LEVELS IN SPEECH COMMUNICATION: RELATIONS AND INTERACTIONS
Edited by C. Sorin, J. Mariani, H. Meloni, J. Schoentgen
1995 Elsevier Science B.V., Amsterdam, The Netherlands

Perceiving Visual and Auditory Information in Consonant-Vowel and Vowel Syllables

Michael M. Cohen and Dominic W. Massaro

Program in Experimental Psychology, University of California - Santa Cruz,
Santa Cruz, CA 95064, USA. mmcohen@fuzzy.ucsc.edu

It is an honor to dedicate this paper to Max Wajskop, a valued member of our research and teaching community. His approach to the study of spoken language emphasized the important relationships and interactions among different levels and domains: particularly articulatory, acoustic, and perceptual. Our present research follows in this tradition because it is based on several related premises: a) performance is influenced by multiple sources of information, b) these sources of information exist in several modalities, such as auditory, visual, and kinesthetic, and c) these sources of information exist in at several levels, including sensory, perceptual, phonological, lexical, syntactic, and semantic. In this paper, we report three new experiments aimed to extend our understanding auditory-visual speech perception in face-to-face speech perception.

When visible speech is added to auditory speech, people find it natural to perceive bimodally—that is, to use both the audible and visible information (McGurk & MacDonald, 1976). In an earlier study (Massaro & Ferguson, 1993) subjects identified as /ba/ or /da/ speech events consisting of synthetic auditory syllables varying along a continuum from /ba/ to /da/ combined with a videotape of a person articulating a /ba/ or /da/ syllable or with no visible articulation. Although subjects were instructed specifically to report what they heard, the visible information also influenced identification. The left panel of Figure 1 gives the proportion of /da/ identifications as a function of the auditory syllable; the visual condition is the curve parameter. Both modalities contributed to the judgments and the contribution of the visual source was larger when the auditory source was more ambiguous.

The design and analysis of the Massaro and Ferguson (1993) experiment allowed the test of two competing quantitative models of the results. The fuzzy logical model of perception (FLMP) is based on three operations in perceptual recognition: feature evaluation, feature integration, and decision (Massaro, 1987). It is assumed that independent, continuously-valued features are evaluated, integrated and matched against prototypical descriptions of syllables in memory. The integration process multiplicatively combines the auditory and visual modalities as independent sources of evidence for the occurrence of syllable prototypes. An identification decision is made on the basis of the relative goodness of match of the stimulus information with relevant prototype descriptions.

An alternative set of predictions corresponds to three different models of bimodal speech perception. According to a categorical model of perception (CMP), each modality is first perceived categorically. The identification decision is based on these two separate categorizations. If the categorizations of the two sources of information agree, the identification agrees with their outcome. When the categorizations of the two different sources disagree, the identification of the speech event agrees with the categorization of one of the two sources with some fixed probability. The predictions of the CMP are mathematically equivalent to a simple weighted averaging of the two sources (Massaro, 1987). This weighted averaging

This research was supported, in part, by grants from the National Institute of Health (NINCDS Grant 20314), and the National Science Foundation (BNS 8812728) to Dominic W. Massaro.

model (WAM) is similar in all respects to the FLMP, except that the integration is additive rather than multiplicative. Finally, the CMP and WAM are mathematically equivalent to a single-channel model (SCM) in which only a single modality contributes to the judgment on a given trial (Thompson & Massaro, 1989). The SCM is important because it is grounded in the assumption that the two modalities are *not* integrated—a clear alternative to the FLMP. We call this class of models additive models of perception (AMP).

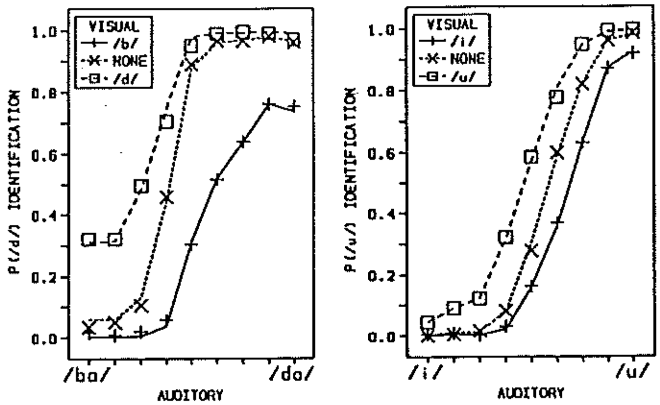

Figure 1 Proportion of observed (points) and predicted (lines) identifications as a function of the auditory and visual levels of the speech event. Left panel: results of 24 subjects following the procedure of Massaro & Ferguson, 1993. Right panel: results of Experiment 1. The lines give the predictions of the FLMP.

It is useful to describe the similarities and differences in the predictions of the identification judgments by the FLMP and AMP. Both models can predict main effects of the auditory and visual syllables on the identification judgments. According to the AMP, the effects of the two syllables should be additive; that is, the AMP predicts that the curves in the left panel of Figure 1 must be parallel to one another. In contrast, the FLMP predicts the effect of one source of information is largest when the other source is at its most ambiguous level. Thus, the FLMP predicts that the different curves in Figure 1 should be most distant from one another in the middle region of the auditory continuum. In fact, the results showed a significant interaction because the effect of the visual variable was smallest at the endpoint levels of the auditory dimension. When the models were tested against the data and the degree of fit for each model was assessed, the FLMP provided a significantly better fit to the data, thus supporting the FLMP over the AMP.

In a recent paper, Braida (1991) tested several models against confusion matrices from multimodal speech identification experiments. In these experiments, subjects were tested under three presentation conditions: two unimodal and one bimodal. Braida (1991) concluded that "Measurements of multimodal accuracy in five modern studies of consonant identification are more consistent with the predictions of the pre-labeling integration model than the FLMP (p. 1991)."

In the taxonomy of Massaro and Friedman (1990) and Cohen and Massaro (1992), Braida's pre-labeling model (PRLM) is a multidimensional version of the theory of signal detectability (TSD). A presentation of a stimulus in a given modality locates that stimulus in a multidimensional space. Given that the process is noisy (Gaussian), the location may be displaced from the stimulus center. There is also a *response center* (prototype) in the multidimensional space. The multidimensional space for a bimodal presentation is simply the combination of the spaces for the two unimodal presentations. For example, if the auditory and visual sources are each represented in 2-dimensional space, the bimodal information is represented in 4-dimensional space. In all cases, the subject chooses the response alternative whose response center (or prototype) is closest to the location of the stimulus in the multidimensional space.

In his tests of the PRLM, Braida used a multidimensional scaling (MDS) technique to find the optimal locations of stimulus centers in order to minimize the errors in prediction of each unimodal condition. The response prototypes were assumed to be equal to their respective stimulus centers. The bimodal judgments were predicted from the combined spaces of the unimodal judgments. For his fits of the FLMP, Braida simply used the unimodal data to directly predict the bimodal points. Neither of these two tests is optimal test because only the unimodal results are used to evaluate the fit. In Braida's test of the PRLM, bimodal results cannot influence the location of the stimulus centers in the multidimensional space. In the test of the FLMP, he assumed that the unimodal results are an error-free measure of the parameters of the FLMP. In the current paper, however, minimization model-fitting techniques will be applied to both the unimodal and bimodal results for the tests of both the PRLM and FLMP. Thus, we should have a direct comparison between these two models when both models are performing as optimally as possible.

It is important to replicate the results of the bimodal perception of consonant-vowel syllables with vowels because of the potential differences between the two classes of sounds. There has been a long tradition of analyzing differences between consonants and vowels in auditory speech perception. Some authors once argued that vowels are perceived differently from consonants (Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967). For example, the "categorical perception" observed for consonants did not extend to vowels, unless the vowels were made relatively short (Pisoni, 1973). The two classes of sounds also appear to differ in terms of their auditory memory (Fujisaki & Kawashima, 1970), their discriminability along the respective speech continua (Ades, 1977), their psychophysical boundaries (Pastore, 1987), or the presence of perceptual anchors (Macmillan, 1987). Finally, a major difference between the auditory form of vowels and consonants is that the acoustic information for vowels is not as transient as that defining consonant sounds (Studdert-Kennedy, 1976). A common property of all of these differences is that there is less auditory information for consonants (at least stop consonants) than for vowels. Thus, all of these accounts of consonant-vowel differences would seem to predict that visible speech would have a smaller effect with vowels than stop consonants. The reason is the known tradeoff between auditory and visible speech: the contribution of one source is attenuated to the extent that the other source is unambiguous. If vowels provide more robust auditory information, then the perception of bimodal vowels might not be influenced very much by visible speech (Summerfield & McGrath, 1984).

Vowels and consonants also differ in terms of their visible characteristics. Vowels involve slower articulatory gestures and less specific articulator positions than stop consonants. The first property should give better discrimination of visible vowels relative to visible consonants, whereas the second property implies the opposite. Compared to consonants, it is easier to articulate the same vowel with different vocal tract configurations (Ladefoged, Harshman, Goldstein, & Rice, 1978), and this often occurs because of coarticulation. This flexibility occurs in part because reductions in lip movement are possible without altering the acoustic form of vowels. Thus, one might expect fairly poor identification of visible vowels. When context is held constant, however, as with the fixed context /h-g/ used by Montgomery and Jackson (1983), all vowels can be recognized at better than chance accuracy. On the other hand, one might argue that the visual influence should be greater for vowels than consonants. There tends to be a larger perceptual range between one endpoint and another along a continuum between two vowels relative to a continuum between two consonants (Ades, 1977). Thus, vowel and consonant perception might be differentially influenced by visible speech.

These putative differences in the auditory and visible properties of vowels and consonants are large enough to justify asking whether the same model describing bimodal perception of consonants also describes the perception of vowels. That is, will the processes in the FLMP that describe auditory-visual perception of consonants also describe vowel perception? We compare both consonant-vowel and vowel syllables in both binary-choice and multiple choice identification. The binary-choice task with consonant-vowel syllables has already been carried out by Massaro and Ferguson (1993). The first experiment extends their task to study the perception of the vowels /i/ and /u/ (as in "beet" and "boot") with binary-choice responses. The results will be tested against the FLMP and AMP to determine which can best fit the data. In Experiments 2 and 3, subjects will be allowed eight responses to provide a stronger test of the models.

1. EXPERIMENT 1

1.1 Method

Subjects. Twenty-four subjects were recruited from an introductory psychology class. These subjects also served in an additional experiment reported in Massaro (1987, pp. 155-162) in which subjects identified /ba/ and /da/ syllables as quickly as possible. Half of the subjects served in that experiment first, and for the other half of the subjects, the experimental order was reversed. The subjects were given extra course credit for their participation.

Stimuli and Procedure. Prior to the experiment, a video-audio master tape was recorded, as described in Massaro (1987). On each trial the speaker said either /i/, nothing, or /u/. The lips were closed at the beginning and end of each vowel syllable. The mean auditory durations of five tokens each of the /i/ and /u/ vowels were 483 and 437 msec, respectively.

Table 1. Parameter values for nine vowel stimuli from /i/ to /u/ used in Experiment 1.

Vowel	F1	F2	F3	B1	B2	B3	AV+
/i/ 1	315	2245	3135	50	200	400	0
2	320	2112	3029	51	188	367	0
3	325	1980	2924	53	177	335	1
4	330	1848	2819	55	166	302	1
5	335	1716	2714	57	155	270	2
6	340	1583	2609	59	143	237	2
7	345	1451	2504	61	132	205	3
8	350	1319	2399	63	121	172	3
/u/ 9	355	1187	2294	65	110	140	4

Nine auditory stimuli on a continuum from /i/ to /u/ were created using a software formant serial resonator speech synthesizer (Klatt, 1980) with a sampling frequency of 10 KHz. Table 1 gives the synthesis parameters used for each 500 msec vowel. The voicing amplitude rose from 24 db at time 0 to 50 db at 70 msec to 60 db at 150 msec and then fell linearly to 48 db at 450 msec and finally to 0 db at 500 msec. To keep the overall loudness equal for the nine vowels, the voicing amplitudes were adjusted upwards for the stimuli at the /u/ end of the continuum by the amount AV+ given in Table 1. The pitch of the vowels rose from 150 Hz at time 0 to 172 Hz at 300 msec, remaining at that value until the last 100 msec during which it descended to 160 Hz. The amplitude and pitch contours closely approximated those of the natural vowels of the speaker as analyzed using linear prediction with cepstrally based pitch estimation. The F4 frequency was fixed at 4100 Hz. An experimental tape was created according to the following design. Each trial on the tape consisted of one of the nine auditory stimuli on the continuum from /i/ to /u/ paired with one of the three possible visual stimuli, /i/, neutral, or /u/. There were 11 blocks of the 27 possible speech events, sampled randomly without replacement according to a prearranged order determined at the time of recording. A partial block of 10 practice trials was created before the 297 experimental trials for a total of 307 trials. The dubbing of the synthetic speech was synchronized with the original audio track on the videotape, as in Massaro (1987) and Massaro and Cohen (1983). The dubbing was accurate to within one ms. The synthetic speech was played at a rate of 10000 samples per second and filtered 20-4900 Hz.

During the experiment, the experimental tape was played to the subjects over individual 12" inch color monitors. Four subjects could be tested simultaneously in individual sound attenuated rooms.

The audio portion of the experimental tape was presented over the built-in speakers of the video monitors at a comfortable listening level of about 67 dB-A. The audio signal from the videotape was monitored by the schmitt trigger of the DEC PDP-11/34a computer to sense the beginning of a response interval. The subjects had 3 sec to make their response by pressing adjacent buttons labeled "EE" or "OO" on a detached terminal keyboard. For half of the subjects, the "EE" button was to the left of the "OO" button. For the other half of the subjects, the positions of the two buttons was reversed. Each subject was instructed to "watch a speaker and listen to what is spoken. Your task will be to identify what you heard.

1.2 Results and Discussion

The right panel of Figure 1 gives the proportion of /u/ responses as a function of the auditory and visual levels of the stimuli. As with consonants, perception of bimodal vowels is influenced by both the visual and auditory sources, and the effect of the visual speech is attenuated at the end regions of the auditory speech.

One question of interest is the relative contribution of audible and visible speech to vowel compared to stop consonant perception. An ANOVA on the proportion of /da/ judgments from 24 subjects in Massaro and Ferguson's (1993) consonant-vowel experiment and the proportion of /u/ judgments from the current vowel experiment showed a difference in the effect of visible speech with a significant interaction of experiment and visual level, $F(2,92)=13.48$, $p<.001$. This difference can be seen by comparing the two panels of Figure 1. Although the magnitude of the influence of audible and visible speech was found to differ for consonants and vowels, the two types of speech categories could have been processed in the same manner. As described by Massaro (1987, 1989), the FLMP allows an important distinction between information and information processing. Information corresponds to how much a given stimulus characteristic supports the various alternatives. Information processing corresponds to the three processes in the model. Consonants and vowels might differ with respect to either or both of these characteristics. The FLMP makes a strong prediction, however, that the perception of consonants and vowels might differ in information, but not information processing. Thus, testing the FLMP against the results also tests whether the observed differences between consonants and vowels can be located entirely in information.

It appears that the consonants /ba/ and /da/ provide more visible information than the vowels /i/ and /u/. We hypothesize, however, that the processing of visible and auditory information occurs similarly for both types of syllables. One test of this hypothesis is to compare the description of the two type of items by the FLMP. As will be presented in the next section, the parameter values derived in the fit of the FLMP indicate that the consonants provided more visible information than the vowels. Even with this difference, however, the FLMP gave an equally good description of the identification of consonant-vowel and vowel syllables.

1.3 Model Analysis

The identification data from Experiment 1 were used to test the FLMP, AMP, and PRLM. The predictions of the FLMP and AMP are derived in Massaro (1987) As described earlier, the PRLM assumes a multidimensional representation of response prototypes. A speech stimulus is also located in this multidimensional space, and the appropriate response is determined by choosing which response prototype is the closest to the stimulus representation. In the confusion matrices analyzed by Braida, the number of responses was equal to the number of unique stimuli. In the current task, there are more auditory stimuli than responses. Therefore, we must allow the stimulus centers to differ from the response prototypes.

Because of variability in the stimuli and in perceptual processing, a given stimulus is represented as a distribution of values centered around its mean location. To determine how often a stimulus is identified as one or another response, we must determine the *proportion of cases* (i.e. the proportion of each stimulus distribution) that falls into each response region.

As demonstrated by Cohen and Massaro (1992), the PRLM model (there called the TSD-N model) makes almost identical predictions to the FLMP. If we assume logistic rather than normal distributions in the binary-response case, its predictions are identical to the FLMP. As with the FLMP, the binary-response PRLM model requires a total of 12 parameters: 3 parameters for the visual feature locations corresponding to the 3 levels of visual information in the experiment and 9 parameters for the auditory feature locations corresponding to the 9 levels of auditory information in the experiment.

The quantitative predictions of the three models were fit to the observed proportion of /u/ identifications for each subject using the 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. Initially, all parameters are set to some starting value such as .5. By iteratively adjusting the parameters of the model, STEPIT minimizes the sum of squared deviations between the 27 (3 auditory x 9 visual levels) observed and predicted points. Thus, STEPIT finds a set of parameter values which when put in the model, come closest to predicting the observed data for each subject. One might question why a unique set of free parameters must be estimated for each new experiment and for each subject. Why can't the same parameters from one experiment or from one subject be used for another experiment or another subject? The reason is that the models do not predict individual differences in terms of how /ba/-like a particular auditory (or visual) stimulus will be. The models simply predict how the two sources of continuous information are integrated and how an identification decision is made given the outcome of integration. Of course, as will be apparent in the results, there is a substantial uniformity among subjects and experiments.

The lines in Figure 1 give the predictions of the FLMP. The FLMP provided an excellent fit, whereas the AMP gave a poor description of the observed results. The FLMP gave an average RMSD of .0277 across the 24 subjects compared to an average RMSD of .1134 for the AMP. The FLMP gave a better fit than the AMP for every subject. An analysis of variance, carried out on the RMSDs showed significantly lower RMSDs for the FLMP compared to the AMP, $F\,(1,23)=70.52, p<.001$. Thus, we can reject the AMP in favor of the FLMP.

For the PRLM, the average RMSD was .0278, as expected not significantly different from the RMSD of the FLMP. The predictions of the PRLM are not given because they are essentially equivalent to the those of the FLMP. Thus, with an equivalent number of parameters, the binary-response auditory-visual paradigm does not allow us to discriminate between these two models. It should be pointed out that the version of the PRLM employed in testing binary response data uses a closed form (see Cohen and Massaro, 1992) with fixed response prototypes rather than a Monte Carlo solution with variable locations for the response prototype. With binary responses and one dimension per information source (or in general, twice the number of responses as the number of dimensions), the closed form provides an exact solution and also allows a significant reduction in computation time.

As noted in the discussion of the potential differences between identifying consonants and vowels, the FLMP makes the prediction that these two classes of items are processed in the same manner. One test of this prediction is to compare the FLMP description of vowel identification with its description of consonant-vowel identification. The fit of the FLMP to the 24 subjects in Massaro and Ferguson's consonant-vowel experiment was compared to its fit to the current vowel experiment. The FLMP provided a good fit to the consonant-vowel identifications, with an average RMSD of 0.0345. An ANOVA on the RMSD values showed no difference in the fit of the model to the two types of stimuli, $F(1,46)=2.116, p=.149$.

The next two experiments further investigate the processes of visual auditory perception for consonants and vowels respectively, using a less constrained set of response alternatives. The increase in the number of responses is motivated by the fact that percepts resulting from combinations of various auditory and visual consonant components are not limited to the percepts of the components. Rather, a variety of other identifications, including consonant clusters, occur (Massaro, 1987). Thus, the processes involved when subjects are permitted to make a variety of responses might differ from those involved in making two-choice responses. The superiority of the FLMP over the AMP might not hold when put to the test of multiple response alternatives. When cluster responses are permitted, the ability of the same model to describe both consonant-vowel and vowel syllables may no longer hold true. Furthermore, the use of a less constrained set of response alternatives might permit a definitive test between the FLMP and PRLM.

2. EXPERIMENT 2

Experiment 2 replicates Massaro and Cohen's (1983) Experiment 3, using 27 auditory-visual speech events created by combining 3 visual stimuli, /ba/, /da/, or nothing with 9 auditory stimuli on a continuum from /ba/ to /da/. The present experiment differs from the original in that the stimuli are presented using color rather than monochrome equipment. In addition, the results are used to test the FLMP, CMP, and PRLM, allowing a further comparison of these competing theories.

2.1 Method

Subjects. A group of 12 students from introductory psychology and psychology statistics classes served in both Experiment 2 and an experiment manipulating the onset asynchrony between the auditory and visual speech. Half of the group served in Experiment 2 first and for the other half of the group, the order was reversed. Some of the subjects participated for class credit and the remainder were paid five dollars for their time.

Stimuli and Procedure. The stimuli were identical to those in Experiment 1, except that the consonant-vowel syllables /ba/ and /da/ were used instead of the vowels /i/ and /u/. The auditory stimuli for the experiment were taken from the set of nine stimuli used in Massaro and Cohen's (1983) Experiments 2 and 3. The experimental procedure was identical to that used in Experiment 1. The instructions to the subjects were the same as those used for Experiment 1, except for the description of the response alternatives. Subjects identified the test items as one of eight alternatives by pressing one of eight buttons labeled /ba/, /da/, /bda/, /ða/, /dba/, /va/, /ga/, or "other".

2.2 Results

Figure 2 gives the probability of each of the eight possible responses as a function of the auditory and visual variables. The eight responses correspond to the eight labeled panels in the figure. The results show that the percepts were not limited to just /ba/ (22.1%) and /da/ (31.4%). The percepts /bda/ (14.2%), /ða/ (11.8%), and /va/ (11.3%) were frequent response alternatives in this situation. These results are very similar to those obtained by Massaro and Cohen (1983) except that the present results gave a greater proportion (11.3% vs 3.5%) of /va/ identifications. As in Massaro and Cohen (1983), the percept /bda/ occurred most often when a visual /ba/ was paired with an auditory /da/. Given a visual /ba/, the response /bda/ increased from 0 to 56% with changes from an auditory /ba/ to an auditory /da/. The symmetrical situation did not occur; subjects did not tend to hear /dba/ when a visual /da/ was paired with an auditory stimulus toward the /ba/ end of the continuum. Rather, subjects tended to hear /ða/ or /va/. The proportion of /ða/ responses reached a peak at about the fourth level of the auditory continuum. This is probably due to the acoustic similarity of the intermediate formant patterns to those for /ða/. When the more /ba/-like members of the auditory continuum was paired with visual /da/, /va/ was most often heard. This alternative might be considered a compromise between these two sources of information. Except for somewhat fewer /ba/ identifications, the effect of the neutral visual condition was essentially the same as the effect of visual /da/. This result reinforces the impression that no articulation is more similar to a /da/ articulation because /da/ can have a less noticeable articulation. Ventriloquists, for example, tend to use alveolar rather than labial consonants.

2.3 Model Analysis

The predictions of the AMP for eight responses closely follow those for the two choice data as described in the analysis of Experiment 1. The CMP assumes that separate categorizations are made to the auditory and visual sources and the identification decision is based on these separate categorizations. The predicted probability of a /ba/ identification response, P(/ba/), given a particular auditory/visual speech event, $A_i V_j$, would be:

$$P\,(/ba/\,|\,A_i\,V_j)=p\;aB_i+(1-p)(vB_j) \tag{1}$$

where i and j index the levels of the visual and auditory stimuli, respectively. The aB_i value represents the probability of a /ba/ categorization given the auditory level i and vB_j is the probability of a /ba/ categorization given the visual level j. In the WAM, aB_i corresponds to the support given by the auditory level i for the alternative /ba/, vB_j corresponds to the support given by the visual level j for the alternative /ba/, and p corresponds to the weight given the auditory modality. In the SCM, aB_i corresponds to the probability of a /ba/ identification given the auditory level i, vB_j corresponds to the probability of a /ba/ identification given visual level j, and p corresponds to the probability of using the auditory modality on that trial. In the AMP, each unique level of the auditory stimulus requires a unique parameter aB_i, and analogously for vB_j. The modeling of /ba/ responses thus requires 9 auditory parameters plus 3 visual parameters. Each of the seven other response alternatives requires an analogous equation to that given above with 12 parameters. An additional p value would be fixed across all conditions, giving a total of 97 parameters. For any particular auditory-visual combination, the sum of the eight response probabilities was constrained to be less than or equal to one.

For the FLMP, each of the response alternatives requires a prototype defined as the conjunction of some visual feature information v and some auditory feature information a. For example, the prototype for /ba/ might be defined as "forward motion of lips" and "rising F2-F3 formant pattern". The other seven alternatives would have analogous prototypes. The predicted probability of a /ba/ identification response, P(/ba/), given a particular auditory/visual speech event, $A_i V_j$, is:

$$P\,(/ba/\,|\,A_i\,V_j)=\frac{aB_i\,vB_j}{\sum aX_i\,vX_j} \tag{2}$$

where aB_i and vB_j now represent feature values supporting the /ba/-ness of the auditory and visual modalities, respectively, and aX_i and vX_j represent the feature values supporting each of the X responses. For the other responses there would be analogous equations. The model requires 3 parameters for the visual feature values and 9 parameters for the auditory feature values for each of the 8 response alternatives, giving a total of 96 parameters. This provides a fair comparison to the AMP which requires 97 parameters.

In the modeling the PRLM with eight responses, we assumed a 3-dimensional space for each of the two modalities, and thus a 6-dimensional space overall. For each of the 8 response prototypes, we therefore require 3 auditory space location parameters and 3 visual space location parameters, for a total of 48 response location parameters. Similarly, we require 3 visual space parameters for each of the 3 levels of visual information and 3 auditory space parameters for each of the 9 levels of auditory information for a total of 9 + 27 = 36 response location parameters. This brings the total number of parameters for the PRLM to 84. As with the binary-response paradigm, it is assumed that the stimuli are noisy and sometimes fall in different response regions. As with the simple case, these response regions are determined by which response prototype location is closest. However, given the multiple response locations and multidimensional space, complex multidimensional region boundaries occur and one cannot realize the model in closed form. Instead, Monte Carlo simulation must be employed. To compute the proportion of responses given each stimulus on each iteration of the model fit, we generated 1000 cases of random multidimensional normal noise which perturbed the stimulus location from its mean location. Then, the response category which was closest to that location was incremented by 1/1000. As in the other models, STEPIT adjusts the parameter values to minimize the fit of the model. To allow STEPIT to make reliable estimates, the same starting random seed was used on each iteration.

Figure 2 also gives the predicted results of the FLMP. The AMP gave a poor description of the observed results relative to the FLMP. For all 12 subjects, the FLMP gave a better fit than the AMP. The FLMP gave significantly lower RMSDs compared to the AMP, $F(1,11)=88.26$, $p<.001$. The FLMP gave a mean RMSD per point of .0346 averaged across the fit of the 12 subjects compared to an average RMSD of .1215 for the AMP. Once again, we can reject the AMP in favor of the FLMP. The same model which accounts for two-choice identification performance can account for essentially open-ended identifications. Comparing the RMSDs of the PRLM (average 0.0434) with that of the FLMP, we see a significant advantage for the FLMP, *F(1,11)=51.992, p<.001*. Although the PRLM gives a reasonable description of the results, it is still significantly poorer than the fit given by the FLMP.

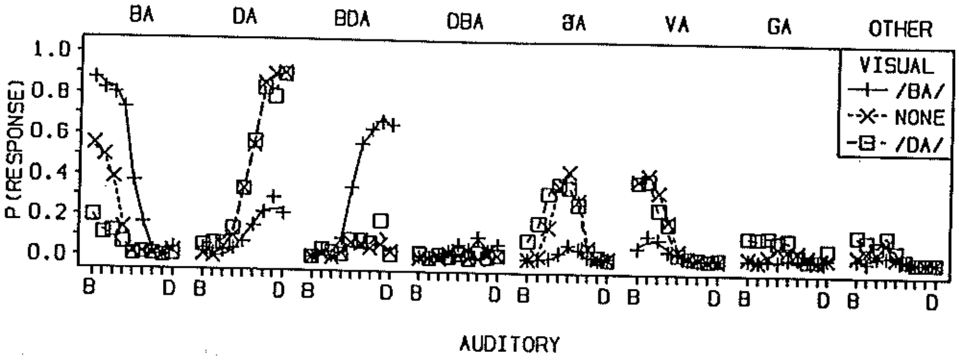

Figure 2 Proportion of observed (points) and predicted (lines) identifications each of the eight responses for Experiment 2 as a function of the auditory and visual levels of the speech event. The lines give the predictions of the FLMP.

Given the relatively small difference in the RMSDs of these two models, and the different number of parameters of the models we also compared the two using the Akaike Information Criterion (AIC) statistic (Akaike, 1974; Sakamoto, Ishiguro & Kitagawa, 1986). This formal theory takes into account the likelihood of a model fit and also the number of parameters used by the model. When several models give an approximately equally likely fit of the observed data, the AIC statistic would say that we should choose the model with the fewest parameters. In this sense, the inclusion of the number of parameters in computing the AIC allows us to contrast different models with a varying number of free parameters.

We note first that our model fits using STEPIT minimize the squared deviations of the observed and predicted data which yields a maximum likelihood fit; thus the likelihood of the obtained fit is the maximum likelihood. The general form for the exact likelihood (L) of this obtained fit is given by the product of the multinomial distributions for each stimulus condition:

$$L = \prod_{s} \left[\frac{(\sum_{r} f_{sr})!}{\prod_{r} f_{sr}!} \times \prod_{r} p_{sr}^{f_{sr}} \right] \tag{3}$$

where f_{sr} is the observed frequency of response r to stimulus s and p_{sr} is the predicted proportion of response r to stimulus s. The log-likelihood (LL) of the fit is given by:

$$LL = \sum_{s} (ln((\sum_{r} f_{sr})!) - \sum_{r} ln(f_{sr}!) + \sum_{r} f_{sr}\, ln(p_{sr})) \quad . \tag{4}$$

The AIC statistic is computed as:

$$AIC = -2(maximum\ LL) + 2(number\ of\ parameters). \tag{5}$$

Smaller AIC values are preferred. From the relationship between the AIC quantity and entropy, if the difference in AICs between models is at least 1 or 2, then the difference is considered to be significant. If the difference is much less than 1, then the models are equally good in describing the data.

In computing the AIC values for the FLMP and PRLM models, we must take into account the true number of free parameters of the models. For the FLMP, although the actual number of parameters used in the fit was 96, one could reduce the number of free parameters needed by certain constraints. For example, by assuming that the fuzzy support for the 8 responses sums to one for each of the 27 conditions, one can reduce the number of free parameters to 84. Additionally, one can fix one of the remaining parameters to arrive at a free parameter count of 83. For the PRLM, one could fix one of the values for each dimension (e.g. of the stimulus centers) to 0, which would save 6 parameters, and perhaps also impose 4 constraints from space rotations, for a total free parameter count of 74. These "free parameter" counts (83 for FLMP and 74 for the PRLM) were then used in the AIC values comparing the models. By the AIC test, the fits for all 12 subjects favored the FLMP model.

3. EXPERIMENT 3

The next experiment extends the eight-response procedure to the perception of the vowels /i/ and /u/. Although consonant clusters have clearly been observed with conflicting auditory and visual stimuli, the same cannot be said for vowels. Summerfield and McGrath (1984) reported a set of experiments on the auditory-visual perception of vowels. In one experiment three vowel continua were used. In separate blocks, 11 member series between pairs of the three point vowels, /i/, /a/, and /u/ were presented in a /bVd/ environment approximately synchronized with a visual articulation from either end of the given continuum. Vowel clusters were not among the response set and were not reported. The responses made by the subjects were transformed into points in a F1′-F2′ equivalent space representation of the four formant values of the response categories. The influence of the visual articulations was measured in terms of the average vector in this space between auditory alone and auditory-visual identifications. In general, the length of the vectors seemed to be larger to the extent that the auditory-visual event was ambiguous, with a bias toward the visual vowel. Although the results are compatible with the FLMP, it is difficult, given their analysis, to determine exactly what responses were made. Thus a quantitative test is not possible.

Experiment 3 explicitly tests whether vowel clusters will occur analogous to those occurring with incongruous auditory-visual consonants. For example, when visual /u/ and auditory /i/ are combined, will /ui/ be observed analogous to /bda/? Both /u/ and /b/ are similar in having somewhat earlier and more salient visual articulations at their onsets. Also, will there be few /iu/ percepts analogous to the rarity of /dba/ percepts? In Experiment 3, subjects are presented with the same stimuli as in Experiment 1. The visual-auditory events are composed of /i/, neutral, or /u/ articulations combined with one of nine vowels from an /i/-/u/ continuum. In addition to /i/ and /u/ responses, subjects could identify the auditory-visual event as /iu/, /ui/, /a/, /I/, /o/, or "other". The latter two are close to /i/ and /u/ respectively in the F1-F2 vowel space and there is some indication that subjects made a number of responses in these directions in the Summerfield and McGrath (1984) study.

3.1 Method

Subjects. Twelve students from introductory psychology and psychology statistics classes served in both Experiment 3 and another experiment on modality asynchrony (Massaro & Cohen, 1993). Half of the group served in Experiment 3 first and for the other half of the group, the order was reversed. Some of the subjects participated for class credit and the remainder were paid 5 dollars for their time.

Stimuli and Procedure. The stimuli and experimental setup was identical to that used in Experiment 1 except for the number and labeling of the response buttons and the instructions to the subjects. Subjects identified the test items as one of the vowels or clusters "EE", "OO", "EE-OO", "OO-EE", "AH", "IH", "OH", or as "OTHER". The response alternatives were carefully pronounced for the subjects in the instructions.

3.2 Results

Figure 3 gives the probability of each of the eight possible responses as a function of the auditory and visual variables. Most of the responses were either /i/ (39.1%) or /u/ (45.2), whereas the responses /iu/ (4.5%), /ui/ (5.6%), /I/ (5%), and "other" (.5%) occasionally occurred for the more ambiguous sounds between /i/ and /u/. The pattern of /i/ and /u/ responses as a function of the auditory and visual levels of the stimulus closely resembled the results of Experiment 1. In contrast to the results with consonant-vowel syllables in Experiment 2, few cluster responses are observed and one cluster category did not occur more than the other.

3.3 Model Analysis

The AMP, FLMP, and PRLM and their tests were identical in form to those used in Experiment 2. Figure 3 also gives the predicted results of the FLMP. The FLMP gave a mean RMSD per point of .0198 averaged across the 12 subjects compared to a mean RMSD of .0614 for the AMP. The FLMP gave a better fit than the AMP for every subject. An analysis of variance on the RMSDs showed that the FLMP gave significantly lower RMSDs compared to the AMP, $F(1,11)=83.085$, $p<.001$. For the PRLM comparison, the FLMP (average RMSD=.0198), was significantly better than the fit of the PRLM (average rmsd=.0226), *F(1,11)=8.87, p=.012*. As for Experiment 2, we also used the AIC to compare the FLMP and PRLM models. This time the test was somewhat closer, with the FLMP winning for seven of the twelve subjects.

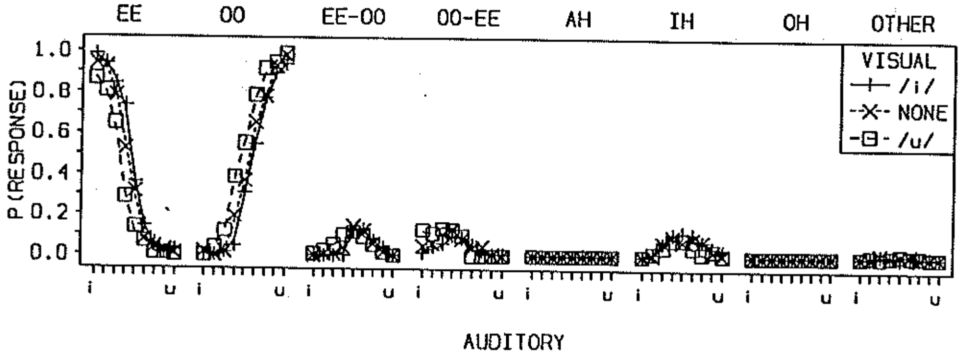

Figure 3 Proportion of observed (points) and predicted (lines) identifications each of the eight responses for Experiment 3 as a function of the auditory and visual levels of the speech event. The lines give the predictions of the FLMP.

Why did vowel cluster percepts occur so rarely in this experiment and in Massaro and Cohen (1993)? One interpretation might be that the auditory and visual information for the vowel persists for too long a time to allow a cluster to be heard. That is, while the transient auditory and visual information associated with consonants is short enough to be perceived serially, the vowel auditory and visual information overlap so much in time that they are perceived simultaneously. A related explanation is that visual /u/ does not precede auditory /i/ by enough to get an /ui/. However, Cohen (1984, Experiment 3) showed that the visual information for /u/ is available about 150 ms sooner than the auditory /i/. A third explanation is that the visual information associated with these vowel articulations is simply not compatible with a vowel cluster. This idea proved to be not only consistent with the parameter values for the visual features associated with each alternative but also reinforces our preferred explanation for consonant cluster perception. The visible articulation of some consonants is perceptually similar to the articulation of certain consonant clusters. Thus, a /ba/ articulation is very similar visually to a /bda/ articulation and subjects might hear /bda/ when visual /ba/ is paired with auditory /da/. (Visual /ba/ or /da/ articulations are not similar to /dba/ articulations, however, and few /dba/ responses are observed.) The analogous result does not occur for /u/ and /i/ since the visual articulation of /u/ differs significantly from that of /ui/

4. GENERAL DISCUSSION

The results support the idea that consonants and vowels are processed similarly in bimodal speech perception. The results of three experiments support the FLMP view that auditory and visual information is processed in a three stage process of feature evaluation, integration, and decision. The results indicated a somewhat larger visual effect for consonants than for vowels. This result is compromised, however, because the influence of a given source of information is dependent on the quality of that information as well as the quality of the other available sources of information. Consistent with previous research (Ades, 1977), we also found a somewhat larger auditory effect for vowels than consonants. This fact will necessarily decrease the impact of the visible speech for the vowels relative to the consonants. We can only conclude that the relative influence of visible to auditory speech appears to be larger for the consonant-vowel syllables /ba/ and /da/ than for the vowels /i/ and /u/.

Although consonant-vowel syllables and vowels might differ in terms of the relative impact of visible and audible speech, we can conclude that the two sources of information are integrated in the same manner in both types of segments. The good description of the FLMP shows that the multiplicative integration process is capturing something fundamental about the combination of visible and auditory speech. The model is not simply a quantitative description of ceiling and floor effects in probability judgments.

In addition to finding support for the FLMP, the results also falsified several viable models of speech perception. For both vowels and consonants, and for both binary and multiple responses, the fit of the FLMP was superior to that for the AMP. Given that the AMP includes a categorical model, a weighted averaging of the two sources of information, and a single channel model in which only a single source influences the judgment on a given trial, the results are equally damaging to all three of these models. While the PRLM could not be discriminated from the FLMP in predictive power for the binary-response paradigm, we found that the PRLM was significantly inferior to the FLMP in predicting multiple responses in Experiments 2 and 3.

REFERENCES

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, AC-19,* 716-723.

Ades, A. E. (1977). Vowels, consonants, speech, and nonspeech, *Psychological Review, 84,* 524-530.

Braida, L. D. (1991). Crossmodal integration in the identification of consonant segments. *The Quarterly Journal of Experimental Psychology, 43A,* 647-677.

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

Cohen, M. M. (1984). Processing of visual and auditory information in speech perception. Dissertation, University of California, Santa Cruz.

Cohen, M. M. & Massaro, D. W. (1992). On the similarity of categorization models. In F. G. Ashby (Ed.) *Multidimensional Models of Perception and Cognition* Hillsdale, NJ: Lawrence Erlbaum Associates.

Fujisaki, H., & Kawashima, T. (1970). Some experiments on speech perception and a model for the perceptual mechanism, *Annual Report of the Engineering Research Institute, 29,* Tokyo: Faculty of Engineering, University of Tokyo, 206-214.

Klatt, D.H. (1980). Software for a cascade/parallel formant synthesizer. *Journal of the Acoustical Society of America, 367,* 971-995.

Ladefoged, P. Harshman, R., Goldstein, L. and Rice (1978). *Journal of the Acoustical Society of America, 364,* 253-257.

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

McGurk, H., & MacDonald (1976) Hearing lips and seeing voices. *Nature, 264,* 746-748.

Macmillan, N. A. (1987). Beyond the categorical/continuous distinction: A psychophysical approach to processing modes, in S. Harnad (Ed.) *Categorical perception*, New York: Cambridge University Press.

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

Massaro, D. W. and 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.

Massaro, D. W. and Cohen, M. M. (1993). Perceiving Asynchronous Bimodal Speech in Consonant-Vowel and Vowel Syllables *SPeech Communication, 13* 127-134.

Massaro, D. W. and Friedman, D. (1990). Models of integration given multiple sources of information, *Psychological Review, 97,* 225-252.

Massaro, D.W. & Ferguson, E. L. (1993) Cognitive style and perception: The relationship between category width and speech perception, categorization, and discrimination, *American Journal of Psychology, 106,* 25-49.

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

Pastore, R. E. (1987). Categorical perception: Some psychophysical models, in S. Harnad (Ed.) *Categorical perception*, New York: Cambridge University press.

Pisoni, D. B. (1973). Auditory and phonetic memory codes in the discrimination of consonants and vowels, *Perception and Psychophysics, 13,* 253-260.

Sakamoto, Y., Ishiguro, M., & Kitagawa, G. (1986) *Akaike Information Criterion Statistics*, Dordrecht, Holland: D. Reidel Publishing.

Studdert-Kennedy, M. (1976). Speech Perception, in N.J. Lass (Ed.) *Contemporary issues in Experimental Phonetics*, New York: Academic Press.

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