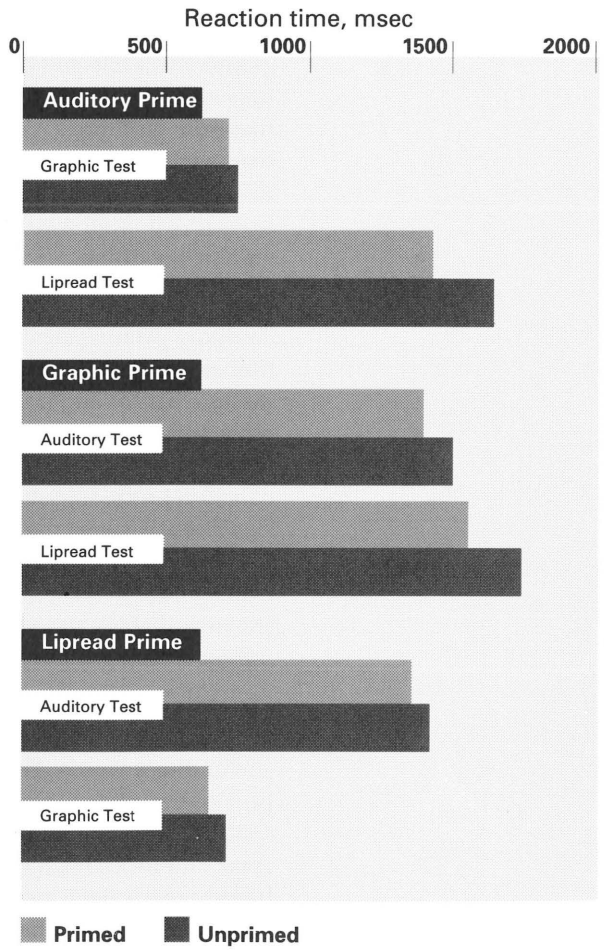
Experiment 2

Method

Subjects. The subjects were 20 unpaid volunteers recruited from the undergraduate population of Macquarie University. There were 14 females and 6 males. All of the subjects had normal hearing and (corrected) vision. They were aged between 18 and 45 years of age. None of the subjects had participated in experiment 1.

Procedures and Materials. The procedure was similar to that used in experiment 1. Subjects were presented with a list of words which had to be categorized as plant or animal in a priming phase and a test phase. The test phase consisted in one case of the list being presented graphically on the computer visual display, and in the other case, lipread from a television monitor (see experiment 1 for details). In both cases the priming task was auditory. The order of presentation of the graphic test and lipread test was alternated. The two word lists were alternated across conditions. In the priming task, half the subjects were presented with 10 of the words (5 animal and 5 plant), whereas the other subjects were primed with the remaining ten words.

Figure 2 Cross-Modal Priming



Results

Data for each subject were analyzed to provide a mean reaction time for the ten primed and ten unprimed words in each condition (figure 2). A two-factor Analysis of Variance was conducted on these scores (lipread versus graphic test; primed versus unprimed). The results showed that the conditions term was significant ($F = 263.2$, d.f. 1, 19; $p < 0.001$). Post hoc tests showed that graphic reaction times were significantly shorter than lipread reaction times. The effect of priming was also significant ($F = 5.77$, d.f. 1, 19; $p = 0.025$). Priming again

resulted in faster reaction times. However, the interaction term was significant ($F = 5.317$, d.f. 1, 19; $p = 0.03$), indicating that the condition of presentation (lipread or graphic) was affected differentially by the auditory priming encounter. Post hoc Newman Keuls testing showed that prior auditory experience primed the perception of words when they were lipread but not when they were read. The mean number of categorization errors for the lipread condition was 3.6, while it was 0.65 for the graphic condition.

Discussion

Auditory priming significantly speeded the categorization of subsequently lipread words (mean increase of 216 msec). This finding is consistent with the results from other paradigms, e.g. serial ordered recall, showing that lipread and auditory stimuli appear to share a common processing stage. Auditory priming did not significantly speed the categorization of graphically presented words (mean increase 10 msecs). This finding is in agreement with those of Monsell (1985) and others, showing that the processing of graphic stimuli is not influenced by prior auditory experience.

The third experiment investigated the effect of a graphic priming task on the semantic categorization of lipread and auditory words. The findings of experiment 2 and previous reviews of research (e.g. Allport and Funnell, 1981) suggest that graphic priming should not affect the speed of processing heard words. No previously published reports have measured the effect of graphic priming on the reaction times for lipread word semantic categorization.

Experiment 3

Method

Subjects. The subjects were 20 undergraduates enrolled at Macquarie University: 13 females and 7 males. No subject took part in more than one of the experiments. All subjects were aged between 18 and 45 and had normal hearing, and (corrected) vision.

Procedures and materials. The procedure was similar to the other experiments although in this case, graphic stimuli were used in the priming encounter and the test

encounter comprised a lipread and an auditory condition. The order of presentation of the test and priming lists was randomized across subjects.

Results

The data were analyzed with a two factor Analysis of Variance. The conditions term was significant ($F = 6.1$, d.f. 1, 19; $p < 0.025$), indicating that reaction times in the lipread task were significantly longer than those in the auditory task (see figure 2). The effect of priming was also significant ($F = 25.6$, d.f. 1, 19; $p < 0.001$). The interaction term was not significant ($F = 2.4$, d.f. 1, 19; $p = 0.138$), indicating that the extent of the priming effect did not differ for the auditory and lipread tests. The mean number of categorization errors for the lipread task was 3.6, while it was 1.3 for the auditory task.

Discussion

The results indicate that a graphically presented priming task increased the speed of semantic categorization of both auditory (mean increase 69 msec) and lipread (mean increase 179 msec) stimuli. The finding that reading words could prime their processing when they were heard is at odds with some previous findings (e.g. Clarke and Morton, 1983) and reviews of the literature (e.g. Allport and Funnell, 1981). However, it is not the first reported case of cross-modal priming. Monsell (1985) reported that graphic priming speeded the processing of auditory stimuli in a lexical decision task. He commented that although cross-modal effects may be numerically small compared to within-modality effects, they should not be dismissed as null effects. Experiments 2 and 3 indicated that the priming effect was asymmetrical: while a graphic stimulus primed auditory categorization, the reverse was not true. This pattern replicated Monsell's (1985) findings. Experiment 4 investigated whether a lipread input, which showed an increased reaction time when primed by auditory and graphic stimuli, could in turn prime semantic categorization of heard and read words.

Experiment 4

Method

Subjects. 20 undergraduates, 12 female and 8 male, volunteered for the experiment. They were all aged between 18 and 45 and possessed normal hearing and (corrected) vision.

Procedures and materials. The procedure followed that of the other experiments. In this experiment the priming stimuli were lipread words and the test stimuli were auditorally and graphically presented words.

Results

An Analysis of Variance was performed on the mean reaction time scores of subjects. The conditions term was significant ($F = 188.3$, d.f. 1, 19; $p < 0.001$) indicating that reaction times were shorter for the graphic test than for the auditory test (see figure 2). The effect of priming was also significant ($F = 5.981$, d.f.1, 19; $p < 0.025$). The interaction term was not significant ($F < 1$). That is, lipread priming equally increased reaction times in an auditory categorization task and a graphic categorization task. The mean number of categorization errors for the graphic task was 0.75, while it was 1.25 for the auditory test.

Discussion

Prior lipread experience of words facilitated their semantic categorization when they were heard and read. While the interactive relationship between lipread and heard speech was predicted from the results of other paradigms, the priming of read words by lipreading was surprising. Since lipreading is a more difficult, and less familiar, task than hearing or reading words, it is likely that any prior information about what a lipread stimulus might be would be used, irrespective of its modality of input. It is more difficult to explain why lipread experience should enhance the processing of read words. A comparison of the four experiments might clarify the pattern of findings.

While the interactive relationship between lipread and heard speech was predicted from the results of other paradigms, the priming of read words by lipreading was surprising.

Comparison of experiments 1, 2, 3 and 4

Table 1 sets out the mean difference scores (unprimed minus primed), and the percent difference expressed as a proportion of total reaction time, for each condition. A two-factor (prime mode, and test mode) Analysis of

Variance using percent difference scores, where each condition was treated as an independent group, was used to compare the four experiments. The priming term was not significant ($F < 1$), i.e. the type of priming (audi-tory, graphic and lipread) did not influence the extent of the priming effect. Obviously, the large analysis, treating the scores for each condition independently, swamped the finding that auditory priming does not affect the semantic categorization of read words (see experiment 2). This is hardly surprising, as all other conditions show a significant priming effect. The test term was significant ($F = 3.7$, d.f. 2, 65; $p < 0.025$), i.e. mode of test stimuli presentation affected the extent of the priming effect. Inspection of Table 1 shows that the priming effect was strongest for lipread test stimuli. The interaction term did not reach significance.

Table 1
Unprimed/Primed mean difference scores (msec) according to test and priming mode, and percentage increase due to priming

	Test Mode					
Increase:	Δ	%	Δ	%	Δ	%
Graphic Prime	64	9.4	69	4.3	179	10.5
Auditory Prime	10	1.4	90	6.1	216	14.3
Lipread Prime	55	6.7	47	1.9	164	7.7

General Discussion

The experiments reported investigated the processing of lipread words in comparison to that of heard and read words using the repetition priming paradigm. The results were somewhat unexpected, and are difficult to explain. Lipreading was primed by everything, and primed everything. The only case when cross-modal priming did not arise was when the priming encounter was auditory and the test encounter was graphic. This finding is asymmetrical in that graphic presentations primed auditory tests. The findings cannot be explained by any *one* of the three hypotheses set out in the introduction.

A comparison of lipread and heard stimuli shows that each primed the other. This fits with the hypothesis that lipread and heard speech share a processing code at the level of lexical access. However, this hypothesis cannot account for all the priming effects found. Lipreading also primed reading, and reading primed auditory perception. The results suggest that at the level of lexical access, verbal information is coded differently from its form in short term memory, where it has been shown that lipread and heard speech share a code that excludes orthographic information (Campbell, Dodd and Brasher, 1983).

The second hypothesis, that same modality of perception would underpin the priming effects found, must also be rejected, since lipread and heard stimuli primed each other and reading primed auditory perception. The third hypothesis, that lipreading will be primed by everything because it is partial information, is rejected because although this was true, lipreading also primed everything.

Lipread stimuli behave differently from both heard and read stimuli. Lipreading primed reading, but hearing did not prime reading. Lipreading was primed by hearing, but reading was not primed by hearing. The pattern of findings cannot be simply explained by saying that lipread information is processed like heard information because they share a code; or like read information because they share a common modality of perception. Further, the fact that lipread stimuli are difficult to perceive has no general explanatory power because lipreading primed both heard and read stimuli.

The pattern of findings cannot be simply explained by saying that lipread information is processed like heard information because they share a code; or like read information because they share a common modality of perception.

Perhaps more than one factor is operating to produce the pattern of results found. Although the comparison across experiments must be considered with caution, an inspection of table 1 shows that the two largest percent priming advantages were both for a lipread test encounter, primed by auditory and graphic stimuli. Because lipreading is a more difficult task than hearing and reading, the lipreader has to "fill in" missing information. That is, lipreading involves more top-down processing. The fact that reaction times were longer for lipread stimuli than for heard and read stimuli may be explained by subjects' need to access stored information that aids identification of a perceptually unclear stimulus. Knowing

which particular words were to be lipread dramatically improved reaction times. A large fraction of the priming advantage must be attributed to that factor, and consequently obscures the findings. If the effect of stimulus "incompleteness" could be partialled out, the results might be different, and easier to interpret.

Another factor that may have influenced the results was that repetition priming would be reduced in conditions where the prime was lipread because subjects did not recognize all of the stimuli during the priming phase. Despite these difficulties, lipread stimuli are still useable in the repetition priming paradigm. The within-modal priming advantage (experiment 1) was consistent across conditions. Even though lipreading takes longer, priming produced an advantage no greater than that found for hearing or reading. This result suggests that the priming effects found for lipread stimuli cannot be solely attributed to uncontrolled variables like task difficulty.

Further evidence that lipread stimuli are effective comes from the finding that lipreading primes both auditory and graphic processing (experiment 4). Since lipread information is partial, and the stimuli in both auditory and graphic test conditions are easily discriminable, the results seems counter-intuitive. One possible explanation is that during a lipread priming task, subjects access stored information from a variety of sources, including graphic and phonological representations of words, in order to identify the partial stimuli. This "deep processing" would be likely to enhance recognition memory, and speed the processing of identified words when they appeared in the auditory and graphic test phases.

The asymmetrical priming effect found for auditory and graphic stimuli is not new (see Monsell, 1985). Reading primed hearing, but hearing did not prime reading. This finding may reflect different reading strategy use in the two auditory/graphic cross-modal priming conditions. Written words can be processed by the brain in two ways. Studies of patients with brain injury have shown a double dissociation. Some patients can only read words using the grapheme-phoneme conversion route, others can only read words that they recognize as an orthographic gestalt (Coltheart, 1980). Subjects with normal

brain function could use either route according to task demands. In a task where speed of response is the known measure, and the decision is one of semantic categorization of read words, recognizing gestalts would be more efficient, and the auditory priming could be ignored. When words are read as a priming task for a list that will be heard, subjects may choose to code the phonological shape of the stimuli because it would provide an advantage during the auditory test phase. This speculative hypothesis is not necessarily weakened by the finding that prior lipreading experience of words speeded response in a graphic test. The "deep processing" required for lipread stimuli during the priming phase may have accessed graphic representations that would contribute to, or account for, the priming advantage for read words in the test phase.

Written words can be processed by the brain in two ways. Studies of patients with brain injury have shown a double dissociation. Some patients can only read words using the grapheme-phoneme conversion route, others can only read words that they recognize as an orthographic gestalt. Subjects with normal brain function could use either route according to task demands.

The experiments reported are an initial exploration of the repetition priming effect using lipread stimuli. The processing of lipread stimuli differed from that of both read and heard stimuli. This is not necessarily surprising. Lipread stimuli share some characteristics of both heard and read stimuli. They also differ from heard and read stimuli in important ways. Logically, the pattern of cross-modal priming effects gained when using lipread stimuli should show similarities and differences with the two other means of verbal communication. Current experiments are investigating two hypotheses: that normal and degraded auditory and graphic stimuli will show different pattern of cross-modal priming, and that forcing subjects to code graphic stimuli phonologically will give rise to symmetrical cross-modal priming effects between hearing and reading.

Acknowledgements

We thank the N.H. & M.R.C. for financial assistance. We are grateful to Kim Kirsner for critically reviewing the manuscript.

Barbara Dodd, Ph.D.

About the authors

Barbara Dodd is a Senior Lecturer in the School of English and Linguistics at Macquarie University in Australia. She is a qualified speech pathologist trained at Sydney University; she completed her doctorate at London University in 1974 on the acquisition of phonological skills in children. She teaches at Macquarie in the areas of psycholinguistics and language disorders. Her research interests include the characterization and treatment of spoken phonological disorders and phonological dyslexia, developing audio-visual methods for teaching lipreading, and the investigation of the behavior of lipread speech in short term memory.

Michael Oerlemans is a research assistant in the Speech, Hearing and Language Research Centre at Macquarie University and holds a Bachelor of Arts (Honors) degree in Linguistics. His interests center on the description of phonological dyslexia and the investigation of short-term memory coding of stimuli and the implications for this to models of memory.

Ray Robinson is a professional Electronic Engineer and is in charge of the Speech, Hearing and Language Research Centre at Macquarie University. He holds a Bachelor of Engineering degree from the New South Wales Institute of Technology and is presently completing a Master of Arts (Honors) degree in Applied Linguistics and Computing Science.

References

Allport, D. A. and Funnell, E. 1981. Components of the mental lexicon. *Philosophical Transactions of the Royal Society, London.* B 295, 397–410.

Campbell, R., 1987. Lipreading and immediate memory processes. In B. Dodd and R. Campbell, (Eds.) *Hearing by Eye.* London: Erlbaum.

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

Campbell, R. Dodd, B. and Brasher, J. 1983. The sources of visual recency: movement and language in serial recall. *Quarterly Journal of Experimental Psychology,* 35A, 571–587.

Clarke, R. and Morton, J. 1983. Cross-modality facilitation in tachistoscopic word recognition. *Quarterly Journal of Experimental Psychology,* 35A, 79–96.

Coltheart, M. 1980. Reading, phonological recoding and deep dyslexia. In M. Coltheart, K. Patterson, and J. C. Marshall (Eds.) *Deep Dyslexia.* London: Routledge and Kegan Paul.

Christoffanini, P., Kirsner, K. and Milech, D. 1986. Bilingual lexical representation: the status of Spanish-English cognates. *Quarterly Journal of Experimental Psychology,* 38A, 367–393.

Dodd, B. 1977. The role of vision in the perception of speech. *Perception,* 6, 31–40.

Dodd, B. and Campbell, R. 1984. Non-modality specific speech coding: the processing of lipread information. *Australian Journal of Psychology,* 36, 171–179.

Gardiner, J. M. Gathercole, S. and Gregg, V. H. 1981. Lipreading and auditory memory. *Paper to Annual Meeting of the Psychonomic Society,* Philadelphia.

Kirsner, K. and Dunn, J. C. 1985. The perceptual record: a common factor in repetition priming and attribute retention. In M. I. Posner and O. S. M. Marin (Eds.), *Mechanisms of Attention: Attention and Performance XI.* Hillsdale, New Jersey: Erlbaum.

Meyer, D. E. and Schvanevelt, R. W. 1971. Facilitation in recognizing pairs of words: evidence of a dependence between retrieval operations. *Journal of Experimental Psychology,* 90, 227–234.

Monsell, S. 1985. Repetition and the lexicon. In A. W. Ellis (Ed.) *Progress in the Psychology of Language, vol. 2.* London: Erlbaum.

Morton, J. and Holloway, C. M. 1970. Absence of a cross modal 'suffix effect' in short term memory. *Quarterly Journal of Experimental Psychology,* 22, 167–176.

Summerfield, A. Q. 1987. Some preliminaries to a comprehensive account of audio-visual speech perception. In B. Dodd and R. Campbell, (Ed.s) *Hearing By Eye.* London: Erlbaum.

Thorndike, E. L. and Lorge, I. 1944. *The Teachers' Handbook of 30,000 Words.* New York: Columbia University.

Wood, L. E. 1974. Visual and auditory coding in a memory matching task. *Journal of Experimental Psychology,* 102, 106–113.

Perception of Facial Movements in Early Infancy

Perception of Facial Movements in Early Infancy: *Some Reflections in Relation to Speech Perception*

Annie Vinter

Faculty of Psychology and Educational Sciences, University of Geneva, 24 rue du Général Dufour, 1211 Geneva 4, Switzerland; and Scientific Institute Stella Maris, Via dei Giacinti, 56018 Calambrone (Pisa), Italy

Visible Language XXII, 1
Annie Vinter, pp. 78–111

Some aspects of the literature dedicated to the study of perception of facial features and movements by infants are examined. More particularly, we try to analyze the kind of visual information infants can process at different ages, and how this may be linked to their developing speech perception. Empirical data related to imitation of facial movements, to pre-speech activity, to lip-reading ability and auditory-visual integration are reviewed. These data show that the ability of young infants to encode face features and process facial information undergoes a complex development in the first year of life. In the final part of this paper, we discuss briefly the relationships between face perception processes and visual speech perception within a developmental and cognitive framework. A central concern in this discussion is related to the "segmentation" problem, i.e. to the nature of the unit of perception used when speech is processed.

Recently people working on language have started to focus their attention on the fact that to produce speech one has to move the lips, and that some information about what is actually said may be derived from these movements. Of course the auditory/acoustic characteristics of language have attracted the major interest of (developmental) linguists and psycholinguists, for blind infants tend to develop normal speech, while deaf infants experience major difficulties in acquiring language. But, as a matter of fact, Mills (1987) reports that blind children do not show normal phonemic production development, but in their own speech, tend to confuse sounds that "look alike" (with regard to lipread information), for a more protracted period than hearing children. They also babble less than sighted infants, which may be indicative of the role that lipread information plays in speech production. Moreover, as mentioned by Dodd (1983), it is a current practice among teachers of deaf children to start to emphasize lipreading when infants are of a very early age, recognizing in this way the role of vision in speech perception. Such a practice seems to take for granted that young infants are able to process lipread information.

In this paper, we will examine some aspects of the literature dedicated to the study of perception of facial features and movements by infants. We will try to analyze the kind of visual information they can process at different ages and that may be related to their developing language perception. This paper will be closely focused on data indicative of the pre-linguistic infant's ability to discriminate visual information in faces. Moreover, when available, developmental data will be presented. The final part of the paper will speculate about the relationships between facial movements and language perception.

Different behaviors are meaningful with regard to the infant's ability to perceive facial movements, associated or not with speech:

Imitation of Facial Gestures: the repertoire of facial movements that have been used in imitation studies includes the tongue protrusion, lip protrusion, mouth opening-closing, eye blink, eye opening-closing and cheek movements.

Pre-Speech Movements: these refer to the infant's facial and manual movements when she is confronted with a talking partner.

Lipreading Behavior: these are some studies which have directly investigated the infant's ability to integrate visual and auditory information from lip movements.

In addition, it is worth analyzing the studies that show the infant's more general ability to coordinate vision and audition in relation to the face, as for instance to conceive that face and voice share a common spatial location.

Gaze Co-Orientation or Reciprocal Gaze Between Infant and Adult: this seems mainly related to pragmatic aspects of language, which is of minor interest; we will report this ability to interpret adult gaze only very briefly.

But before considering these different behaviors, it may be interesting to briefly mention some data on the infant's visual preferences with respect to faces as well as some indications on the developing psychophysical sensitivities of the infant's visual skills.

The Infant's Visual Preferences and Visual System Maturation

A summary of the main research results on early visual preferences follows. We will assume that these preferences indicate which stimuli are more efficiently processed by young infants (see Fantz, 1966; Banks, 1985). From a very early age, infants seem to prefer faces to other stimuli (see Vinter et al., 1986 for a review). Newborns track a moving schematic face in preference to other similar stimuli (Goren et al., 1975), and 2-week-old infants prefer to look at a still human face rather than at other stimuli (Fantz, 1966).

From a very early age, infants seem to prefer faces to other stimuli.

The stimuli which are attended to most by newborns are large, of high contrast, and with sharp contours (Fantz & Fagan, 1975; Fantz et al., 1975; Karmel, 1969; Salapatek, 1975). The hairline and the eyes are the most distinctive or high contrast features of the face, and seem to be the first features to be discriminated. In fact, habituated with a schematic face, 4-month-olds do not notice modifications in the mouth-nose configuration but do notice those of hairline-eyes, suggesting that, at this age, the lower

part of the face is not well discriminated probably because of lower contrast and softer contour. Changes in nose and mouth are noticed later, at 5 months (Caron et al., 1973).

Given patterns with a fixed number of elements, infants prefer larger to smaller elements, and given patterns of fixed element size, they prefer the one with more elements in it. Later the number of elements in a pattern, in comparison with the size, becomes a more salient feature. At birth, a flat representation of an object is preferred to a three dimensional object and to photographs. Preferences for photographs and then for the real three dimensional object appear later. Newborns prefer schematic faces to photographs of faces. The preferences are reversed at 5 months. Finally a preference in curvature of external contour also exists at birth. This preference seems to decrease to a minimum at 4 weeks and increases again by 7 weeks. Specific curvatures do in fact characterize the external contours of the face, such as either the border of the face, which may be the first attended to, or in a less pronounced way, the eyes or the mouth.

In short, these data give support to the view that, from birth, infants are attracted by the human face, are senstive to different optical-perceptual parameters at different times, and do spend time observing faces. They also suggest that, in a static face, the mouth is discriminated relatively late in development, not before 4 months.

In short, these data give support to the view that, from birth, infants are attracted by the human face, are sensitive to different optical-perceptual parameters at different times, and do spend time observing faces. They also suggest that, in a static face, the mouth is discriminated relatively late in development, not before 4 months.

Several aspects of developing visual capacities may be important with regard to these visual preferences. Development of the oculomotor systems might explain the appearance of the ability to detect a small pattern, to explore many details successively, and to hold foveal fixation with both eyes. Such abilities are important for the exploration of the fine internal features of the face. Development of visual acuity is also relevant for an understanding of face perception development, since pattern elements which are smaller than the resolution limit of the visual system cannot be detected. At birth the resolution power of the infant's visual system is low (around 2c/deg) and increases steadily for at least the first 6 months of life (Banks & Salapatek, 1978; Marq et al., 1976). Fine facial details are therefore unlikely to be

discriminated in the very first months of life. Finally the development of contrast sensitivity appears to be another important factor of the developing ability to recognize pattern. The CSF (contrast sensitivity function) plots the minimal contrast necessary to just detect a sine wave grating with the grating's spatial frequency. Sensitivity to middle and high spatial frequencies undergoes a marked development during the first months of life (Banks & Salapatek, 1981; Banks, 1982). Atkinson et al., (1977) showed a large increase in contrast sensitivity from 1 to 2 months mostly at high spatial frequencies. By contrast, no noticeable change occurs between 2 and 3 months (Banks & Salapatek, 1978).

According to Bank's visual preference model (Banks, 1982), the linear systems model, the most preferred pattern is that one which best fits the infant's visual "window". This model uses the CSF as the description of this window, and can be summarized by a very simple rule: infants aged less than 3 months look at the pattern whose filtered output is greatest. After 3 months, other dimensions (perceptive-cognitive, attention-memory) also become important to account for visual preferences. In fact, as pointed out by Banks (1985), the linear-systems model is completely insensitive to the meaningfulness of a visual pattern, whereas it is very likely that as the complexity of visual perception grows with age, what a visual pattern "tells" to the infant may become a major determinant of his visual perception or preference.

A review of some specific behaviors that demonstrate that infants are sensitive to facial information at a very early age follows.

Perception of Facial Movements in Infancy

Imitation of facial movements

Following Piaget (1946), it has long been assumed that very young infants are poor at imitating gestures, either manual or facial. Yet, the first experimental study of early imitation, was carried out as early as 1928 by Guernsey. She observed 214 infants aged from 2 to 21 months, and analyzed their imitations of different models. Of interest are what she called "expressive mimic movements", which include the mouth opening–closing,a large open-

ing of the mouth, the tongue protrusion, and which are contrasted with models such as vocal models or movements performed with objects, toys and so on.
She concluded that (ibid, p.143):

"Die einzigen wirksamen Nachahmungsreize unter 4 1/2 Monaten sind mimische Ausdrücksbewegungen. Die Reaktionen sind vorherrshend reflexartig and bleiben es auch während des ganzen Lebens".

"The only items which are effectively imitated under 4 1/2 months of age are the expressive mimic movements. The reactions are mainly reflex, and remain so throughout life" *(our translation).*

The mimic movements which are reproduced from the age of 2 months onwards are essentially the large opening of the mouth, the mouth opening-closing, the tongue protrusion, and a lateral rotation of the head. Moreover she observed a progressive disappearance of these imitative responses between 4 and 6 months, whereas imitation of other kinds of movements progressively develops after 6-7 months of age. Even if she considered these early imitations as reflex, it is astonishing to realize that she identified as first imitative behaviors imitative responses to the "Aüsdruckbewe-gungen", i.e. facial gestures.

Finally, several anecdotal reports have accumulated during the last three decades of observation of early imitation occurrences (Brazelton &Young, 1964; Gardner & Gardner, 1970; Zazzo, 1957).

Current knowledge of early imitation has not progressed very much since 1928! Maratos' experiment (1973, 1982) marked the beginning of a new line of research in infancy but basically confirmed Guernsey's results. At 1 month, infants imitate a tongue protrusion movement, an opening–closing of the mouth, and a lateral head movement. Furthermore, imitation of the tongue protrusion movement disappears between 2 and 3 months, and that of the mouth opening-closing at around 3 months. But now, our American cousins, with their high technology and sensitive methods, have considerably complicated the situation! Some authors have argued in favor of the existence of early imitation ability and have extended the

repertoire of facial movements that very young infants are able to reproduce: lip protrusion at 2-3 weeks (Meltzoff & Moore, 1977), facial expressions of sadness, surprise and happiness at less than 1 week (Field, et al., 1982), eye blink and cheek movement at 2 months (Fontaine, 1982), opening-closing of the eye within three-quarters of an hour of birth (Kugiumutzakis, 1985a, 1985b). On the other hand, others have denied that young infants can imitate facial movements, after failing to replicate Meltzoff and Moore's results (Hayes &Watson, 1981; McKenzie & Over, 1983). Lewis and Wolan-Sullivan (1985) did not observe any facial imitation either at 2 weeks, 3 months or 6 months. Finally, whereas Abravanel and Sigafoos (1984) described a very restricted capacity to imitate the tongue protrusion movement, the specificity of this task may be put in doubt by Jacobson's (1979) finding that tongue protrusion is elicited no more frequently by a person's protruding tongue than by a pen moving toward and away from the infant's mouth.

Despite some inconsistencies, then, (see Vinter, 1985, for further discussion), these studies attest quite well to the capacity of newly born infants to reproduce at least two facial movements: that of tongue protrusion and that of mouth opening-closing. It is worth mentioning that one cannot conceive of imitation without postulating the existence of some selective perceptuo-motor linkage, which integrates different sensory modalities and permits an identification of some facial features. Further research is needed to confirm the neonate's ability either to imitate facial expressions or to selectively imitate facial gestures other than tongue protrusion or mouth opening and closing.

Up to now, we have little understanding of which features or combination of features infants respond to when they imitate. Jacobson's study suggests that properties of shape and movement are fundamental to the elicitation of imitative responses, which has been confirmed by Vinter (1986) with regard to the role of movement in neonatal imitation. Infants exposed to kinetic (dynamic) facial and manual actions emitted higher rates of the modeled acts in the interval during which it was modeled than in any other condition. In contrast, infants exposed to the static version of the same act failed to show evidence of selec-

Further research is needed to confirm the neonate's ability either to imitate facial expressions or to selectively imitate facial gestures other than tongue protrusion or mouth opening and closing.

tive reproduction of the modeled behavior; they did, however, spend relatively more time visually fixating both the facial and the manual positions. This suggests that, at this age, a static face is more likely to elicit visual exploration than imitation, and that the role of movement may be a fundamental criterion that differentiates early imitation from late imitation.

It would also be interesting to know to what extent the "faceness" of the model plays a role in this phenomenon, and whether, for instance, the presence of features such as the eyes or the nose are also important in eliciting imitation of mouth movements.

The fact that newborn infants can reproduce mouth movements such as tongue protrusion and mouth opening-closing may appear to contradict data on how they process the features of the face. It has in fact been claimed that infants do not initially discriminate internal fea-tures of the face. Studies of infant eye movements show that 1-month-olds examine only the external contours of a real face (hairline, chin, ear) whereas 2-month-olds also scan the internal features, particularly the eye region (Maurer & Salapatek, 1976; Hainline, 1978). But this "externality effect", which is also observed with a compound figure, appears to be influenced by at least three parameters: the size of the internal figure, its salience, and the presence of relative motion (Banks & Salapatek, 1983; Milewski, 1976). This last parameter is of interest to us here. Bushnell (1979) showed that infants at 1 month discriminate changes of the internal figure when it flickers or is moved within the external figure, but not when both move together or when the component is static. In the imitation studies, an "internal feature" of the face (the mouth) is in fact in motion. Thus there is sufficient evidence to support the view that infants in the very first month of life are able to process visual information related to internal features of the face, in particular to mouth movement (if not to a still mouth).

Interestingly, the developmental studies of early facial imitation all report a gradual disappearance of this ability during the first months of life (Dunkeld, 1978; Maratos, 1973; Fontaine, 1982; Vinter, 1985). Imitation of the tongue protrusion movement is no longer observed at 3

months, that of mouth opening-closing disappears at around 3-4 months (Vinter, 1985). Also Jacobson (1979) who did not agree on the existence of a selective imitative capacity at birth since "imitative" movements are equally elicited by inanimate models sharing some characteristics with human models, observed a "disappearance" of matching responses between the inanimate model's movements and the infant's movements after 2 months. Simultaneously, some authors described the progressive appearance of a new imitative ability, essentially related to vocal imitation, in the period between 2 and 6-8 months (Papousek & Papousek, 1979, 1982; Kugiumutzakis, 1985b).

In relation to this later period of development, Razran (1971) quoted a very interesting Russian study of imitation carried out by Lyakh (1968a and b). In this experiment, infants aged from 2 to 8 months are confronted with an adult performing two mouth movements: one corresponds to the articulatory movement typical of the vowel "a" (similar to a mouth opening-closing), and the other of the vowel "o" (close to a lip protrusion movement), but both were produced without sound. The author reported that imitation of these movements is more frequent in the 2-to 4-months age group than in the 4-to 8-months age group, but does nevertheless exist in the latter group.

The loss of selective imitative responses corresponds with the appearance of new selective reactions to the modeled acts, either facial or manual. As far as facial imitation is concerned, several authors report that the infant tries to reach for the experimenter's tongue (Fontaine, 1983; Kugiumutzakis, 1985b).

Social reactions also are obtained. Vinter (1985) has shown that 3-month-olds smile and vocalize in response to the facial model (tongue protrusion), but look at their own hand in response to the manual model. She has called these reactions "analogical imitations" in the sense that, to some extent, the infant takes into consideration the body part involved in the modeled act. These reactions also make clear that during the first months of life, the infant seems to be involved in a process of (re)discovery or (re)identification of his body parts. This

process is necessary for an intentional imitative ability to occur.

The way in which facial imitation develops later in childhood has been well analyzed by Piaget (1946), and confirmed by Uzgiris & Hunt (1975). Successful imitation of mouth movements (such as mouth opening-closing or tongue protrusion) seems to appear again between 9 and 14 months. In this period, there seems to be great interindividual variability in the order of acquisition of these facial imitative responses.

Pre-speech movements

With regard to language as both auditory and visual input, some astonishing kinds of imitative behavior have been described. Condon & Sander (1974) showed that a precise temporal and rhythmical synchrony can be established between the infant's movements (arm-hand) and adult speech when talking to the infant. They showed that arm displacement by the infant coincided with syllable or word pronouncement by the adult. Furthermore, infants tend to move when the speech sounds are modified and keep their posture constant all the time that a speech sound is produced. According to the authors, this ability to synchronize one's own movements with heard speech units may account for the relationships between articulation and audition, in particular the fact that speech stream is perceived in discrete units. At the least, it suggests a general sensitivity to spoken syllable structure in the prelingual child. Infants aged 2-4 months also seem to mimic lip and tongue movements associated with speech articulations when stimulated to communicate with an adult partner. These movements have been called "pre-speech" movements (Trevarthen, 1974, 1984), and are often produced without vocalisations. If the imitative nature of such movements could be established, this behavior might show that young infants are sensitive to visual information derived from adult mouth movements when talking. Sensitivity of infants as young as 2 months to the mother's facial expressions as a whole (not just the mouth) has been noted by several authors (see Schaffer, 1977; Trevarthen, 1980). The more often the mother smiles, vocalizes, looks at her infant, the more frequent are infant's smiles, vocalizations, positive

With regard to language as both auditory and visual input, some astonishing kinds of imitative behavior have been described.

We lack the experimental studies that would give us a real understanding of these phenomena. We need to know to what aspects of adult speech these prespeech movements are related, for example, whether they can be elicited by hearing or vision.

orientations to the mother. In complementary fashion, a blank or silent mother's face elicits negative reactions or responses of avoidance from her infant.

We lack the experimental studies that would give us a real understanding of these phenomena. We need to know to what aspects of adult speech these pre-speech movements are related, for example, whether they can be elicited by hearing or vision alone. Similarly, it may be important to establish whether responsiveness of the infant to the mother's facial expressions is based upon a wholistic perception in which kinetic information is essential (the infant would react by a moving face to a moving face) or is more dependent on static face features (the infant may smile at a wide mouth for instance). In addition, indications about the development of this ability to mimic speech activity are required; in particular the age at which it appears and whether or not its development follows a course similar to that of early imitation.

Lipreading ability and auditory-visual integration

A large number of experiments have been dedicated to the study of how infants can integrate auditory with visual information. Some of them are specifically related to speech perception (experiments on lip reading for instance– see Campbell, 1986, for a review), others are aimed at a more general understanding of how infants conceive of the relationships between face and voice (see Butterworth, 1980, for a review).

Aronson & Rosenbloom (1971) have shown that l-month-old infants are distressed when confronted with a voice coming from a spatial location different than that of the face, as if they expected face and voice to share a common location. Other authors described other kinds of behavior in this situation of spatial discordance. According to Castillo & Butterworth (1981), neonates systematically orient to the face and thus seem to "resolve" the spatial conflict in favor of vision. Vinter et al. (1984) showed that neonates orient more frequently face after having turned their head toward the voice than toward the voice after the face has been the first preferred stimulus. It might mean that upon an auditory stimulus (voice) a visual stimulation (face) is expected, but not

vice-versa. These studies demonstrate that infants younger than 1 month are sensitive to a spatial discordance between vision and audition. But so far, data are lacking in order to know to what extent this sensitivity may be specific to face perception, in contrast to perception of any audible object, as far as very young infants are concerned. At around 3-4 months, it seems established that a similar sensitivity can be observed both with human faces and inanimate displays (see Spelke, 1976, 1985).

An intriguing developmental course has been revealed both with regard to the ability to integrate vision and audition from face perception and to orient toward a voice (see Muir & Clifton, 1985, for further discussion). Muir et al. (1979) found that at around 2 months, a response of orientation to sound is very difficult to elicit, whereas it is much easier to obtain either at birth or at 3-4 months. Similarly, Vinter et al. (1984) described a U-shaped development in infant's response to spatial discordance between face and voice. In particular, two-month-olds did not seem to notice when face and voice were displaced, since they rarely, if ever, turn their head in the direction of the voice. This change in responsivity at 2 months recalls a similar failure to react which was noted in the imitation studies (see above).

While attention to a single face or voice is one way to study sensitivity to their concordance, a preferential looking paradigm can provide more detailed information. In such studies, the experimental procedure consists of neighbouring presentation of two films (or slides) with a central soundtrack that correctly matchs one of the films. Relative durations of the infant's looking at the two films are measured. Confronted with the mother's face and a strange female face, 8-month-olds look at the mother when the central sound corresponds to the mother's voice, at the stranger when the sound emits her voice, whereas 5-month-olds do not look preferentially to one of the faces according to the voice heard (Cohen, 1974). This study suggests that it is rather late in development that infants are sensitive to shared identity of face and voice. But other experiments seem to demonstrate that such an ability exists earlier. In Spelke & Owsley's study

(1979), the mother's face was contrasted to the father's face whereas either the mother's voice or the father's voice was centrally emitted. Three-month-olds are able to associate the face with the voice correctly in this situation. But it is true that mother and father are auditorally and visually more different than mother and female stranger. This difference in the degree of similarity of the two streams of information may account for the difference observed between Cohen's and Spelke & Owsley's studies. In sum, Spelke's research (1976; see Spelke, 1985, for a review) suggests that 3 or 4-month-olds are able to coordinate auditory and visual information correctly from very different displays, not uniquely from faces.

Infants can be shown to be aware of an even more precise concordance between face movements and voice. Dodd (1979) demonstrated that 10 to 16-week-old infants can detect that a voice is in or out-of-synchrony with respect to the mouth's movements. Spelke & Cortel you (1981) confirmed that 3-month-olds look more at the face whose mouth movements are synchronized with the heard voice. Moreover 5-month-old infants look preferentially at the face that matches a heard voice in expressed emotion, rather than one that does not match (Walker, 1982).

From these studies, it may be inferred that what the face is saying can in some way be processed at an early age. Kuhl and Meltzoff's study (1982, 1984) is a first attempt to indicate how speech-specific such abilities may be. They first showed that 18 to 20–week–olds looked longer at the face whose lip movements matched either the heard vowel "a" or "i" than at the face which articulated the other vowel. They then asked whether this ability is due simply to the detection of temporal asynchronies between the onset and offset of acoustic input with lip opening and closing or is specific to the recognition of particular correspondences between a speech sound and its precise articulatory format. By removing the sound-spectral information from the same vowels "a" and "i" but preserving their temporal properties, these authors showed that purely temporal factors were insufficient to produce the preference patterns for seen and heard vowels. Spectral information seems necessary, which

suggests that when infants distinguish lipread /a/ and /i/, they are sensitive to the acoustic correlates of these speech sounds. These authors conclude that speech is likely to be supramodally represented, i.e. that auditory and visual speech information are related to a common supramodal phonetic representation (see Studdert-Kennedy, 1983).

MacKain et al. (1983) also demonstrated a sensitivity of 5 to 6-month-old infants to auditory and visual correlates of speech structures. They showed that infants looked longer at a woman's face articulating CVCV syllables to which they were listening (e.g. "mama") than to the same woman repeating a synchronized competing CVCV ("lulu" in this case), but only when infants were looking at the video display on their right. They concluded that left hemisphere activation facilitated perception of auditory-visual speech correspondences, and argued for the existence of a left hemisphere perceptuo-motor mechanism.

There is considerable experimental evidence that young infants discriminate auditorilly presented phonemes and treat discriminably different members of the same vowel category as equivalent (Eimas et al., 1971; Kuhl, 1983). No similar categorization ability has yet been shown with respect to language as lipread visual input (but for a discussion of this notion of categorization, see Massaro & Cohen, 1983) . Yet it would be very interesting to investigate whether infants are able to differentiate visually discriminable phonemes categorically, i.e. phonemes that differ with respect to the place of articulation feature, when no auditory information is provided.

To investigate this issue further, it may be important to know if infants are subject to auditory-visual illusions as adults and children are. With regard to lipreading, two forms of illusion have been revealed in adults and children; one is often called the McGurk effect, the other the blend illusion (McGurk & MacDonald, 1976; MacDonald & McGurk, 1978; Dodd, 1977; Massaro, 1984, and see Massaro, this volume). Such an experiment is currently in progress (Dodd & Dennis, personal communication), and it should indicate the innate basis of lip reading ability.

The role of lipreading with regard to language acquisition has been explored by Dodd (1987), who reported that access to lipread information has an effect on some aspects of babbling (increase of the number of utterances containing consonants) in 9 to 12-month-old infants.

With regard to the question of how heard and seen speech are integrated, it is interesting to note that the candidate hypotheses are of the same kind as those suggested to account for gestural imitation. Dodd (1983) discussed three possible alternatives and argued in favor of the existence of a nonmodality specific phonological code, i.e. of a common code for processing auditory-visual as well as articulatory speech information from early infancy. Summerfield (1979) also argued that lipreading ability constitutes a convincing argument for information used is essentially provided by articulatory movements. This theoretical position corresponds to that of Spelke with regard to auditory-visual coordination, and more generally has to do with some basic Gibsonian principles in perception. In such a line of reasoning, it would be necessary to demonstrate that visual and articulatory speech perception processes are similar to those that govern auditory speech perception.

Gaze co–orientation between infant and adult

One specific facial act, gaze orientation, is often used by children and adults as a valid cue of the act of " referring", i.e. act by means of which we make use of words or gestures in order to communicate or share a particular knowledge or state of affair. Recent interest in the pragmatics of communication, in particular in the study of the pre-linguistic period as a preparatory period for linguistic communication, has led some authors to examine at what age infants are able to understand the adult's gaze direction (Bruner, 1975; Scaife & Bruner, 1975). The comprehension by infants of the uses of pointing as a referential gesture has also been studied within the same perspective (Bates et al., 1975; Pechman & Deutsch, 1982).

In natural settings the mother very often tracks the gaze of her infant and tries to establish in this way moments of gaze co-orientations (Collis & Schaffer, 1976). Scaife

& Bruner (1975) showed that the gaze direction of infants as young as 4 months can be influenced by the gaze direction of the adult. The role of different spatial indicators in determining this ability, such as landmarks located in the infant's environment, has been analyzed in infants in their first year of life (Butterworth & Cochran, 1980; Butterworth, 1982; Churcher & Scaife,1982; Lempers, 1976). It appears that at first the young infant has only a roughly differentiated notion of where his mother is looking, turning to look to one side only, and not precisely where the mother looks. More finely differentiated reactions are not apparent before 10 or 12 months, more or less at the same time that the hand pointing gesture begins to be used and understood. What is of interest is the fact that from around 3-4 months, infants are able to get information from eye direction whose meaning goes beyond the actual behavior and is related to interindividual communication.

Reflections in Relation to Speech Perception

Imitation, perception of facial gestures and lipreading

The reviewed data shows that the ability of young infants to encode face features and process facial information undergoes a complex development in the first year of life. Moreover , the data is contradictory at least at first glance. In brief, neonates and infants aged less than 2 months can imitate some facial gestures and expressions, they are sensitive to the fact that a face and a voice should share a common spatial location, and can mimic the lip movements of a speaking person. Movement of the seen face plays an important role in eliciting these performances. From a developmental point of view, two of these abilities, coordination between vision and audition imitation — "disappear" between 2 and 3 months to reappear later (at around 4 months for the former ability, 9-14 months for the latter). Other kinds of behavior, not specifically related to face perception, such as reaching, also "disappear" in the first months of life.

Different hypotheses can be suggested to account for such a developmental trend, focusing on changes in peripheral processes (as, for instance, an asychronical development between modifications of the weight of

Our proposal is to radically differentiate between the neonatal level of behavioral organization, called sensori-motor organization, and the new behavioral organization, called perceptuo–motor organization, that develops progressively in the first two years of life. These organizations differ with regard to the code used to process incoming information (sensory versus perceptual code).

some body parts and of their muscular strengths, see Thelen & Fisher, 1982) or focusing on changes in central processes (Mounoud, 1979).

Our proposal (Mounoud &Vinter, 1981) is to radically differentiate between the neonatal level of behavioral organization, called sensori-motor organization, and the new behavioral organization, called perceptuo-motor organization, that develops progressively in the first two years of life (Mounoud, 1984; Vinter, 1986b). These organizations differ with regard to the code used to process incoming information (sensory versus perceptual code).

In this theory it is postulated that neonates possess an innate body representation (or schema), in which information is coded by means of the sensory code. They then construct new representations, a new body representation for instance, by means of the perceptual code. "Representation" or "schema" is defined as an internal organization of contents, of the different properties of objects, situations or events, i.e. as the result of a top–down process. It can also be seen as the result of information selection and information-processing processes. The term "code" is used to mean the set of formal operations or rules that transform or translate the information related to objects or actions. And a representation is understood as a translation of information by means of a particular code.

Within this framework, we claim that the perception of facial movements at birth is qualitatively different from the perception of facial movements appearing later (Vinter, et al., 1986). This is how we explain why neonates are able to process information coming from internal features of the face, as evidenced by the imitation ability, whereas other studies suggest that internal features of the lowerpart of the face are not discriminated before 4 months. In our view, imitation at birth is based on a sensorial coding of information, in which kinetic information is of prime importance, and which does not permit any facial movement to be produced in isolation. Mouth movements for instance are integrated in a more complex sequence in which head movements (and probably arm-hand movements) also intervene

(Vinter, 1985b). By contrast, the ability to perceive facial features demonstrated in a 3 to 4-month-old infant by scanning or preferential studies is based on a perceptual coding of facial information. With this code, movement is not a crucial determinant of feature detection. A specific facial movement can be reproduced in isolation, without other associated movement. Thus the same part of the face may be processed by different codes, and it may be that at the beginning of life, both processes can occur, which may explain some apparent contradictions in the literature. Within such a view, it is crucial to define precisely the specific elicitators of each code. This has yet to be done.

In relation to lipreading ability, we do not yet know whether or not it is present from birth, and we are also ignorant about how it develops. Sensitivity of infants to the integration between lip movements and the sound produced seems to be evident at around 4 months. Whatever its developmental course may be, it is intriguing that lip reading is present when imitation of facial movements has disappeared. More precisely, at the age when infants are able to match the lip movements that produces an "a" (i.e. more or less an opening closing of the mouth) with the sound "a", and moreover are able to produce the sound "a", spontaneously as well as in response to the same auditory input (occurences of vocal imitation are reported by Kuhl & Meltzoff's study), they are at the age when they seem to be unwilling or unable to imitate that movement of mouth opening- closing. By contrast, while some mouth imitation is present at birth (i.e. under 3 weeks of life), speech imitation as a visual-auditory input (i.e. mouth movements and sound) has not been reported*.

Campbell (1986) discusses another provocative contrast by pointing out that the age at which imitation of facial movements is no longer elicitable is precisely the age at

* Kugiumutzakis (1985a) mentioned some cases of imitation of the vowel "a" but without any analysis of the neonate's vocal emission, which appears very doubtful.

4-month-olds can be characterized by a complex behavioral panorama. They can discriminate acoustically different speech units, vocally imitate, lipread, but they no longer imitate a visually perceived mouth or lip movement.

which infants are very sensitive to auditory phonetic contrasts (Eimas et al., 1971). In short, to the extent that the data are reliable**, 4-month-olds can be characterized by a complex behavioral panorama. They can discriminate acoustically different speech units, vocally imitate, i.e. associate a heard sound with an articulatory movement, lipread, i.e. visually perceive that a particular lip shape goes together with a heard acoustic input, but they no longer imitate a visually perceived mouth or lip movement, i.e. produce an articulatory movement that conforms to the perceived model. Infants are about 10 to 14-months before they can again imitate a visually perceived mouth movement, while it seems established that lipreading occurs around 4 months earlier. Thus a dissociation at least partial and temporary between lipreading and imitation must be postulated.

These contrasts raise different fundamental questions which refer to the relationships between perception (auditory perception, visuo-auditory perception of speech), production (speech production, articulatory speech and nonspeech movements production) and imitation (of speech or nonspeech movements). Imitation constitutes a very particular ability since perception and production must closely interact for precise imitation to occur.

The question that arises now is, in what way does the ability to derive speech from seen faces depend on, and relate to, the various mouth imitative skills of the youngest infants, and the demonstrated sensitivities of slightly older ones.

It may first be argued that although both are closely linked to perception of facial movements and are based on intermodal coordination, lipreading ability and facial imitation do not share any common process. A basic

** Experiments on imitation of speech movements produced without associated sounds in comparison with vocal imitation are still needed for a valuable understanding of these questions. Moreover it would be interesting to know if infants of this age are able to visually discriminate different articulatory movements, for instance the visual form associated to the production of "ba" from that associated with "da" or "ga".

difference may be that differentiated face schema in which the mouth and its movements are represented is a necessary condition for imitation, but not for lipreading. If this is so, it is what infants at 4 months lack. Three-month-olds, who no longer reproduce a facial movement, do nevertheless react to such a movement in a specific way, by smiling and vocalizing, i.e. in the same manner that they re-spond to any human moving face. On the other hand, to a manual movement, they respond by looking at their own hand (Vinter, 1986). These reactions have been interpreted as demonstrating that the infant, at this age, is progressively discovering his body.

To what extent can we support the view that lipreading does not require the existence of an abstract face representation, through which one's own mouth may be conceived of as corresponding to another person's mouth? In fact, such close correspondence between one's own body and the body of another may not be necessary for lipreading to occur.

Within a different framework, Campbell (1986) also proposed that lipreading ability must be distinguished from other skills related to face perception. Campbell et al. (1986) showed a complete dissociation between face recognition and classification processes and lipreading in two unilaterally lesioned patients. They concluded that most aspects of lipreading are likely to be more related to language processes than to processing of non- linguistic properties of faces, and as a consequence are likely to be left-hemisphere lateralized.

A different distinction between lipreading and imitation may reflect their different relationship with perception and action. Lipreading may not require an integration between perception and production, i.e. between vision and proprioception (articulatory processes). Lipreading ability would be essentially a perceptual act. This hypothesis does not fit one interpretation of the Motor Theory of speech perception (Liberman & Studdert-Kennedy, 1977), which has been suggested by MacDonald & McGurk (1978) to account for audition-vision fusion "illusions". Within this theory, lipread information is processed in a code derived from articulatory feedback, and thus speech perception cannot be dissociated from speech production.

But such a theory cannot apparently easily account for the differences between lipreading and imitation of facial movements on which we are focusing, since these theorists also reject any idea of mediated perception through internal schemas. Lipreading and imitation could not thus differ with respect to the necessary presence of a differentiated face schema for the latter. By contrast, Straight (1980) favors a theory that would distinguish different mechanisms and different representational basis for speech perception processes on one hand and speech production processes on the other. He suggested a distinction between auditory phonological processes (we could add visual phonological processes) and articulatory phonological processes, imitation being an essential mechanism for the latter processes. Within such a framework, we expect to observe a dissociation during development between lipreading ability as a perceptual process and imitation of mouth movement as a productive process.

Speech perception within a developmental and cognitive model.

In conclusion we will discuss briefly the relationships between face perception processes and visual speech perception within a developmental and cognitive perspective. An account of lipreading ability requires an understanding of how heard and seen speech can be integrated, and we may ask ourselves to what extent speech may be a special phenomenon, i.e. to what extent such auditory-visual integration is special with regard to similar intermodal coordinations governing object perception. In relation to somewhat different problems, we have argued elsewhere that speech constitutes a cognitive system among others, no more specific than others (Mounoud, 1986; Vinter, 1987). This means that the development of lipreading might be described by the same cognitive model that is used to understand the development of face perception (Vinter et al., 1986).

Different hypotheses about visual-auditory speech integration can be generated from this model (see Mounoud (1984) or Vinter (1986b) for a presentation of the model):

- integration between seen and heard speech in the neonatal period (0-1 month) should be based on physical,

i.e. acoustic and visual properties of heard speech and seen faces and not on more cognitively mediated perceptual properties such as phonemic categorisation;

- this ability may follow a U-shaped development, between birth and 6 or 8 months, i.e. may disappear sometime after the first month of life, and reappear later;
- when infants are again sensitive to seen and heard speech in the middle of the first year, this visual-auditory speech integration is likely to be qualitatively different from the form present at birth, and is possibly based on spectral information, as demonstrated by Kuhl & Meltzoff (1984). It means that an experiment such as Kuhl & Meltzoff carried out with infants aged less than 1 month may demonstrate their sensitivity to sound-face synchrony, but *not* to specific vowels (not on the basis of spectral information);
- as far as a complex visual-auditory speech integration such as the McGurk effect requires that the phonemes are perceived as interrelated units, and not as independent speech units (Massaro's model argues against this assumption), we might not expect to observe it before the end of the first year or the second year of life.

If such hypotheses can be empirically validated, it may be possible to sustain the idea of the nonspecificity of speech perception processes, at least in infancy and with regard to processing of visual-auditory information.

Generally, we argue that lipreading ability, as a particular speech perception skill, qualitatively changes during the first years of life, and its relationships with imitation or other speech perception processes may also vary during development. With regard to speech perception, MacKain (1987) discusses in a very interesting way three theoretical alternatives that are currently being taken up by psycholinguists:

- the phonetic view, which claims that infants perceive phonetic structures, i.e. are sensitive to the abstract phonetic features of phonetic segments (Eimas, 1975);
- the auditory view, which postulates the existence of an auditory mechanism, not specific to human speech but common to all mammals, and which considers that

Generally, we argue that lipreading ability, as a particular speech perception skill, qualitatively changes during the first years of life, and its relationships with imitation or other speech perception processes may also vary during development.

infants are sensitive to the acoustic attributes that distinguish phonetic features without having an abstract phonetic code at their disposal (Stevens, 1975). We think that this view may very well be extended to visual information related to speech: infants may recognize visual speech patterns through a general mechanism of visual pattern recognition. Both sensory information sources may then be integrated on the basis of the principles described by Massaro & Cohen (1983) for instance;

- the perceptuo-motor view, which closely links speech perception processes with the motor activity of speech production (articulation) and thus suggests that infants are sensitive to the phonetic articulatory information in the speech spectrum, i.e. that auditory and visual speech information provide directly information about the underlying articulatory gestures (Studdert-Kennedy, 1986).

To be able to argue for or against any one of these views, which suggest different levels of speech processing, it appears essential to know the unit of perception with which infants process speech (e.g. syllable, word, phoneme), and whether or not this unit changes with development.

Such questions are in no way language-specific. We are confronted with exactly the same questions when trying to analyze the development of a psychomotor ability, such as reaching (Vinter, in press *a*). We need to briefly develop this point before going back to speech perception. With regard to reaching (see Mounoud, 1983), we have suggested that three different levels of processing (or "codes") can be distinguished between birth and 2 years, and which are successively predominant without ever disappearing: a sensory level (predominant from birth until around 4 months), a perceptual level (from 1 month to 24 months), and a conceptual level (from 16 to 18 months until 10 years). The crucial dimension of differentiation between the sensory and perceptual levels is related to the "referential relationship" between object, subject and meaning; this relationship is necessarily undifferentiated at the sensory level (for example, incoming information cannot be related to the object's

properties by the subject) but is differentiated at the perceptual level. Moreover, the unit of perception (the "segmentation problem") evolves with development and always through the same steps, whatever the level of processing:
1. uncoordinated and partial segments;
2. wholistic and nondecomposable units; 3. units partially and then completely decomposable in their constitutive segments.

Within this framework, it maybe suggested that:

- speech is initially (at the beginning of life) processed at a sensory level, i.e. at the acoustic auditory or auditory-visual level. Infants can discriminate auditory speech contrasts, independently of any segmentation specific to their language, without any meaning associated to these sound contrasts. Auditory perception maybe categorical, as held by Eimas (1975), without involving the existence of a phonetic code.

- speech will then be progressively processed at a perceptual level, and different steps in the processed units of perception can be distinguished. The kind of speech processing postulated by the perceptuomotor view necessarily belongs to this level, because of the implicit assumption of a differentiated subject-object (articulation-perceptual speech information) relationship. The auditory view may also belong to this level, for incoming information can be processed at a sensory as well as at a perceptual level (i.e. without or with the ability to refer information to the object's properties).

- the *syllable*, which can be understood as an elementary and independent unit (in contrast to phonemes, whose definitions are based on their interrelationships), may be the first speech unit of the perceptual level. Visuo-auditory information may very well specify an articulatory pattern, although the processed unit is not the word but a syllable (see MacKain, who argues that the perceptuomotor view requires a wholistic unit such as the word), but in no way is this specification "direct" in our opinion. An internal representation of the incoming auditory-visual information must be postulated to account for speech perception and speech production. Language pathology shows cases in which neither perception nor

production of speech stimuli (acoustic discrimination versus spontaneous production) are distorted when assessed independently, but are disturbed when they must be integrated.

- then, the *word* is likely to be the unit of speech perception (by around 9-12 months). Meaning conveyed by the speech sound contrasts plays a crucial role in determining specific auditory-visual-articulatory associations.
- finally, the unit at which speech is processed may be *phonemic*. We fully agree with MacKain (1987) that the knowledge infants acquire about phonetic segments results from analysis subsequent to their sensitivity to the whole word. Mounoud (1986) described a similar transition from syllabic to phonemic segmentation in reading.

The various assertions stated above, if valid, make it clear that speech perception processes are not developmentally different from general object perception processes. Moreover, the different theoretical views of speech perception should not be considered as competing alternatives. They are all valid, depending on the developmental step under consideration. The fact that speech perception processes have mainly been studied in adults is probably responsible for this state of "competition" between theories. We think that detailed analysis of the speech stimuli and of the experimental situations should reveal that, in adults too, qualitatively different levels of processing and different units of speech perception can be contemporaneously observed.

About the author

Annie Vinter is a "Maître d'Enseignement et de Recherche" at the University of Geneva, Faculty of Psychology. She was trained in Geneva, and spent some time in Italy (Scientific Institute Stella Maris of Pisa) and Germany (Interdisciplinary Research Center of Bielefeld). Her basic research interest is in human development. She has been involved in rsearch projects with Pierre Mounoud (Geneva). She studied the development of self-image from childhood to adolescence, early imitative ability, and auditory-visual coordination in the first months of life. She is presently carrying out a study of handwriting development in children.

References

Abravanel, E., Sigafoos, A.P. 1984. Exploring the presence of imitation during early infancy. *Child Development,* 55, 381–392.

Atkinson, J., Braddick, 0., Moar, K. 1977. Development of contrast sensitivity over the first three months of life in the human infant. *Vision Research,* 17, 1037–1044.

Aronson, E., Rosenbloom, S. 1971. Space perception in early infancy: perception within a common auditory-visual space. *Science,* 172, 1161–1163.

Banks, M.S. 1982. The development of spatial and temporal contrast sensitivity. *Current Eye Research,* 2, 191–98.

Banks, M.S. 1985. How should we characterize visual stimuli? In G. Gottlieb & N.A. Krasnegor (Eds.). *Measurement of audition and vision in the first year of postnatal life.* Norwood: Ablex.

Banks, M.S., Salapatek, P. 1978. Acuity and contrast sensitivity in 1-, 2-, and 3-month-old human infants. *Investigative Ophthalmology and Visual Science,* 17, 361–365.

Banks, M.S., Salapatek, P. 1981. Infant pattern vision: a new approach based on the contrast sensitivity function. *Journal of Experimental Child Psychology,* 31, 1–45.

Banks, M.S., Salapatek, P. 1983. Infant visual perception. In P.H. Mussen (Ed.). *Handbook of Child Psychology. Vol. II.* N.Y.: Wiley.

Bates, E., Camaioni, L.,Volterra,V. 1975. The acquisition of performatives prior to speech. *Merrill Palmer Quarterly,* 21, 205–226.

Brazelton, T.B.,Young, G.C. 1964. An example of imitative behavior in a nine-week-old infant. *Journal of the American Academy of Child Psychiatry,* 4, 53–67.

Bruner, J.S. 1975. The ontogenesis of speech acts. *Journal of Child Language,* 2, 1–19.

Bushnell, I.W.R. 1979. Modification of the externality effect in young infants. *Journal of Experimental Child Psychology,* 28, 211–229.

Butterworth, G. 1980. The origins of auditory-visual perception and visual proprioception in human development. In H. Pick & K.Walk (Eds.). *Perception and Experience.* N.Y.: Plenem Press.

Butterworth, G. 1982. Object permanence and identity in Piaget's theory of infant cognition. In G. Butterworth (Ed.). *Infancy and Epistemology.* Brighton: Harvester Press.

Butterworth, G., Cochran, E. 1980. Towards a mechanism of joint visual attention in human infancy. *International Journal of Behavioural Development.* 3, 253–270.

Campbell, R. 1986. Lip reading. In A.W. Young & A.D. Ellis (Eds.). *Handbook of Research in Face Processing.* Amsterdam: North Holland.

Campbell, R., Landis, T. Regard M., 1986. Face recognition and lipreading: a neurological dissociation. *Brain,* 109, 509–521.

Caron, A.J., Caron, R.F., Caldwell, R.C., Weiss, S.J. 1973. Infant perception of the structural properties of the face. *Developmental Psychology,* 9, 385–399.

Castillo, M., Butterworth, G. 1981. Neonatal localisation of a sound in visual space. *Perception,* 10, 331–338.

Churcher, J., Scaife, M. 1982. How infants see the point. In G. Butterworth & P. Light (Eds.). *Social cognition.* Brighton: Harvester Press.

Cohen, S. 1974. Developmental differences in infants' attentional responses to face–voice incongruity of mother and stranger. *Child Development,* 45, 1155–1158.

Condon, W.S., Sander, L.W. 1974. Synchrony demonstrated between movements of the neonate and adult speech. *Child Development,* 45, 456–462.

Dodd, B. 1977. The role of vision in the perception of speech. *Perception,* 6, 31–40.

Dodd, B. 1979. Lipreading in infants attention to speech presented in- and out-of-synchrony. *Cognitive Psychology,* 11, 478–484.

Dodd, B. 1983. The visual and auditory modalities in phonological acquisition. In A.E. Mills (Ed.). *Language acquisition in the blind child.* San Diego: College-Hill Press, 57–61.

Dodd, B. 1987. The acquisition of lipreading skills by normally hearing children. In B. Dodd & R. Campbell (Eds.). *Hearing by Eye: the Psychology of Lipreading.* N.Y.: Erlbaum.

Dunkeld, J. 1978. The function of imitation in infancy. *Unpublished Ph.D. Thesis of the University of Edinburgh.*

Eimas, P.D. 1975. Auditory and phonetic coding of the cues for speech: Discrimination of the (r-1) distinction by young infants. *Perception & Psychophysics,* 18, 341–347.

Eimas, P.D., Siqueland, E.R., Jusczyk, P., Vigorito, J. 1971. Speech perception in infants. *Science,* 171, 303–306.

Fantz, R. 1966. Pattern discrimination and selective attention as determinants of perceptual development from birth. In A. Kidd and J. Rivoire (Eds.). *Perceptual development in children.* New York: International University Press.

Fantz, R., Fagan, J.F. 1975. Visual attention to size and number of pattern details by term and preterm infants during the first six months. *Child Development,* 16, 3–18.

Fantz, R., Fagan, J.F., Miranda, S.B. 1975. Early visual selectivity. In L.B. Cohen & P. Salapatek (Eds.). *Infant perception: from sensation to cognition.* Vol. 1. N.Y.: Academic Press.

Field, T.M., Woodson, R., Greenberg, R., Cohen, D. 1982. Discrimination and imitation of facial expressions by neonates. *Science,* 218, 179–182.

Fontaine, R. 1982. Conditions d'évocation des conduites imitatives chez l'enfant de 0 à 6 mois.*Unpublished Ph.D. Thesis of The University of Paris.*

Gardner, J., Gardner, H. 1970. A note on selective imitation by a six-week-old infant. *Child Development,* 41, 1209–1211.

Goren, C.C., Sarty, M., Wy, P.Y.K. 1975. Visual following and pattern discrimination of face-like stimuli by newborn infants. *Pediatrics,* 56, 544–549.

Guernsey, M. 1928. Eine genetische Studie über Nachahmung. *Zeitschrift für Psychologie,* 107, 105–178.

Hainline, L. 1978. Developmental changes in the scanning of face and non-face patterns by infants. *Journal of Experimental Child Psychology,* 25, 90–115.

Hayes, L.A., Watson, J.S. 1981. Neonatal imitation: fact or artifact? *Developmental Psychology,* 17, 655–660.

Jacobson, S.W. 1979. Matching behavior in the young infant. *Child Development,* 50, 425–430.

Karmel, B.Z. 1969. The effect of age, complexity and amount of contour on pattern preferences in human infants. *Journal of Experimental Child Psychology,* 7, 339–354.

Koepke, J.E., Hamm, M., Legerstee, M. 1983. Neonatal imitation: two failures to replicate. *Infant Behavior and Development,* 6, 97-102.

Kugiumutzakis, J. 1985a. Imitation in newborns 10-45 minutes old. *Uppsala Psychological Reports,* 376, 1–16.

Kugiumutzakis, J. 1985b. Development of imitation during the first six months of life. *Uppsala Psychological Reports,* 377, 1–21.

Kuhl, P.K. 1983. Perception of auditory equivalence classes for speech in early infancy. Infant Behavior and Development, 6, 263–285.

Kuhl, P.K., Meltzoff, A.N. 1982. The bimodal perception of speech in infancy. *Science,* 218, 1138–1141.

Kuhl, P.K., Meltzoff, A.N. 1984. The intermodal representation of speech in infants. *Infant Behavior and Development,* 7, 361–381.

Lempers, J.M. 1976. Production of pointing, comprehension of pointing, and understanding of looking behavior in young children. *Unpublished Ph.D. Thesis of the University of Minnesota.*

Lewis, M., Wolan-Sullivan, M. 1985. Imitation in the first six months of life. *Merrill Palmer Quarterly,* 31, 315–333.

Liberman, A.M., Studdert-Kennedy, M. 1977. Phonetic Perception. In R. Held, H. Leibowitz & H.L. Teuber (Eds.). *Handbook of sensory physiology. Vol. 8.* Heilderberg: Springer-Verlag.

Lyakh, G.S. 1968a. Articulatory and auditory mimicry in the first months of life (in Russian). *Zhurnal Vysshey Nervnoy Deyatel'nosti imeni I.P. Pavlova,* 18, 831–835.

Lyakh, G.S. 1968b. Characteristics of conditioned connections in mimo-articulatory and auditory components of speech stimuli in the first year of life (in Russian). *Zurnal Vysshey Nervnoy Deyatel'nosti imeni I.P. Pavlova,* 18, 1069–1071.

MacDonald, J., McGurk, H. 1978. Visual preferences of speech perception processes. *Perception & Psychophysics,* 6, 230–245.

McGurk, H., Lewis, M. 1974. Space perception in early infancy: perception within a common auditory–visual space? *Science,* 186, 649–650.

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

McGurk, H., Turnure, C., Creighton, S.J. 1977. Auditory-visual coordination in neonates. *Child Development,* 48, 138–143.

MacKain, K.S. 1987. Filling the gap between speech and language. In M.D. Smith, J.L. Locke (Eds.). *The emergent lexicon: the child's acquisition of a linguistic vocabulary.* N.Y.: Academic Press.

MacKain, K.S., Studdert-Kennedy. M., Spieker, S., Stern, D.S. 1983. Infant intermodal speech perception is a left-hemisphere function. *Science,* 219, 1347–1349.

McKenzie, B., Over, R. 1983. Young infants fail to imitate facial and manual gestures. *Infant Behavior and Development,* 6, 85–89.

Maratos, 0. 1973. The origin and the development of imitation during the first six months of life. *Unpublished Doctoral Dissertation of the University of Geneva.*

Maratos, 0. 1982. Trends in the development of imitation in infancy. In T.G. Bever (Ed.). *Regressions in Mental Development: Basic phenomena and theories.* Hillsdale: Lawrence Erlbaum.

Marq, E., Freeman, D.N., Peltzman, P., Goldstein, P.J. 1976. Visual acuity development in human infants: evoked potentials measurements. *Investigative Ophthalmology in Visual Sciences,* 15, 150–153.

Massaro, D.W. 1984. Children's perception of visual and auditory speech. *Child Development,* 55, 1777–1788.

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.

Maurer, D., Salapatek, P. 1976. Developmental changes in the scanning of faces by young infants. *Child Development,* 47, 523–527.

Meltzoff, A.N., Moore, K.M. 1977. Imitation of facial and manual gestures by human neonates. *Science,* 198, 75–78.

Meltzoff, A.N., Moore, K.M. 1982. The origins of imitation in infancy: paradigm, phenomena, and theories. In L.P. Lipsitt & C.K.Rovee-Collier (Eds.). *Advances in Infancy research.* Norwood: Ablex, 263–299.

Milewski, A. 1976. Infants' discrimination of internal and external pat-tern elements. *Journal of Experimental Child Psychology,* 22, 229–246.

Mounoud, P. 1979. Développement cognitif: construction de structures nouvelles ou construction d'organisations internes. *Bulletin de Psychologie,* 36, 107–118.

Mounoud, P. 1983. L'évolution des conduites de préhension comme illustration d'un modèle de développement. In S. de Schonen (Ed.). *Le développement dans la première année de la vie.* Paris: PUF.

Mounoud, P. 1984. A point of view on ontogeny. *Human development,* 27, 329–334.

Mounoud, P. 1986. Similarities between developmental sequences at different age periods. In I. Lewin (Ed.). *Stage and Structure.* Norwood: Ablex, 40-58.

Mounoud, P., Vinter, A. 1981. Representation and sensori-motor development. In G. Butterworth (Ed.), *Infancy and Epistimology.* Brighton: Harvester Press, 200–235.

Muir, D., Clifton, R.K. 1985. Infants' orientation to the location of sound sources. In G. Gottlieb & N.A. Krasnegor (Eds.). *Measurement of audition and vision in the first year of postnatal life.* Norwood: Ablex, 171–194.

Muir, D., Abraham, W., Forbes, B., Harris, L. 1979. The ontogenesis of an auditory localization response from birth to four months of age. *Canadian Journal of Psychology,* 33, 320–333.

Papousek, H., Papousek, M. 1979. Early ontogeny of human social interaction: its biological roots and social dimensions. In P. von Cranach, K. Foppa, W. Loprnies, D. Plooj (Eds.). *Human Ethology.* Cambridge: Cambridge University Press.

Papousek, H., Papousek, M. 1982. Vocal imitation in mother-infant dialogue. *Paper presented at the International Conference of infant studies,* Austin.

Pechman, T., Deutsch, W. 1982. The development of verbal and nonverbal devices for reference. *Journal of Experimental Child Psychology,* 34, 330–341.

Piaget, J. 1945. *La formation du symbole chez l'enfant.* Neûchatel et Paris: Delachaux et Niestlé.

Razran, G. 1971. *Mind in evolution, an East-West synthesis of learned behavior and cognition.* Boston: Houghton-Mifflin.

Salapatek, P. 1975. Pattern perception in early infancy. In L.Cohen & P. Salapatek (Eds.). *Infant perception.* N.Y.: Academic Press.

Scaife, M., Bruner, J. 1975. The capacity of joint visual attention in the infant. *Nature,* 253, 265–266.

Spelke, E.S. 1976. Infants' intermodal perception of events. *Cognitive Psychology,* 8, 553–560.

Spelke, E.S. 1985. Preferential looking methods as tools for the study of cognition in infancy. In G. Gottlieb & N.A. Krasnegor (Eds.). *Measurement of audition and vision in the first year of postnatal life.* Norwood: Ablex, 31–52.

Spelke, E.S., Cortelyou, A. 1981. Perceptual aspects of social knowing: looking and listening in infancy. In M.E. Lamb & L.R. Sherrod (Eds.). *Infant social cognition.* Hillsdale, N.J.: Erlbaum.

Spelke, E.S., Owsley, C.J. 1979. Intermodal exploration and perceptual knowledge in infancy. *Infant Behavior and Development, 2,* 13–28.

Stevens, K.N. 1975. The potential role of property detectors in the perception of consonants. In G. Fant & M.A.Tatham (Eds.). *Auditory analysis and perception of speech.* N.Y.: Academic Press.

Studdert-Kennedy M. 1983. On learning to speak. *Human Neurobiology, 2,* 191–195.

Studdert-Kennedy, M. 1986. Sources of variability in early speech development. In J.S. Perkell & D.H. Klatt (Eds.). *Invariance and variability of speech processes.* Hillsdale, N.J.: Erlbaum.

Straight, H.S. 1980. Auditory versus articulatory phonological processes and their development in children. In G.H. Yani-Komshian, J. F. Kavanagh & C.A. Ferguson (Eds.). *Child Phonology. Vol. 1.* N.Y.: Academic Press, 43–71.

Summerfield, 0. 1979. Use of visual information for phonetic perception. *Phonetica, 36,* 314–331.

Thelen, E., Fisher, D.M. 1982. Newborn stepping: an explanation for a "disappearing" reflex. *Developmental Psychology, 18,* 760–775.

Trevarthen, C. 1974. The psychobiology of speech development. In E.H. Lenneberg (Ed.). *Language and Brain: Developmental aspects.* Boston: Neurosciences Research Program.

Trevarthen, C. 1979. Communication and cooperation in early infancy: a description of primary intersubjectivity. In M. Bullowa (Ed.). *Before Speech: the beginning of interpersonal communication.* Cambridge: Cambridge University Press, 321–347.

Trevarthen, C. 1980. The foundations of intersubjectivity: Development of interpersonal and cooperative understanding in infants. In D. R. Olson (Ed.). *The social foundation of language and thought.* N.Y.: Norton.

Trevarthen, C. 1982. Basic patterns of psychogenetic change in infancy. In T.G. Bever (Ed.). *Regressions in mental development: basic phenomena and theories.* Hillsdale, N.J.: Erlbaum, 7–46.

Uzgiris, I.C., Hunt, J. 1975. *Assessment in infancy.* Urbana, Ill: University of Illinois Press.

Vinter, A. 1985. *L'imitation chez le nouveau-né.* Paris and Neuchatel: Delachaux and Niestlé.

Vinter, A. 1986a. The role of movement in eliciting early imitations. *Child Development, 57,* 66–71.

Vinter, A. 1986b. A developmental perspective on behavioral determinants. *Acta Psychologica,* 63, 337–363.

Vinter, A. 1987. Les fonctions de représentation et de communication dans les conduites sensori-motrices. In J. Piaget, P. Mounoud & J.P. Bronckart (Eds.). *La Psychologie.* Encyclopédie La Pléiade, Paris: Gallimard.

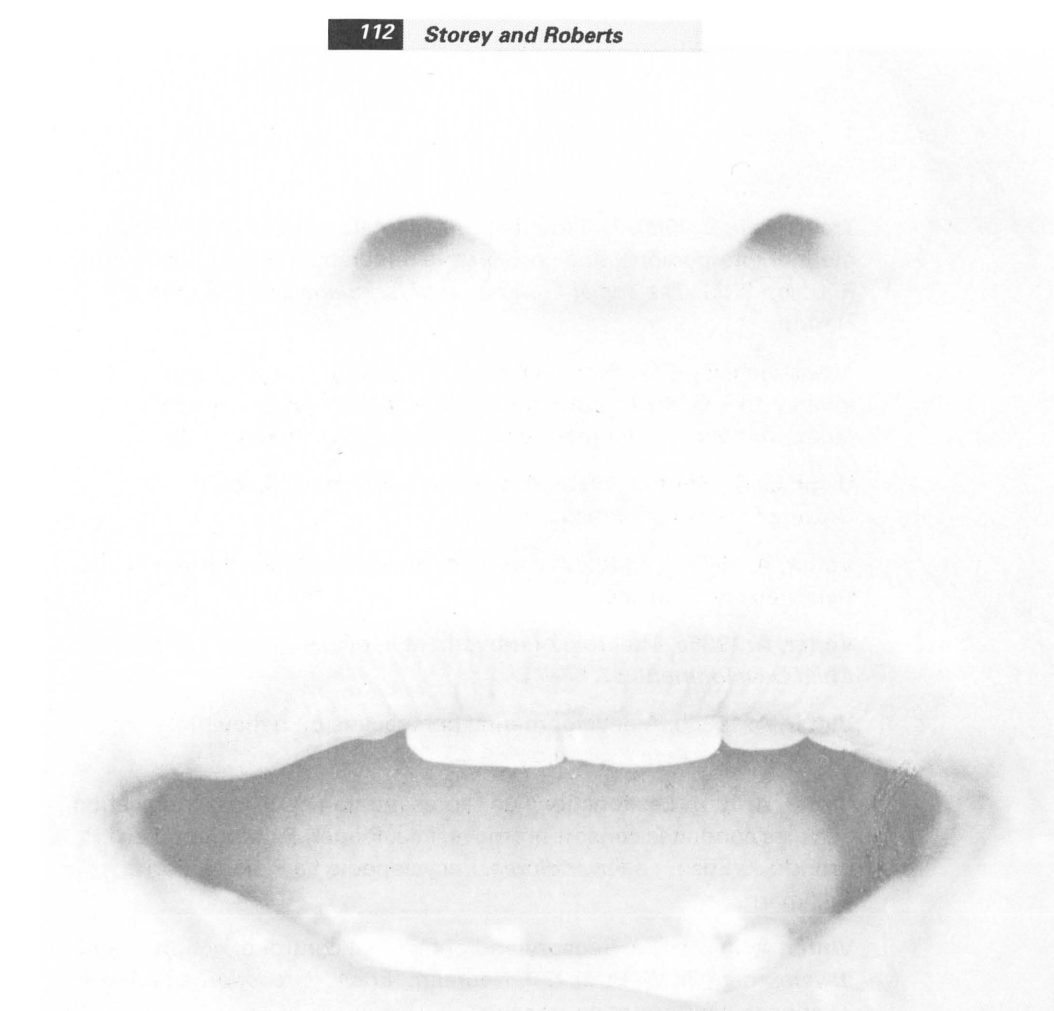
Vinter, A. in press *a.* Sensory and perceptual control of action in early development. In W. Prinz & 0. Neuman, (Eds.). *Perception and Action relationships: Current approaches.*

Vinter, A., Lanares, J., & Mounoud, P. 1986. Development of face perception. *Reports of the Perception and Action Research Group of The Bielefeld University.* # 114.

Vinter, A., de Nobili, G.L., Pelligrinetti, G., & Cioni, G. 1986. Auditory–visual coordination: does it imply an external world for the newborn? *Cahiers de Psychologie Cognitive,* 4, 309–322.

Walker, A.S. 1982. Intermodal perception of expressive behaviors by human infants. *Journal of Experimental Child Psychology,* 33, 516–535.

Zazzo, R. 1957. Le problème de l'imitation précoce chez le nouveau-né. *Enfance,* 10, 135–142.



Reading the Speech of Digital Lips

Reading the Speech of Digital Lips: *Motives and Methods for Audio – Visual Speech Synthesis*

Darryl Storey and Martin Roberts

Department of Computer Studies, Loughborough University of Technology

Visible Language XXII, 1
Darryl Storey and Martin Roberts, pp. 112–127

The widespread practice of lipreading among the hearing impaired has, for a number of years, stimulated research into the feasibility of transmitting visible images of articulation to accompany acoustically conveyed speech, in those circumstances where visual reinforcement of the speech signal is typically lacking. Although there already exist several systems which, exploiting computer graphics, are capable of generating animated images of articulation while allowing for eventual audio/visual synchrony, each is open to criticism on the grounds of its perceptual inadequacy and/or cost. This paper offers a brief review of these initiatives to date and describes the recent development of a relatively simple, effective, and hence economical method of audio/visual speech synthesis.

Introduction

Aspects of visible speech

The skill of lipreading (speech reading) is practised by many partially hearing listeners in order to offset, at least to some degree, their own aural limitations. For the hearing-impaired exponent, lipreading provides a perceptual supplement by means of which the intelligibility of heard speech may be substantially enhanced. It is important to recognize however, that such improvement is not brought about as an consequence of the lipreader's experiencing some visual analog of amplification. Speech which is both seen and heard frequently appears more intelligible than any wholly acoustical counterpart, because the separate visual and acoustical representations are perceptually complementary.

Those of us with normal hearing can, of course, hold intelligible conversations without necessarily facing each other in order to do so. Unimpaired, the auditory system is perfectly able to detect and resolve all of the changes in acoustical frequency, and their relationships in time, which typify human speech. All too often however, hearing loss manifests itself as a reduced sensitivity to specific frequency regions of the acoustical spectrum; those very regions within which a variety of linguistic/phonetic distinctions are portrayed in sound.

Under such circumstances one's auditory perception of the temporal order of acoustical events may remain reasonably acute while the 'identity' of certain phonetic events may be obscured. The benefit of lipreading resides in the perceptual restoration of such lost phonetic identities. For example, a commentator's pronouncement to the effect: "I sought the President's opinion. . ." could easily be misconstrued by a listener with a hearing loss as "I thought President's opinion. . ." or worse still, "I fought the President's opinion. . ." due entirely to the perceived similarity between the productions of "s", "th", and "f". Visually, however, these elements of our phonetic repertoire are distinct. This situation pertains virtually across the board as far as place of articulation is concerned. For a hearing–impaired listener to distinguish "pot" from "tot", and "cot", seeing the word spoken is at least as useful as hearing it, if not more so.

Simply being able to view a talker's face in action however, is not necessarily the *sine qua non* of lipreading.

Simply being able to view a talker's face in action, however, is not necessarily the *sine qua non* of lipreading. Even visible enunciations may be 'more' or 'less' clear to interpret. Moustaches and beards can be detrimental to the lipreader's art and poor lighting conditions, coupled with an idiosyncratic articulatory style, can result in relatively little of a talker's speech being visually informative. Notably, for the purpose of lipreading, it is particularly helpful for the viewer to be able to discern the behavior of the talker's tongue!

In point of fact, perceptual reference to the visual aspect of speech communication is not a trait of the hearing-impaired exclusively. The capabilities of highly skilled lipreaders actually reflect something approaching the asymptote in exploitation of a tendency which normal hearing viewer/listeners share, i.e. the *de facto* integration of synchronous auditory and visual images of speech. Such a tendency is born out, not so much by the perceptual integration of ostensibly compatible components (there were many connected with the silent film industry who doubted that 'the talkies' would work) as by convincing experimental demonstrations of the ease with which seemingly incompatible elements could be fused, resulting in audio-visual 'illusions' of the type first reported by McGurk and Macdonald (1976). For many observers, the perceptual consequence of attending to an audio-visual presentation of the acoustically unambiguous syllable /ba/ synchronized with a video image of the syllable /ga/ is not /ba/ nor /ga/ but /da/.

Considerable research has gone into mapping out the common articulatory ground which visual and acoustical instances of speech clearly share (e.g., Summerfield, 1979; Campbell and Dodd, 1980). Running parallel with such examinations have been various endeavours to analyze the visual concomitants of speech, both for the purpose of understanding articulatory processes in speech production (e.g., Perkell, 1986; Abry and Broe, 1986), and in the hope of providing the fundamental data from which 'artificial' representations of human talkers might be reliably reconstructed (e.g. Montgomery, 1983; Montgomery and Jackson, 1983; Brooke and Summerfield, 1983).

Computer faces: The problems that computers face

A variety of methods for computer-controlled synthesis of visible articulatory gestures have been explored to date (e.g. Boston, 1973; Erber and De Filippo, 1978; Montgomery, 1978; Brooke and Summerfield, 1983), although not as contributors to audio-visual productions necessarily. Arguably the most sophisticated of such synthetic faces, both computationally and graphically, is that developed by Parke (1975, 1982). This three computationally dimensional model *has* been incorporated into a system for synchronous audio-visual synthesis (Pearce, Wyvill, Wyvill and Hill, 1986).

None of these methods are entirely without limitation however, either as regards their visual sufficiency or their cost-effectiveness. The vector (outline) graphic images generated by the system of Montgomery (1978) for example, deal somewhat less than adequately with the problem of representing the behavior of the talker's tongue. Brooke's (1982) method, and the infinitely more complex model of Parke (1975), each avoid this issue, although ultimately they cannot escape it, by omitting the tongue altogether. It could of course be argued that this actually constitutes a more, rather than less authentic interpretation, since, during face-to-face communication, the talker's tongue is largely obscured within the shadow of the oral cavity. However, since the objective is to supplement the listener's acoustical analysis with a complementary visual one, strategies for the development of visual prostheses aimed at enhancing the perception of speech, and which exclude from the outset certain known and powerful visual cues, must be open to question.

The readiness with which the hearing-impaired appeal to lipreading, and its demonstrable efficacy as a perceptual strategy (Summerfield, 1987), even among 'normal' listeners (Reisberg, McLean and Goldfield, 1987), argue strongly for the visual reinforcement of speech in those settings where it is not usually provided, e.g., during public address annoucements, telephone con-versations, and radio broadcasts. Nevertheless, success at such a task would require a robust methodology for inferring

high-level articulatory 'primitives' from the acoustical outcome of articulatory gestures, i.e., the capability for working 'backwards' toward some explicit, unambiguous, and continuously changing specification of an articulatory source from what are, in effect, residual acoustical data. These primitives would then have to be reinterpreted as graphical, rather than acoustical coordinates. However, there is, as yet, a complete absence of any appropriate repertoire of such articulatory parameters.

The difficulties these objectives present have been emphasized elsewhere by others (e.g. Sondhi and Resnick, 1983; Levinson and Schmidt, 1983). Of the many methodological and engineering problems, two in particular need to be overcome. The first is the one-to-many relationship between speech events and the variety of possible configurations of the sound source from which these events might plausibly originate. The second is an analogous, and inverse, many-to-one mapping. Listeners commonly assign an equivalent perceptual label to quite discrepant acoustical structures (the different members of any linguistic community are perfectly intelligible, each to the other, despite their acoustically distinct spoken versions of particular language tokens. Although the phonological rules of language offer some hope for one's being able to constrain the number of options which might need to be considered by automated 'decision' processes, any undertaking of the kind considered here would remain formidable.

Of the many methodological and engineering problems, two in particular need to be overcome. The first is the one-to-many relationship between speech events and the variety of possible speech configurations of the sound source from which those events might plausibly originate. The second is an analogous, and inverse, many-to-one mapping.

Given that graphical enhancement of acoustically realized speech would be both useful and desirable, the most straightforward approach to these issues would be to deal, in the first instance, with synchronous audio-visual synthesis. With the output of both graphical and acoustical media under one's direct control, it is pos-sible to address, purposefully, at least one significant problem; that of achieving flexible, yet synchronous audio-visual output.

First Principals

A cornerstone of much speech science research is the well documented discrepancy between the phonetic nature of our perceptual interpretation of speech, and the acoustical signal's being wholly lacking in any

correspondingly discrete elements (e.g. Liberman, Cooper, Shankweiler and Studdert-Kennedy, 1967). When speech is articulated it emerges as a dynamic, continuously changing pattern of sound. When speech is perceived it registers as a succession of phonologically defined components of the listener's language. Thus a more-or-less continuous signal is, apparently, *analyzed* into a sequence of more-or-less descrete percepts. An analogous, though converse 'discrete-continuous' dichotomy pertains with respect to our perception of apparent motion. It has been known since the earliest days of animation with the Victorian Zoetrope, that discrete images, when presented in sufficiently rapid succession, can induce the perception of continuous movement. In this instance discrete events are synthesized into a perceptual representation of seemingly continuous activity [see Ramachandran and Anstis (1986) for a more detailed discussion].

A more-or-less continuous signal is, apparently, *analyzed* into a sequence of more-or-less discrete percepts.

Taking these two observations together, suggests a way in which a graphical simulation of articulation might be arranged to coincide with an acoustical complement. A set of individual frames, representing discrete phonetic states, and presented in rapid succession, could be perceptually representative of articulatory motion. The complexities inherent in alternative approaches, and the force of 'apparent motion' phenomena, encouraged our exploration of this particular class of audio-visual relationship.

An Audio-Visual Synthesis System

The components

Real-time audio-visual synthesis has been accomplished using a BBC microcomputer (plus associated disk drive and monitor) in conjunction with a digital speech synthesizer provided by Loughborough Sound Images (LSI) limited. Acoustical productions are defined by a number of parameters stored on, and retrieved from disk. These parameters are interpreted by the synthe-sizer's digital circuitry, at a uniform sample rate, as settings for a series of electronic filters, which, in turn, simultaneously pass modified signals to the loudspeak-er. Each sample is separated from its predecessor by an interval of 10 milliseconds. The filtered signals, when realized in analog form, constitute speech sounds.

The graphical constituents of the system owe their origin to a Micro-Robotics 'Snap Camera'. This small device connects directly to the microcomputer. It focuses light, not onto a film plane, but onto an array (256 x 128) of light-sensitive cells. The inital, stable voltage output of each cell defines its state to the computer as 'on'. Light striking the cells of the camera causes them to discharge until they reach a voltage 'ceiling' at which point they turn themselves 'off '. These 'on/off ' states are interpreted by the computer as 'black' and 'white' respectively. The resultant distribution of black and white locations within the 256 x 128 point (pixel) array can then be stored, manipulated, and displayed to the VDU.

These are essential devices in themselves, but the key to operational success is a computer program which allows a recorded image to be edited interactively. The values stored within the global array can be changed, i.e., reversed, by manipulating them within 8 x 8 pixel blocks, a block at a time. By editing the data in this way, any recorded image can be completely transformed, as figures 1- 3 illustrate. The result can be made to appear as 'nice' or as 'nasty' as may be necessary.

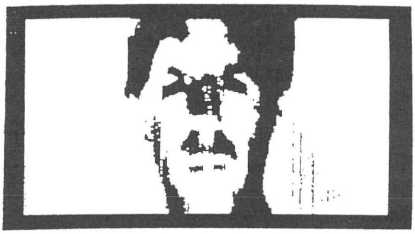


Figure 1
Initial face image recorded via the 'Snap Camera'.

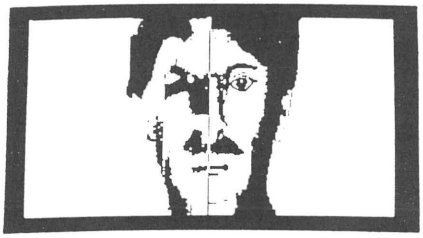


Figure 2
Transitional image achieved through the use of a 'pixel editor '.

Figure 3
Final image for animation.

Illumination and image enhancement

Under normal daytime conditions, the light we receive originates entirely from the one source, i.e., the sun. Although diffused, diffracted and scattered to some extent by the earth's atmosphere, natural sunlight tends to illuminate objects differently according to their position relative to the light's origin. The perceptual experience of 'highlight' and 'shadow' in our visual interpretation of objects in the world is therefore commonplace. In view of the inadequacy of natural daylight as a source of illumination for digital images captured via the Snap Camera, the face eventually recorded as a basis for graphical articulation, was illuminated in the first instance using two artifical light sources (100 watt standard lamps). This resulted in an altogether 'unnatural' image in terms of inherent light and shade. The artificial effect was counteracted through subsequent editing of the image, which was reconstructed to give the more naturalistic impression of its having been illuminated from one angle rather than two. Only those 'shadows' were retained which were consistent with this interpretation.

It was considered fundamental that the image area representing the mouth should be enhanced so as to give prominence to the lips, teeth, and tongue. With only black and white pixels available, this proved rather difficult at first. However, careful study of the monochrome photographs reproduced in newspapers, confirmed that reasonable effects may be obtained by highlighting black lips with white edges. The teeth, for example, could be represented as a few white pixels with a black outline, while the tongue could be made to appear white upon a black background, representing the oral cavity. Individual static positions for the mouth were determined by scrutinizing that of a subject asked to prepare, as it were, the articulation of various phonemes. These were modelled in an exaggerated fashion initially in order to arrive at an adequate approximation of the phonemes for storage by the computer. 'Fine tuning' of the various representations was done later with 'highlights' and 'shadows' touched in via the pixel editor where necessary.

It was considered fundamental that the image area representing the mouth should be enhanced so as to give prominence to the lips, teeth, and tongue.

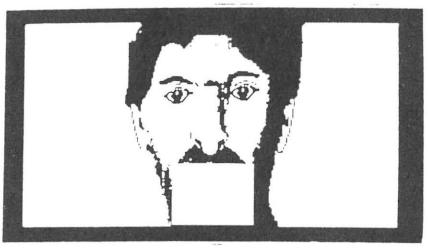


Figure 4
Enhanced facial image with animation window 'cleared'.

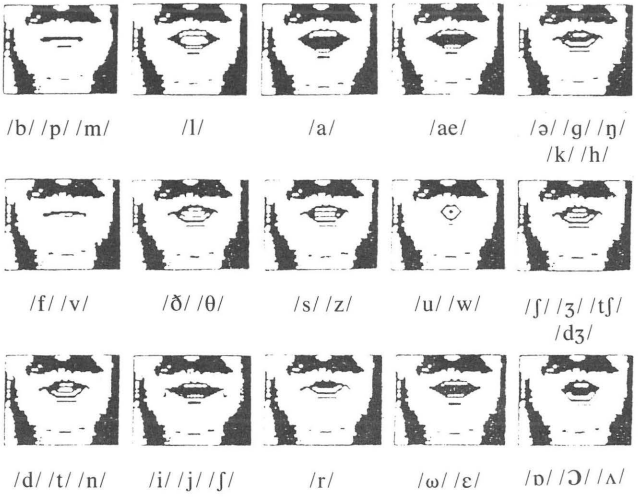


Figure 5
Individual frames representing 15 articulatory positions.

Animation and audio-visual synchrony

Animation of the digital face centers upon changes made to the values stored, and hence the images produced, in a small (72 x 64 pixel) 'window', the relative screen location of which is shown in figure 4. Fifteen static articulatory configurations (of lips, teeth, and tongue) have been predefined, as in figure 5. They are each stored at run time in separate areas of memory. From these locations the 'frames' can be recalled and transferred to the area supporting the graphical window, their insertion simply overwriting the prior resident data and concomitant screen image in less than 10 milliseconds. (The repertoire of fifteen discrete articulatory positions attempts to exploit the intuitive observation that groups of speech sounds, at least as phonetically defined, are visually similar if not identical. English equivalents of the phonetic captions employed are given in Table 1.) The perceptual effect of a change from the graphical realization

of an articulatory position appropriate to the bilabial stop consonant /b/, say, to that representing the vowel /a/, is unquestionably appearance of the syllabic gesture /ba/, motion being inferred between the two positional extremes. (This sequence could of course be construed as /ma/ or /pa/.)

The speed with which changes can be wrought upon the display of this smaller window to the VDU assumes particular importance for the eventual synchrony of such changes with complementary instances of acoustical synthesis. No modification of the instruction set passed to the speech synthesizer can become effective in less than 10 milliseconds, whereas shifting of the screen images can. Thus, even the most rapid of consonant-vowel articulations, as portrayed acoustically by the LSI device at least, can be adequately encapsulated within a synchronous graphical event.

Any instance of speech prepared in a format admissable as input by the LSI synthesizer can be coupled with a synchronous graphical interpretation. Manual scrutiny of the parameter values required to drive the synthesizer reveals the locations of disjuncture within this table of values, corresponding to concomitant acoustical, and hence articulatory change. Since the speech synthesizer receives its instructions via the microcomputer it is necessary only to interleave these with the requisite code for generating the appropriate graphical sequence, each frame in the sequence occupying the window only for so long as is necessary to accompany the speech synthesis to its next point of acoustical departure. The preliminary visual analysis implicated in this set of procedures has in fact been automated. A program has been written which distills the discontinuities from within the synthesis parameters automatically. Determining which of the 15 frames has to be invoked at any point however, remains a manual, or rather auditory exercise for the present.

Text–to–(audio-visual) speech conversion

Anyone familiar with the operation of software-controlled speech synthesizers will appreciate how tedious is the preparation of perceptually adequate utterances.

Table 1
phonetic symbols for transcribing English consonants

p	pie	pea	
t	tie	tea	
k	kye	key	
b	by	bee	
d	dye	D	
g	guy		
m	my	me	ra*m*
n	nigh	knee	ra*n*
ŋ			ra*ng*
f	fie	fee	
v	vie	V	
θ	thigh		
ð	thy	thee	
s	sigh	sea	
z		Z	mi*z*zen
ʃ	shy	she	mi*ss*ion
ʒ			vi*s*ion
l	lie	lee	
w	why	we	
r	rye	re	
j		ye	
h	high	he	

Note also the following:

tʃ	chi(me)	chea(p)
dʒ	ji(ve)	G

phonetic symbols for transcribing English vowels

i	heed	he	bead	heat	keyed
ɪ	hid		bid	hit	kid
eɪ	hayed	hay	bayed	hate	Cade
ɛ	head		bed		
æ	had		bad	hat	cad
ɑ	hard		bard	heart	card
ɒ	hod		bod	hot	cod
ɔ	hawed	haw	bawd		cawed
o	hood				could
oo	hoed	hoe	bode		code
u	who'd	who	booed	hoot	cooed
ə	herd	her	bird	hurt	curd
ʌ	Hudd		bud	hut	cut
aɪ	hide	high	bide	height	
aʊ		how	bowed		cowed
ɔɪ		(a)hoy	Boyd		
ɪə		here	beard		
ɛə		hair	bared		cared
aə	hired	hire			

Hence our own demonstrations of audio-visual synchrony utilizing the LSI device have been confined, thus far, to two examples only, i.e. the phrases: "A bird in the hand is worth two in the bush" and "An apple a day keeps the doctor away". Nonetheless, these efforts were sufficient to establish that the principles are fairly robust. Interestingly, informal observation of the graphical animations in isolation reveals their fragmentary nature. It is the synchronous occurrence of an acoustically dynamic signal which lends coherence to what is then interpreted, perceptually, as a unitary event. This observation complements, in a somewhat contradictory way, those illusory audio-visual demonstrations mentioned earlier. The 'McGurk effect' is itself an example of perceptual coherence being maintained despite apparently *discrepant* information arriving at the senses. In such illusions, it is the visual component of the ensemble which tends to impose its overall perceptual structure. This particular situation is the converse of that noted here, although a formal examination of the degree to which our fragmentary animations may compensate for acoustically impoverished speech may yet bring the two rather divergent observations into line.

Informal observation of the graphical animations in isolation reveals their fragmentary nature. It is the synchronous occurence of an acoustically dynamic signal which lends coherence to what is then interpreted, perceptually, as a unitary event.

To provide for faster entry of speech data to the synthesizer than is customarily possible, LSI have implemented a software facility which admits the input of pseudo-phonetic strings. Such an input sequence is converted automatically into the most appropriate, acoustically context-dependent interpretation, in terms of filter source parameters for the machine, the output function of which remains as previously described. A recent project (Strawbridge, 1986), has resulted in the provision, additionally, of a module for 'translating' standard English orthography into the quasiphonetic form of representation adopted by LSI. Although this latter system currently resides within a mainframe computer, it should prove relatively straightforward to replicate its operation using the BBC micro. We therefore have all of the necessary components at least for realizing a system for the conversion of text to fairly high quality audio-visual speech. In the meantime, however, we have been exploring alternative, though admittedly less elegant vehicles for speech synthesis via the BBC microcomputer, in particular the "Speech!" utility

A recent project has resulted in the provision, additionally, of a module for 'translating' standard English orthography into the quasiphonetic form of representation adopted by LSI.

marketed by Superior Softward Ltd., which is both cheap and readily available.

Careful disassembly of the appropriate constituent program from the "Speech!" package revealed that it would admit the predication of our own graphics routines (also written in assembly language). The unadulterated 'speech' program enables, within reason, acoustical synthesis from orthographical input. Since our own graphical extension is designed to follow, explicitly, the sequence of phonetic events as determined by the host program, the fundamental text-to-speech facility is unimpaired by the modification. Indeed, phonetic misinterpretation of any orthographical string is carried over to the graphical aspect also, the two coexisting programs being synchronously compatible, literally to a fault.

Enhancements ought undoubtedly to follow, but the most significant, and demanding challenge; that of driving the graphics from an analysis of acoustical output remains to be confronted. At the very least we have an extremely cost effective medium with which to evaluate any theoretical steps taken in that particular direction in the future.

Acknowledgement

The guidance and encouragement of Dr. Q. Summerfield of the MRC Institute of Hearing Research, Nottingham, is gratefully acknowledged.

About the authors

Darryl Storey gained his B.Sc. in Mechanical Engineering from Lanchester Polytechnic, Coventry in 1982. He then spent a period in industry as a Development Engineer and a year in local government as a Computer Programmer. Since 1984 he has held the post of Experimental Officer in the Department of Computer Studies, Loughborough University of Technology.

Martin Roberts hold a Ph.D. in Experimental Psychology from Nottingham University. His research interests are in distinguishing sensory from cognitive influences upon speech perception and the application of audio-visual experimental techniques to that end. He is currently a Research Assistant in the Department of Computer Studies, Loughborough University of Technology.

References

Abry, C. and Broe, L. J. 1986. Laws for lips. *Speech Communication,* 5, 97–104.

Boston, D. W. 1973. Synthetic facial communication. *British Journal of Audiology,* 7, 95–101.

Brooke, N. M. 1982. Video speech synthesis for speech perception experiments. *Journal of the Acoustical Society of America,* 71, S77(A).

Brooke, N. M. and Summerfield, Q. 1983. Analysis, synthesis, and perception of visible articulatory movements. *Journal of Phonetics,* 11, 63–76.

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

Erber, N. P. and De Filippo, C. L. 1978. Voice/mouth synthesis and tactile/visual perception of pa, ba, ma. *Journal of Acoustical Society of America,* 64, 4, 1015–1019.

Levinson, S. E. and Schmidt, C. E. 1983. Adaptive computation of articulatory parameters from the speech signal. *Journal of the Acoustical Society of America,* 74, 4, 1145–1154.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P. and Studdert-Kennedy, M. 1967. Perception of the speech code. *Psychological Review,* 74, 6, 431–461.

McGurk, H. and Macdonald, J. 1976. Hearing lips and seeing voices: A new illusion. *Nature,* London, 746–748.

Montgomery, A. A. 1978. Generation and evaluation of synthetic facial images for lip-reading. *Paper presented at the annual meeting of the American Speech and Hearing Association,* November, 1978.

Montgomery, A. A. 1983. The search for invariant visible cues in lipreading. *Journal of the Acoustical Society of America,* 73, S15.

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, 6, 2134–2144.

Parke, F. I. 1975. A model for human faces that allows speech-synchronized animation. *Computers and Graphics,* 1, 3–4.

Parke, F. I. 1982. Parameterized models for facial animation. *IEEE Computer Graphics and Applications,* 2, 9, 61–68.

Pearce, A., Wyvill, B., Wyvill, G. and Hill, D. 1986. Speech and expression: a computer solution to face animation. In: *Proceedings of the graphics interface '86 conference*, Vancouver, Canada, May 26th–30th.

Perkell, J. S. 1986. Coarticulation strategies: preliminary implications of a detailed analysis of lower lip protrusion movements. *Speech Communication,* 5, 47–68.

Ramachandran, V. S. and Anstis, S. M. 1986. The perception of apparent motion. *Scientific American,* 254, 6, 80–87.

Reisberg, D., McLean, J., and Goldfield, A. 1987. Easy to hear but hard to understand: A lip-reading advantage with intact auditory stimuli. In: B. Dodd and R. Campbell, (Eds.). *Hearing by Eye: Experimental Studies in the Psychology of Lipreading.* London, Lawrence Erlbaum Associates.

Sondhi, M. M. and Resnick, J. R. 1983. The inverse problem for the vocal tract: numerical methods, acoustical experiments, and speech synthesis.*Journal of the Acoustical Society of America,* 73, 3, 985–1002.

Strawbridge, K. P. 1986. The development of a grapheme-to-phoneme translation algorithm in Prolog for use with the LSI Phonetic Synthesizer. *Unpublished M.Sc. dissertation,* Loughborough University of Technology.

Summerfield, Q. 1979. Use of visual information for phonetic processing. *Phonetica,* 36, 314–331.

Summerfield, Q. 1987. Preliminaries to a comprehensive account of audio-visual speech perception. In: B. Dodd and R. Campbell, (Eds.), *Hearing by Eye: Experimental Studies in the Psychology of Lipreading.* London, Lawrence Erlbaum Associates.

Speaking From Two Sides of the Mouth

Speaking from Two Sides of the Mouth

Roger E. Graves and Susan M. Potter

Department of Psychology,
University of Victoria,
Victoria, B.C., Canada,
V8W 2Y2

Visible Language XXII, 1
Roger E. Graves and Susan M. Potter, pp. 128-137

Differences while speaking from the two sides of the mouth are both visible and audible. Careful observation has shown that the right side of the mouth typically opens wider and moves more during speech. This visible asymmetry reveals the underlying physiology in which expression of speech is controlled primarily by the left side of the brain. Since the left side of the brain has better control of the right side mouth muscles, an asymmetry favoring the activity of the muscles of the right side results during articulation of speech sounds. In contrast, more equal activity from the left side of the mouth can be seen during emotional expression, prosodic expression, and singing which reveals a greater role of the right side of the brain during these latter types of expression. There are also audible manifestations of the physiological asymmetries. In a new study, subjects were required to speak from only one side of the mouth. Better quality of articulation was audible from the right side for most subjects.

"My left hand never learnt what my right one's been doing" (Treat, 1943).

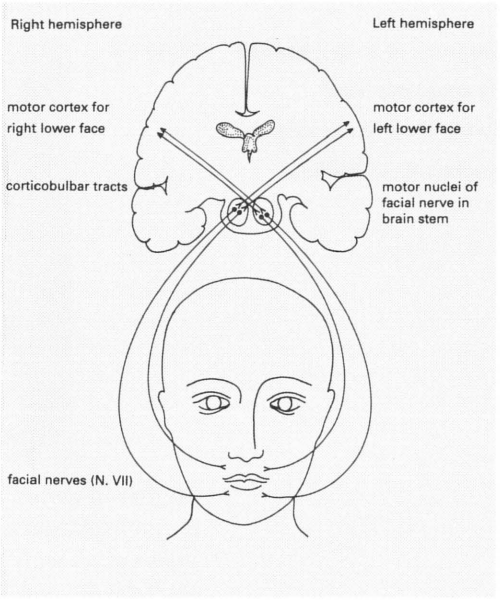
Most of us are quite familiar with the fact that we have a "preferred" or "dominant" hand which is more skilled for many actions, such as writing. Less familiar but equally true is that the non-preferred hand is more skilled for certain other actions. Try, for example, to tie your shoelace backwards.You will probably find that neither the left nor the right hand seems to know what the other one has learnt. None of this will be surprising to anyone save the few ambidexters among us, after all we do have two hands and use them differently. We have only one mouth, however, and we tend to view speech as a unitary act employing a single central organ which has no essential left versus right side difference. Thus, we may be somewhat surprised to discover that the left side of the mouth may also not have learnt what the right side's been doing, and vice versa. Now neither the hand nor the mouth actually has much learning ability, the site of the learning and memory is usually considered to be the cerebral cortex of the brain. Furthermore, the brain is so constructed that skilled control of the right hand depends mainly on motor control areas on the left side of the brain, while control of the left hand depends on the right side of the brain. The superior right hand performance of right handers is thus thought to reflect superior praxic skill representation in left brain motor control areas.

We have only one mouth, however, and we tend to view speech as a unitary act employing a single central organ which has no essential left versus right side difference. Thus, we may be somewhat surprised to discover that the left side of the mouth may also not have learnt what the right side's been doing, and vise versa.

The situation is actually much the same for speech and the two sides of the mouth as for skilled actions and the two hands. The learning and memory for speech, at least as far as words, grammar, and the control of speech articulation is concerned, is for most of us strongly dependent on one side of the brain, namely the left cerebral hemisphere. As with the hands, the right side mouth muscles are mainly controlled by the left half of the brain and vice versa (see figure 1). From this anatomical perspective it is understandable that the right side of the mouth, which is controlled by the side of the brain with the superior speech control ability, might behave differently than the left side of the mouth during speech articulation. Furthermore, the *left* side of the mouth is

controlled by that side of the brain which is superior for the perception and expression of emotional and prosodic aspects of speech (Ley & Bryden, 1981; Ross & Mesulam, 1979). (Prosody is the rising and falling pitch pattern which, for example, conveys information about whether a particular word or string of words is a statement, command, question, etc.) Thus, the left side of the mouth may behave differently than the right side of the mouth during expression of emotion and prosody.

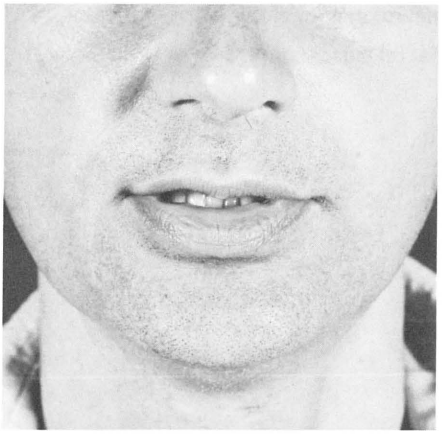
Figure 1 **Diagrammatic representation of the neural control of the lower facial musculature.**



By 1980 a number of studies had discovered that emotional expressions, at least of certain types, were indeed expressed more strongly on the left side of the face (Borod and Caron, 1980; Campbell, 1978; Thompson, 1985). The first investigators looking for mouth asymmetry during speech then confirmed their prediction that the right side of the mouth would open more than the left side during speech articulation (Graves, Landis, & Goodglass, 1982). These studies observed that 150 of the 196 subjects had greater right side mouth opening during speech. Figure 2 shows the photographed asymmetry of one subject. Informal real time observation of speakers on TV or in person can sometimes also reveal an asymmetry, and more often than not it is the right

side which shows the greater opening (Hager & van Gelder, 1985). Anyone tempted to make such observations at the next cocktail party should be forewarned that people usually become quite uncomfortable when they notice someone staring at their mouth, and their discomfort does not decrease if you explain that you are looking at their asymmetries. There appears to be a social taboo against staring at your conversant's mouth, and this may have inhibited earlier discovery of mouth asymmetry during speech.

Figure 2

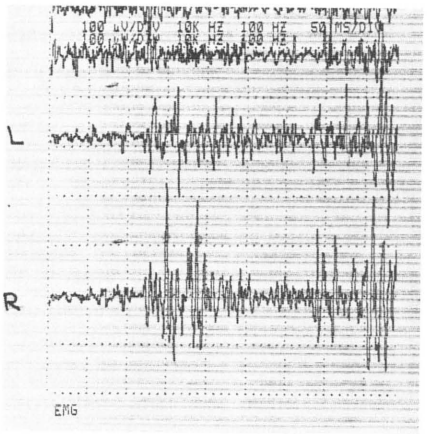


Photograph of a right handed male speaking "Pea" showing greater opening of the right side of the lips.

Studies of mouth asymmetry have shown that, when normal subjects are speaking, the right side of the mouth *opens more widely* (Graves et al., 1982; Graves, 1983; Wyler et al., 1987). This asymmetry was observed using still frame photographic techniques at a time about 50 milliseconds from the initial lip opening of a bilabial consonant (e.g., "B" or "M"). Landis, employing a computerized tracking system (Graves et al., 1982), also showed that the right side of the mouth moves more overall during continuous speech. Preliminary observations with a few subjects (Graves, Landis, & Simpson, 1985) have shown that asymmetry in muscle activity can also be seen in electromyographic recordings from surface electrodes placed around the mouth (see figure 3). The asymmetry in the electrical correlates of muscle activity appeared to be most prominent during fast changes in lip configuration such as with bilabial consonants, especially those in the middle of words.

Figure 3

EMG recordings from the left and right side lip musculature of a right handed male speaking "Bobbing".



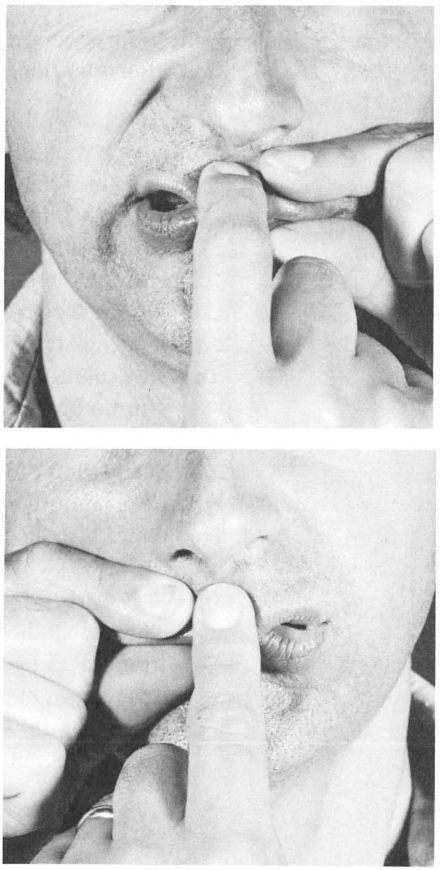
Greater right side mouth opening or mouth movement during speech has been seen in all groups of normal subjects reported to date, these included both left and right handers and both men and women. The incidence of greater right side opening varied depending on the recording technique and the subjects' speech task. The highest incidence (80-90%) was observed with photographic recording and a word list task which discouraged visual, emotional, and prosodic involvement. Left handers in general also typically show greater right sided mouth activity during speech (Graves et al., 1982). This is consistent with the evidence that the percentage of left handers having left hemisphere control of speech is much higher than the percentage having right hemisphere control (Rasmussen & Milner, 1977).

The initial studies also indicated that asymmetry in mouth opening during speech reflected more than just the left hemisphere control of articulation of speech sounds. This and subsequent work has shown a reduction in the strength of the greater right side mouth opening when subjects are describing visual situations of emotional significance (Graves et al., 1982; Wyler, Graves, & Landis, 1987). This effect does not typically involve a reversal to greater left side activity, but rather involves more equal levels of activity on the two sides of the mouth. More equal levels of mouth muscle activity on the two sides has been assumed to reflect participation of both sides of the brain in the expressive control.

Equal level of activity need not mean identical activity, of course, and the two sides of the mouth (and brain) may be expressing qualitatively different things. Smiles may produce a reversal to greater left sided activity (Wyler et al., 1987), although this has been an inconsistent observation (Campbell, in press).

All of the studies of mouth asymmetry which have so far been mentioned have looked for differences in the movements of the two sides of the mouth when people were speaking naturally. The latest study took a different approach. Subjects were asked to speak using one and then the other side of the mouth and differences in the quality of articulation were looked for. Sixteen right handed male university students took part in this study. The task was to say quickly two tongue twisters, "Peter Piper picked a peck of pickled peppers", and "Bobbling babies blowing bubbles burble and babble". These were chosen because they contain bilabial phonemes which require precise coordination of lip movements. Each subject said each tongue twister first from one side, then from the other side of the mouth. Half the subjects began with the left, half with the right side. The method used for speaking out of the right side of the mouth was for the subject to pinch the lips together in the midline using the tips of the thumb and index finger of his right hand and to pinch the lips on the left side of the mouth together using the thumb and index finger of the left hand placed parallel to the lips. For speaking out of the left side, the positions of the hands were changed accordingly (see figure 4). The tape recorded speech samples were judged by a listener who did not know which side of the mouth was which. The result was that, for the 16 subjects, quality of articulation was judged as better from the right side for 10 subjects, as better from the left side for 2 subjects, and as not discernably different for 4 subjects. This simple procedure thus showed that most of these right handed male subjects spoke better out of the right side of the mouth ($t(15) = 2.65$, $p<.01$, one-tail). Some subjects exhibited a virtual hemiparalysis of the left side lips when attempting this task.

Figure 4



Photographs of a right handed male speaking from the two sides of the mouth illustrating the technique used to restrain one side of the lips.

An important implication of the studies with normal subjects is that mouth asymmetry can be used as a tool to reveal underlying distinct components of the organization and control of expression, as well as to reveal the relative involvement of the two hemispheres of the brain with these functional components. One study with patients (Graves & Landis, 1985) employed this tool in order to understand better why some types of speech are more impaired while other types of speech are more preserved following brain damage. The results indicated that spontaneous speech, which is typically the most impaired following left hemisphere damage, showed greater right side mouth activity and thus is strongly dependent on left hemisphere control. Singing and serial speech (counting, for example), which are typically less impaired following left hemisphere damage did not

show greater right side mouth activity and thus are less dependent on left hemisphere control.

Apart from the potential of mouth asymmetry as a research tool, there are possibilities for therapeutic applications. Training of impaired speakers, including deaf children, stutterers, and brain damaged patients, could conceivably be facilitated by providing feedback concerning the relative activity of the two sides of the mouth. The results with aphasic patients suggest that feedback training to increase the relative amount of right side mouth muscle activity might assist the patient in employing the optimal (left hemisphere) system and inhibiting interfering (right hemisphere?) systems. Investigation of whether attention to one or the other side of the mouth would (or does) assist lipreading could also be considered.

Acknowledgement

This research was supported in part by Grant A-1021 from the Canadian Natural Sciences and Engineering Research Council. We thank Dr. Ruth Campbell for her many helpful suggestions which led to substantial improvements in this paper.

About the Authors

Ms. Potter was born in Middlesborough, England and moved to Canada. She attended Queen's University in Kingston, Ontario and then the University of Victoria, Victoria, B.C. where she received the B. Sc. degree in 1986. Her Honor's thesis concerned interhemispheric transfer of different types of information in normal adults. She is currently a doctoral student in Clinical Psychology at McGill University in Montreal, Quebec.

Dr. Graves received the B.S. (Electrical Engineering) and Ph.D. (Psychology) degrees from the Massachusetts Institute of Technology. He received training in Clinical Neuropsychology during a Postdoctoral Fellowship at Sunnaas Hospital in Norway and also during five years of association with Dr. Harold Goodglass in the Aphasia Research Center at the Boston Veterans Administration Medical Center. Since 1980 he has been on the faculty of the Department of Psychology of the University of Victoria.

References

Borod, J.C. & Caron, H.S. 1980. Facedness and emotion related to lateral dominance, sex, and expression type. *Neuropsychologia,* 18, 237–241.

Campbell, R. 1980. Asymmetries in the interpretation and expression of a posed expression. *Cortex,*14, 327–342.

Campbell, R. in press. Asymmetries of facial action; some facts and fancies of normal face movement. In R. Bruyer (Ed.) *The neuropsychology of face perception and facial expression.* Hillsdale, N.J.: Lawrence Erlbaum.

Graves, R. 1983. Mouth asymmetry, dichotic ear advantage and tachistoscopic visual field advantage as measures of language lateralization. *Neuropsychologia,* 21, 641–649.

Graves, R. & Landis, T. 1985. Hemispheric control of speech expression in aphasia. A mouth asymmetry study. *Archives of Neurology,* 42, 249–251.

Graves, R., Landls, T., & Goodglass, H. 1982. Mouth asymmetry during spontaneous speech. *Neuropsychologia,* 20, 371–381.

Graves, R., Landis, T., & Simpson, C. 1985. On the interpretation of mouth asymmetry. *Neuropsychologia,* 23, 121–122.

Hager, J.C. & Van Gelder, R.S. 1985. Asymmetry of speech actions. *Neuropsychologia,* 23, 119–120.

Ley, R.G., & Bryden, M.P. 1981. Consciousness, emotion, and the right hemisphere. *In Aspects of Consciousness,* (pp. 215–240), Vol. 2. G. Underwood & R. Stevens (Eds.). New York: Academic Press.

Rasmussen, T., & Milner, B. 1977. The role of early left brain injury in determining lateralization of cerebral speech functions. *Annals of the New York Academy of Sciences,* 299, 355–367.

Ross, E.D., & Mesulam, M. 1979. Dominant language functions of the right hemisphere? *Archives of Neurology,* 36, 144–148.

Thompson, J.K. 1985. Right brain, left brain; left face, right face: hemisphericity and the expression of facial emotion. *Cortex,* 21, 281–300.

Treat, L. 1943. *0 as in omen* (p. 23). New York: Sloan and Pearce.

Wyler, F., Graves, R., & Landis, T. 1987. Cognitive task influence of relative hemispheric motor control: mouth asymmetry and lateral eye movements. *Journal of Clinical and Experimental Neuropsychology,* 9, 105–116.

Visible Language Advisors, Research Interests, and Upcoming Issues

Sharon Helmer Poggenpohl

Abstract

New and returning Advisory Board members are introduced along with their research interests and their relationship to the Journal. Board members suggested areas of investigation for the future, many of which relate to the problems and opportunities of new technology.

Visible Language is a journal that bridges the sciences and the humanities. It takes the position that both a poem and its typographic presentation and the testing of a scientific hypothesis concerning reading or perception are equally important to understanding the use and development of visible language. The Journal is interdisciplinary by both design and necessity. As such, it is a publication with a mind of its own. No one mind can encompass in depth all the Journal's concerns, consequently a carefully orchestrated interdisciplinary advisory board keeps a check on ideas worth pursuing and the quality of both the pursuers and the pursuit.

Visible Language is a community of scholars. Readers should know who the very important members of the Advisory Board are. They are international in scope and have significant achievement in areas important to the Journal. It is my pleasure, as editor, to introduce the Advisory Board. Many of them are research fellows. All have published and lectured broadly. Collectively, they have received many honors. In presenting their brief biographies, I will focus on their interests, their relationship to *Visible Language,* their work in progress, and in some cases, I will quote them directly.

New additions to the board

Peter Bradford is both a practicing graphic designer and an educator. His New York based firm has done award winning work in the areas of corporate identification, packaging, book and magazine design, exhibits and signage, and educational and institutional programs.

I asked Peter to comment on his interest in visible language. He gave me the following statement.

"It is faintly ludicrous to presume that designers will inevitably enjoy an enlarged role in an 'informational' society. They just don't seem much interested in the essential skills. Content analysis, logic, analogism, human behavior, and articulation are still too rarely taught in design curriculums, and too rarely nurtured in design practice. So, as designers continue to trifle with type history, styling, and other private esoterica, the immense issues of our lives, such as the national debt impacts, rain-forest depletion, and moral dilemmas like abortion, remain in limbo – unvisualized, uncommunicated, limbo.

I once took a bus ride in Mexico, in a quite modern bus driven by what appeared to be a fifteen-year-old boy. My Lord, he loved that bus. He ran it with all the lights on, inside and out, in bright sunlight. He constantly pushed and pulled every lever and button the bus had, just to watch them all work. His preoccupation with the vehicle was feverish. He especially loved to shift. Up and down, up and down through every gear at the slightest opportunity. It was a four-hour trip and I went crazy, everybody went crazy, lurching forward and back whenever the mood took him to explore another gear. While he seemed to know and care a lot about buses, he certainly wasn't much interested in conveying anybody comfortably.

Given the typical designer's passion for his vehicles, is he missing the point of his job, too?"

Dick Higgins is a multi-media artist and poet. "I find I never feel quite complete unless I'm doing all the arts – visual, musical, and literary. I guess that's why I developed the term 'inter-media', to cover my works that fall conceptually between these." His work has been published, performed, and shown internationally. His most recent book is *Pattern poems: guide to an unknown literature* (1987). Currently, he is translating Giordano Bruno's *De imaginum, signorium et idearium compositione* (1591). Dick is the first poet/artist to serve on the board. He guest edited Volume XX, no. 1 (Winter 1986), a special issue on Pattern Poetry.

Kenneth M. Morris is an expert on language simplification. He is President of Siegel & Gale, New York, as such, he oversees the firm's administrative and strategic planning as well as its international expansion. He formerly headed the firm's Simplified Communications Group. During that time, Dr. Morris supervised simplification projects for the U.S. Internal Revenue Service, and major insurance and financial service companies.

Before moving into the corporate arena, Dr. Morris was on the faculty of the English Department at John Jay College of Criminal Justice and also taught

plain English writing at Carnegie Mellon's Document Design Center. His anthology, *Literature in Bureaucracy* (Avery, 1979), explores the communications and managerial problems faced by large corporations.

David R. Olson is Professor of Applied Psychology at The Ontario Institute for Studies in Education and Director of the McLuhan Program in Culture and Technology at the University of Toronto. He is author of the entry on Writing in the latest edition of *Encyclopedia Britannica*; co-editor of *Literacy, language, and learning: The nature and consequence of reading and writing* (1985); and is currently at work on a book on literacy with the working title, *The world on paper*. He is currently President of the Canadian Psychological Association.

David's most recent contribution to the Journal is "Interpreting Texts and Interpreting Nature: The Effects of Literacy and Epistemology", Volume XX, no. 3 (Summer 1986).

Gerard Unger is a graphic designer in Bussum, The Netherlands where he teaches at the Rietveld Academy and has a design practice specializing in type design. Much of his type design (Demos, Praxis, Hollander, Flora, Swift, and Cyrano) has been for the German firm Dr. Ing. Rudolf Hell. More recently, he has been collaborating with the American firm, Bitstream, for whom he designed Amerigo, a digital face for desktop publishing. He has published two monographs: *Text About Text*, and *Typography: Principles and Applications.*

Dietmar R. Winkler combines an active design and consulting practice with educational institutions, organizations, and publishers with the teaching of professional subjects in design, typography, and communication theory. He is Director of Graduate Studies in Visual Design at Southeastern Massachusetts University. He writes critically on educational and professional design issues.

"I am interested in evolving programmatic and systemic approaches to data and information based design as well as the solving of culturally complex design and communication problems. My personal goal is to help move Visual Design from a vocation to a responsible profession by urging expansion of the traditional visual perception research base to include human and social behavioral as well as language and communication research issues."

Continuing board members

Colin Banks is a partner in Banks and Miles, a design firm in London. In addition to large corporate projects such as British Telecom and typeface design and development for the telephone directory in the UK, Banks and Miles serve as graphic advisors to Her Majesty's Stationery Office. They are specialists in security printing and have designed banknotes and stamps for a number of Middle Eastern countries. The firm has a special interest in 'Social Communication'. They have written a book (1979) and assembled a travelling exhibit under the same title. Public information is a particular focus of the firm.

Naomi S. Baron is a linguist with special interest in language acquisition, language change, the theory of signs, technology, and computers. Her publications include *Computers: A Guide for the Perplexed* (1986); *Speech Writing and Sign* (1981); and *Language Acquisition and Historical Change* (1977).

She is currently writing a book on children's acquisition of speech and literacy. She is Professor of Language and Foreign Studies and Associate Dean for Undergraduate Affairs at The American University, Washington, D.C.

Fernand Baudin is Vice-President of the Association Typographique Internationale (ATypI). He was a member of the former Centre de Recherches Typographiques (CERT) under the leadership of Charles Peignot and M. Georges Bonnin, and is currently a member of a committee working within the national French Ministry of Education for the rehabilitation of handwriting and the introduction of 'typography' in the classroom. His most recent book *La typographie au tableau noir* (Paris, 1984) will be reissued in English this year (London: Lund Humphries).

Pieter Brattinga is a Dutch graphic and exhibition designer. He is a senior partner in Form Mediation International, Amsterdam. He has been involved intermittently with design education in the United States and The Netherlands. Books have been published both by and about him. He characterizes his relationship to the Journal as follows:

"In the past I have been...a scout; suggesting authors, experiments, articles which were brought to my attention or which I had heard about. Sometimes I asked [the editor] to receive or meet people who's ideas I found worthwhile. Sometimes he [the editor] was surprised, sometimes aghast, and sometimes very enthusiastic. I am a regular visitor to Asia and North America and hear and see subjects which are of interest to our disciplines."

Gunnlauger SE Briem is a designer in London. He has of late taken an active interest in the introduction of italic handwriting in his native Iceland. His most recent contribution to this Journal was a special Calligraphy Issue, Volume XVII no. 1 (Winter 1983). "With the Second Hand Press, he makes impossible ideas into futile reality."

James Hartley is Reader and Head of the Department of Psychology at Keele University, England. He first became involved with typographic research when he joined forces with the designer Peter Burnhill to carry out joint research between 1970-1980. Two of the many outcomes of this work were the special issue of *Visible Language* entitled The Spatial Arrangement of Text (Volume XV, no. 1) and the book *Designing Instructional Text.* Currently, Dr. Hartley is involved in research on academic writing and the design of effective medical audiotapes.

Dominic Massaro is a Professor of Psychology at the University of California at Santa Cruz. His work focuses on information processing and visible and audible language perception. Prof. Massaro is currently completing *Experimental Psychology: The Study of Mental Processes* for Oxford University Press. He is represented by an article in this issue.

Alexander Nesbitt is an educator and calligrapher. He is Professor Emeritus at Southeastern Massachusetts University. In addition to numerous articles concerning typography and graphics in design journals, he wrote *The History and Technique of Lettering,* a classic book on the subject. With his wife, he

conducts the Third & Elm Press in Newport, which produces small books and printed matter. His life-long interest is the written and printed word.

Thomas Ockerse is Chair of the Rhode Island School of Design's Division of Design. He is known for his theoretical developments in 'semiotics' as this applies to visual language, design practice, and design education. These theoretical explorations find further expression in his visual/linguistic experiments in 'concrete poetry' and 'bookworks'. His most recent article is De-Sign/Super-Sign (Semiotica 52-3/4). He is an advocate for substance and quality in design education nationally and internationally.

To temper his work in theory and education, Tom is a consultant in the areas of communication and design strategies. He is both the practical guide and typographic critic for the graduate student designed issues of this journal.

Charles L. Owen is Professor of Design at the Institute of Design, an academic department of the Illinois Institute of Technology (IIT) in Chicago. He directs the Design Processes Laboratory and the Product Design graduate program. His work spans the fields of product design, computer-supported design, design methodology and computer graphics – teaching, conducting research, and consulting. His current research is directed toward structured planning techniques, computer-supported diagramming processes, and the development of design-support systems employing artificial intelligence techniques.

Sharon Helmer Poggenpohl is a graphic designer engaged in both practice and education. Poggenpohl Design is particularly involved with nonprofit clients who have complex communication problems that address the public agenda. She is an Adjunct Professor in the Graphic Design Department at the Rhode Island School of Design where she teaches in the graduate program. She is currently investigating alternative methods to organize information and text. As editor of *Visible Language*, she functions as a generalist whose real passion is in finding connections between ideas. Computer Graphics: Graphic Design, Volume XIX, no. 2 (Spring 1985), was the last special issue she edited. She became general editor with Volume XXI.

"*Visible Language* is a concept to be delineated. I expect to move the markers out a bit further than my editorial predecessor; I will include word/image relationships. This is an interesting time for visible language as the cultural need for more streamlined information intersects with new technology. It is a time to question our habits and traditions and to look for appropriate performance characteristics. Without loosing its research focus, the Journal should begin to bridge the gap between research and theory and communication practice."

Denise Schmandt-Besserat is Associate Professor of Art and Middle Eastern Studies at the University of Texas at Austin. She is presently studying the origins of writing. Her work is based on archaeological collections of clay counters-tokens-which were the direct precursor of writing in the Middle East. She has published many articles on early writing. Her last contribution to this journal was "Tokens: Facts and Interpretation ", Volume XX, No. 3 (Summer 1986).

For a late 1990 or early 1991 release, she will guest edit a special issue for *Visible Language* on "The First Scripts". These include Sumerian, Proto-Elamite, Ancient Egyptian, Indus Valley, Linear A and B, Luwian (Hittite), Etruscan, and Mayan. She plans several comparative threads which will bring into relief the various differences and similarities among the scripts. These are: the date and circumstance of the script invention, a description of the first documents and their function, the present state of decipherment, and the pictographic or phonetic nature of the sign.

Michael Twyman is Professor of Typography & Graphic Communication at the University of Reading, England, where he has been teaching for thirty years. His major research interests lie in the fields of the history of printing and the theory of typography. His books include *Printing 1770-1970* and *Lithography 1800-1850*. He has a special interest in typographic education and has been Chairman of the Working Party on Typographic Teaching of the Education Committee of ATypI (Association Typographique Internationale). His most recent contribution to the Journal was in the special issue Graphic Design: Computer Graphics, Volume XIX, no. 2.

Richard L. Venezky is Unidel Professor of Educational Studies at the University of Delaware, with a joint appointment in computer and information sciences. He guest edited a special issue on Literacy and Competency, Volume XVI, no. 2. He identifies his professional interests as follows.

"If there is a center or pivot around which my work revolves, it is information – its reception, storage/retrieval, and communication. One dimension of this space involves writing systems and their psycholinguistic properties. Another is literacy – its history, its acquisition, and its social/political consequences. Yet another dimension involves lexicography, particularly electronic access to dictionaries and other reference materials. From an interest in electronic assistance to dictionary making, I have moved to the problems of structuring and representing complex information sets, such as science skills or even the total knowledge base of the world as reflected in a multi-volume encyclopedia. (On the periphery, like faint twinkling stars, are art and technology and computer-assisted learning, with tenuous links to the center.)"

Dirk Wendt is an experimental psychologist, working as a Professor of Psychology at the University of Kiel, West Germany. He has done work on the effectiveness of typographic variables in communication. Most of these studies have been published in this Journal.

Patricia Wright is a cognitive psychologist interested in the ways that the design of visual information influence its usability. She is a member of the scientific staff of the Medical Research Council's Applied Psychology Unit in Cambridge, England. Her empirical research concerns 'technical' rather than 'leisure' materials, but has covered topics as diverse as application forms, tables and graphs, procedural instructions, electronic texts and location maps for use inside buildings such as hospitals. Her most recent article in *Visible Language* was "Investigating Referee Requirements in an Electronic Medium", Volume XVIII, no. 2.

Hermann Zapf is a renown type designer who has created more than 175 designs including Optima and Palatino. Recently he designed a typeface specifically for computer generation "World Book Modern", for the encyclopedia *World Book*. He is currently designing a type-face for the American Mathematical Society to accomodate advanced mathematical formulas. This computer generated face is being developed in conjunction with Stanford University's Computer Science Department.

Prof. Zapf lives in Darmstadt, Germany. He describes himself as a type designer, calligrapher, book designer, teacher, and author. The Society of Typographic Arts (Chicago) recently published a monograph celebrating his work titled *Hermann Zapf and His Design Philosophy.*

Research directions

The Journal both documents completed research and initiates questions and new lines of inquiry. It is in this last capacity that the Advisory Board was asked to recommend content areas that the Journal should explore. Fully half of the responses related to new technology. For example, Hermann Zapf called for an exhaustive look at the problems of low resolution digital type design.

Ken Morris called for an examination of the merging of language and design in electronic formats. Design tools are accessible to the wordsmith but the visual sensibility and the control of scale, space, and emphasis (to name a few visual variables), remains invisible. Default programs are dull or are viewed as a challenge – an opportunity to beat the system. The real danger is the technical finish; this illusory appearance often masks an ill-formed visual result that seems perversely credible. Typographic literacy begs for a definition – a realistic definition. Desktop publishing underscores the need for fundamental understanding of the principles of successful visible language. This goes beyond legibility performance or the identification of simple functional thresholds to a more complete understanding of visible language as a system. Pat Wright underscored the need to understand *principles* of visible language.

Another broad area of investigation was the impact of technology on how we communicate. Dick Venezky made the following observation.

"...video disk, CD-ROM, and other marvels of the current electronic age allow quick easy access to vast stores of information. A single CD-ROM, for example, could store an encyclopedia, dictionary, thesaurus, and almanac, with extensive indexing. The effective use of his medium, however, requires effective knowledge representation, that is, visualization schemes that can convey the structures of the knowledge spaces and subspaces more quickly and adequately than outlines, lists, or texts alone. Biologists are experimenting with 3-D, color representation techniques for molecular structures, as are neurologists for the activities of the cerebral cortex. These visual languages, based on 3-D, color graphics, are a subject for *Visible Language* to explore, as are the flatter representations of knowledge, such as tree-diagrams, flow charts, and semantic nets."