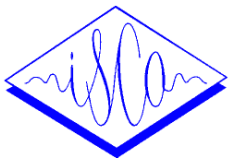

ETRW on Speech Technology
in Language Learning (STiLL)
Marholmen, Sweden
May 24-27, 1998

# Visible Speech and Its Potential Value for Speech Training for Hearing-Impaired Perceivers

*Dominic W. Massaro and Michael M. Cohen*

Department of Psychology University of California,
Santa Cruz Santa Cruz, CA 95064 U.S.A.
massaro@fuzzy.ucsc.edu mmcohen@ranx.ucsc.edu

## Abstract

A theoretical framework, with much research support, is presented as a basis for the use of technology in language learning. According to the FLMP, well-learned patterns are recognized in accordance with a general algorithm, regardless of the modality or particular nature of the patterns. Multiple continuously-valued sources of information are evaluated, integrated, and matched against prototype descriptions in memory, and an identification decision is made on the basis of the relative goodness of match of the stimulus information with the relevant prototype descriptions. Three important properties are 1) the sources are optimally combined, 2) the sources are complementary, and 3) visible speech is a robust contribution to speech perception. The technology developed in our research can be leveraged within our theoretical framework to provide a novel and potentially productive pedagogy for language learning.

## 1. Introduction

It is now common knowledge that there is valuable and effective information afforded by a view of the speaker's face in speech perception and recognition, and understanding. Visible speech is particularly effective when the auditory speech is degraded, because of noise, bandwidth filtering, or hearing-impairment. During the last 15 years, our empirical and theoretical research has evolved a strikingly orderly and parsimonious account of speech perception by ear and eye. This account presents an optimistic framework for language training for hearing-impaired individuals. Our experience is proving the value of Kurt Lewin's dictum that there is nothing so challenging as a practical problem and nothing so practical as a good theory.

## 2. The Fuzzy Logical Model of Perception (FLMP)

We are only beginning to test our theoretical framework in the arena of language training [1]. In this paper, we articulate the theoretical framework, provide reasons why multimodal input is ideal for speech perception, and illustrate how it can serve as the theoretical foundation for language training. In contrast to the majority view that speech and language processing are somehow unique and special, we have established how they can be envisioned as an instance of a general capability of pattern recognition [2,3].

Within our framework of the FLMP, speech perception is robust because there are usually multiple sources of information that the perceiver evaluates and integrates to achieve perceptual recognition. The central assumptions to the model, illustrated in Figure 1, are 1) each source of information is independently evaluated to give a continuous degree to which that source supports the relevant alternatives, 2) the sources are integrated to provide an overall amount of support for each alternative, and 3) perceptual identification follows the relative amount of support among the alternatives. Independent evaluation, multiplicative integration and relative decision predicts that the combination of two imperfect sources of information yield better performance than would be possible using either source by itself.

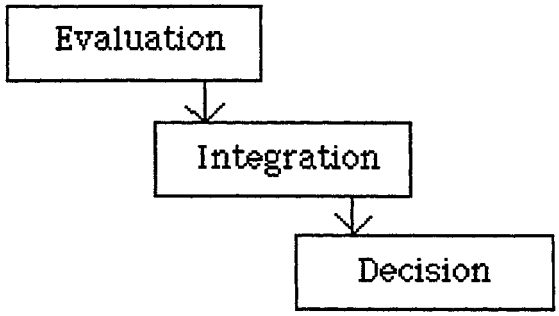


*Figure 1. Three operations assumed by the FLMP.*

We now describe briefly three important aspects of speech processing that provide an optimistic view of the eventual payoff of the use of technology in language training. We end with a description of potential applications of our visible speech technology.

## 3. Three Touchstones for Speech Training

The FLMP provides an optimal algorithm for integrating several sources of information in pattern recognition, learning, and judgment [2,3]. Optimal integration requires that all sources contribute to a decision but that more ambiguous sources be given less of a say in the decision. It has been proven that the FLMP has an op-

timal integration rule. This optimal integration rule was developed over two centuries ago by a Nonconformist minister, Thomas Bayes. Bayes theorem gives the optimal (the most accurate) solution for determining the probability of an event given several pieces of evidence.

### 3.1 Optimal Integration

To illustrate the FLMP's predictions, we present some typical results from Erber [4]. Severely hearing-impaired children identified auditory, visual, and bimodal syllables. The experimental test consisted of a videotape of the eight consonants /b, d, g, k, m, n, p, t/ spoken in a bisyllabic context /aCa/. The children used their hearing-assistive devices during the test. The top row and middle column of Figure 2 gives the confusion matrices under the auditory, visual, and bimodal conditions. An important outcome for our purposes is the performance gain the children show in the bimodal condition relative to either of the unimodal conditions. This outcome reflects the synergy of multiple modalities in speech perception: two ambiguous sources of information can be combined to produce an unambiguous outcome. The observed results match the predictions of the FLMP and reflect the optimal integration of audible and visible speech. These results also illustrate the complementarity of these two sources of information.

### 3.2 Complementarity

For whatever reasons, audible and visible speech are complimentary, and thus can together enhance the information impact of the information. Normally, two sources of information would lead to better performance than just one. This is analogous to the situation in which the perceiver is given two observations of the same information. Complementarity, on the other hand, means that a speech distinction is differentially conveyed by two sources of information. That is, two segments that are distinctly conveyed in one modality are relatively ambiguous in the other modality. For example, the difference between /ba/ and /da/ is easy to see but relatively difficult to hear. On the other hand, the difference between /ba/ and /pa/ is relatively easy to hear but very difficult if not impossible to see. The fact that two sources of information are complementary makes their integration much more informative than would be the case if the two sources were non-complementary, or redundant. One way to see this is to generate some hypothetical results of a case in which the two sources are not complementary and compare the outcome to the typical speech case with audible and visible sources of information.

We carried out this hypothetical analysis on the same results shown in the top row of Figure 2. As noted earlier, perceivers combine several sources of information in an optimal manner, resulting in better performance given two sources of information relative to just one. This optimal integration can serve as a benchmark for assessing the additional benefit of complementarity. The easiest way to implement this situation is to simply duplicate the information from one modality as the contribution of the second source of information. Thus, we simply took the observations for the unimodal auditory condition and duplicated it for our second hypothetical source of information. The two sources were combined according to the algorithm of the FLMP to generate the predicted results for the bimodal condition.

As can be seen in the left plot of the bottom row of Figure 2, there is some advantage of having one source of information duplicated: two sources are more valuable than just one. This same analysis was performed on the visual condition, with the outcome shown in the right plot of the bottom row. A similar advantage is observed. However, the actual situation in the middle column of the bottom row of Figure 2 with complementary auditory and visual speech gives much more of an advantage than what would be provided if observers had two independent sources of the same auditory information or two independent sources of the same visual information. Each modality alone gave poor performance, whereas the bimodal condition, in which these two deficient sources were combined, revealed a huge gain in performance relative to the unimodal conditions. The additional advantage of having auditory and visual speech relative to two observations of the same modality is a quintessential demonstration of the complementarity of audible and visible speech.

One way to assess the added bonus of complementarity is to compare the accuracy of performance between a single source, two noncomplementary sources, and two complementary sources. For the conditions with two sources of information, there was about a 10% gain in accuracy for the noncomplementary condition compared to a whopping 26% gain in accuracy for the complementary case. It is clear, then, that supplementing the auditory signal with visible speech provides a distinct advantage to the hearer. We now turn to the question of whether this advantage holds even under nonoptimal viewing conditions, making visible speech a robust source of information.

### 3.3 Robustness of Visible Speech

One of the functional attributes of auditory speech is that we can talk and listen with our hands full and our eyes closed. Although we have the impression that we can't help but hear speech whereas speechreading requires greater effort, there is good evidence that speechreading is robust in the sense that visible information is obtained even in what might be considered to be nonoptimal situations.

Our most recent findings show that speechreading, or the ability to obtain speech information from the face, is not compromised by a number of variables. To assess robustness, we presented observers with an animated face articulating one of four syllables without sound. The face appeared in one of five locations in the visual field. The subject's task was to identify the test

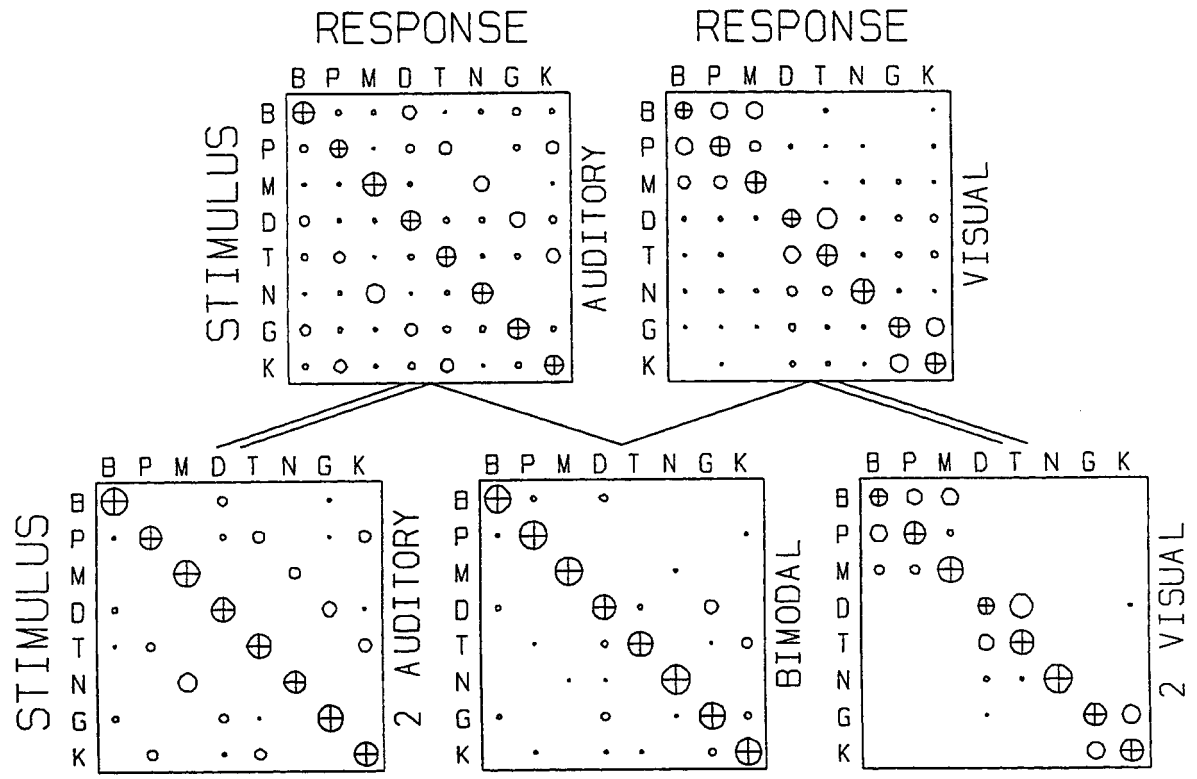


*Figure 2. Confusion matrices from Erber [4]. The area of the circle is proportional to response probability. The top row gives results from the unimodal auditory and unimodal visual conditions. The middle plot of the bottom row give the observed performance in the bimodal condition. The results show both optimal integration and a complementarity of auditory and visual speech because performance on the bimodal is so much better than it is for either the unimodal auditory or unimodal visual conditions. The observations match the predictions of the FLMP almost exactly. The right plot of the bottom row gives the FLMP's predictions for the auditory condition integrated with its duplicate. The same predictions are given in the right plot for the visual condition. The results of this simulation show an advantage of two sources even when there is a lack of complementarity, but one that is much smaller than in the complementary case in the middle column.*

syllable in a speechreading task. To prevent subjects from making eye-movements during the presentation, subjects had to detect and count changes in the size of a test dot, presented at the fixation point, simultaneously with the speechreading task.

Figure 3 shows the average correct performance of the subjects at the five spatial locations for each of the four test syllables. Replicating previous results [3], there were overall differences in performance as a function of the test syllable. Performance on the speechreading task decreased somewhat when the presentation was in the periphery relative to the center. But even with a face presented 9 degrees in the periphery, performance was only about 20% poorer than with a central view. There was also very little influence stemming from which side of the periphery the syllable was presented on.

Thus, persons are fairly good at speechreading even if they are not looking directly at the talker's lips. Furthermore, accuracy is not dramatically reduced when the facial image is blurred (because of poor vision, for example), when the image is rotated towards a profile view, when viewed from above or below, and when there is a large distance between the talker and the viewer, and even when the face is presented upside down. Another example of the robustness of the influence of visible speech is people naturally integrate visible speech with audible speech even when the temporal occurrence of the two sources is displaced by about a 1/5 of a second. These findings indicate that speechreading is highly functional in a variety of everyday (nonoptimal) situations.

## 4. Applications of FLMP to Language Learning

Our theoretical research also had the serendipitous outcome of a new technology not previously available in speech research and application. Almost 15 years ago, to achieve control over visible speech in our experiments, we used a completely animated synthetic talking head. Since that time, we have attempted to make it as realistic as possible. Today, our talking head can be heard, communicates paralinguistic as well as linguistic information, and is controlled by a text-to-speech system. Although visual information is helpful for language acquisition by people with normal hearing, it

should be a godsend for the hearing impaired. For the hearing impaired, oral language is deficient in information, making its acquisition difficult to say the least.

### 4.1 Seeing Speech

A variety of training programs have been devised to aid the deaf in the acquisition of spoken language. These current schemes provide the deaf with only more or less symbolic feedback about the accuracy of their speech production. As one example, a vowel's acoustic quality

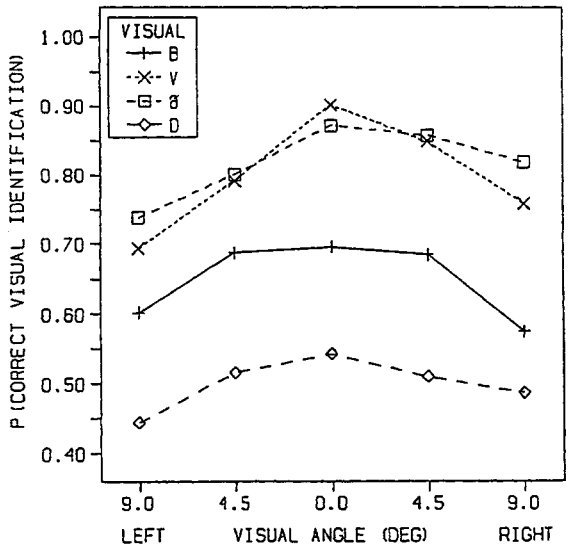


*Figure 3. Proportion correct visual identification of the test syllables /ba/, /va/, /tha/ and /da/ as a function of the side of presentation and the distance from the center of the face to the fixation point.*

might be represented by the location of a marker in a two-dimensional space corresponding to the first two formants. With our talking head, on the other hand, this trial-and-error learning might be accelerated by demonstrating the articulation directly. For example, we are able to show the position and height of the tongue in a midsagittal view of a half-face or a side view of the transparent face for each vowel category. Needless to say, this same type of learning scheme could be used in second language learning, remediation for poor readers, courses in phonetics and phonology, and even in enhancing the child's exposure to first language learning. Our talking head can also be used to provide instruction in speechreading as well as to facilitate training in visible speech production.

### 4.2 Superrealism

Our goal is to not only make our animated agent as realistic as possible, but also to develop it to display much more information than is available on real faces. The synthetic face can be further embellished by including other characteristics not normally apparent in visible speech. The velum could be raised or lowered to convey visible information about nasality. A visible breath stream could be presented during the occurrence of bursts, aspiration and frication. Because the pitch of the voice is also an informative cue, another possibility is to add movement in the neck of the animated talker to signify vocal cord vibration. In conclusion, visible speech is a valuable supplement to the auditory channel, and this belief motivates the application of our technology to language learning.

## Acknowledgments

This research was supported, in part, by a grant from the Public Health Service (PHS R01 DC00236), and a challenge grant from the National Science Foundation, and support from the the University of California, Santa Cruz.

## References

[1] Cole, R. A., et al. (1998). This volume.

[2] Massaro, D. W. (1987). Speech perception by ear and eye: A paradigm for psychological inquiry. Hillsdale. NJ: Lawrence Erlbaum Associates.

[3] Massaro, D. W. (1998). Perceiving talking faces: From speech perception to a behavioral principle. Cambridge, MA: MIT Press.

[4] Erber, N. P. (1972). Auditory, visual, and auditory-visual recognition of consonants by children with normal and impaired hearing. Journal of Speech and Hearing Research, 15, 422-423.