Perception of asynchronous and conflicting visual and auditory speech

Dominic W. Massaro,[a] Michael M. Cohen, and Paula M. T. Smeele
Department of Psychology, University of California, Santa Cruz, California 95064

(Received 26 June 1995; revised 8 May 1996; accepted 17 May 1996)

Previous research has shown that perceivers naturally integrate auditory and visual information in face-to-face speech perception. Two experiments were carried out to study whether integration would be disrupted by differences in the stimulus onset asynchrony (SOA), the temporal arrival of the two sources of information. Synthetic visible and natural and synthetic auditory syllables /ba/, /va/, /ða/, and /da/ were used in an expanded factorial design to present all possible combinations of the auditory and visual syllables, as well as the unimodal syllables. The fuzzy logical model of perception (FLMP), which accurately describes integration, was used to measure the degree to which integration of audible and visible speech occurred. These findings provide information about the temporal window of integration and its apparent dependence on the range of speech events in the test. © *1996 Acoustical Society of America.*

PACS numbers: 43.71.An, 43.72.Ja, 43.71.Ma [RAF]

INTRODUCTION

Previous research has shown that seeing the talker's face can improve the intelligibility of speech, relative to the presentation of only auditory information. This improvement is most easily demonstrated in situations in which the auditory signal is degraded, e.g., due to a hearing impairment, the presence of noise, or bandwidth filtering (Binnie *et al.* 1974; Breeuwer and Plomp, 1984; Erber, 1972; Massaro, 1987; McGrath and Summerfield, 1985; Sumby and Pollack, 1954; Summerfield, 1979). Visual speech predominantly provides information about place of articulation for consonants (Grant and Walden, 1995), but also cues manner of articulation to some extent. In addition, visible speech cues vowel identity (Montgomery and Jackson, 1983) and probably also gives information about segmentation and prosody (Risberg and Lubker, 1978). Speech perception is superior in the presence of additional visual information, whether the speech materials presented are sentences (Reisberg *et al.*, 1987; Risberg and Lubker, 1978; Summerfield, 1979), meaningful words (Campbell and Dodd, 1980), or nonsense syllables (Binnie *et al.*, 1974; Smeele and Sittig, 1990, 1991). The generality of the results indicates that there is a contribution of vision to speech perception regardless of lexical status or sentential context.

Various models of how audible and visible sources of information are utilized have been proposed. They can be classified as either integration models or nonintegration models. Examples of integration models are the fuzzy logical model of perception (FLMP, Massaro, 1987, 1989, 1990) and the "prelabeling" integration model (Braida, 1991), whereas an auditory dominance model (Sekiyama and Tohkura, 1991; Vroomen, 1992) and a single-channel model of perception (Thompson and Massaro, 1989) can be regarded as nonintegration models. The FLMP assumes multiple sources of information—considered as continuously valued features—are evaluated, integrated, and matched against prototype descriptions in memory. In the "prelabeling" integration model, a multidimensional version of the theory of signal detection, continuously valued cues are combined across sensory systems to produce a vector of cues (in a multidimensional space) that is mapped onto a identification response (Braida, 1991). The auditory dominance model assumes that visible speech contributes only in situations when the audible speech is not sufficiently informative. In the single-channel model of perception, decisions are made on each modality separately and the perceiver responds with the decision of either modality with a certain probability.

To gain more insight into audiovisual speech perception, and to discriminate among these models, experiments have presented conflicting auditory and visual information (Green and Kuhl, 1989; Massaro, 1987, 1989; Massaro and Cohen, 1983; McGurk and MacDonald, 1976; Mills and Thiem, 1980). These studies show that vision strongly influences perception and, in certain cases, even leads to perception of items that were presented neither auditorily nor visually. The FLMP has consistently given the best description of results from a wide variety of experiments (Massaro and Cohen, 1990; Massaro *et al.*, 1993; Massaro *et al.*, 1995).

In addition to providing conflicting information, it is also valuable to manipulate the synchronization between the audio and visual channels. This method can provide information about the interval during which the auditory and visual speech sources can be integrated. Koenig (1965) carried out an early study with a single subject with low-pass filtered speech consisting of isolated words and sentences. Performance was not disrupted until the delay exceeded 240 ms. Campbell and Dodd (1980) presented subjects with consonant-vowel-consonant (CVC) words. They used desynchronizations of 400, 800, and 1600 ms in which vision always preceded audition. Although phoneme identification decreased somewhat when compared to the synchronous condition, it was found that identification was invariably better in the asynchronous bimodal conditions than in the audi-

[a]Corresponding author.

 0001-4966/96/100(3)/1777/10/$10.00

tory only control condition. Pandey *et al.* (1986) studied the effect of audio delays of 0, 60, 120, 180, 240, and 300 ms for audiovisual presentation of sentences. They also degraded the auditory signal by mixing it with a multitalker babble. Although the results were somewhat variable, they found that accuracy in the audiovisual conditions at all delay values was significantly better than scores either with vision or audition alone. Moreover, delays of up to 120 ms did not indicate any significant deterioration of scores when compared to the synchronous condition. Munhall *et al.* (1996) found an influence of visible speech even when the audible speech lagged the visual by 180 ms.

Massaro and Cohen (1993) also looked for differences in the integration process with changes in SOA. A visual /ba/ or /da/ was combined with an auditory /ba/ or /da/ with a SOAs of up to plus or minus 200 ms between the auditory stimulus and the voice onset of the original audio. Participants were instructed to watch a talker and listen to what was spoken and to identify what was *heard*. When the visual and auditory syllables were identical, the responses were mostly accurate. When visual /ba/ was combined with auditory /da/, the predominant responses were /da/ (31%) and /bda/ (52%). The likelihood of a /bda/ judgment increased from 31% when the auditory preceded the visual syllable to 69% when the visual syllable preceded the auditory. The large number of /bda/ judgments is consistent with the idea that visual /ba/ is highly similar to a visual /bda/ articulation. The increase in /bda/ judgments as the visual /ba/ was presented earlier than the auditory information might indicate that cluster responses occur because one modality is processed sooner than the other.

Although the temporal difference between the two modalities does not appear to be sufficient to produce a large proportion of cluster responses, the temporal relationship between the auditory and visual information can modulate the number of cluster judgments. The judgment /dba/ occurred only about 10% of the time when visual /da/ preceded auditory /ba/. If consonant clusters occurred simply because of differences in arrival times of the two information sources, then there should have been as many /dba/ judgments as there were /bda/ judgments. The results instead lend support to the hypothesis that clusters occur when both the visual and auditory syllables are consistent with the articulation of a consonant cluster. Judgments of /dba/ seldom occur because a visual /da/ is highly dissimilar to a /dba/ articulation.

Massaro and Cohen (1993) performed a second experiment with the vowels /i/ and /u/. Although cluster responses did not occur, the results indicated that the auditory and visual sources were integrated at all SOAs as large as 200 ms. It remains to be determined when integration of auditory and visual vowel syllables breaks down.

These earlier studies do not establish unambiguously the SOAs at which integration breaks down. The major reason is that the previous research did not design the experiments or use the results to test extant models of bimodal speech perception. Integration is one of the central assumptions of the FLMP, and a good fit of this model to the results insures that integration occurred (Massaro, 1987). The test of this model will allow us to determine when integration of audible and visible speech did occur. Given that the FLMP has been repeatedly shown to give an excellent description of bimodal speech perception (Massaro, 1987), the goodness-of-fit of this model can be used to determine if SOA disrupts the integration of auditory and visual speech. For example, integration might not occur with SOAs greater than some minimal duration and, therefore, the FLMP should give a poor description at SOAs larger than this value.

Within the 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 assumptions central to the model are (1) each source of information is evaluated relative to prototypes in memory to give the degree to which that source specifies the relevant alternatives, (2) the sources of information are evaluated independently of one another, (3) all of the sources are integrated relative to prototypes in memory to provide an overall degree of support for each alternative, and (4) perceptual identification follows the relative degree of support among the alternatives. In bimodal speech perception, both sources are assumed to provide continuous and independent evidence for each of the alternatives, the integration of the sources is multiplicative, and the decision operation determines the support for one alternative relative to the sum of the support for each of the relevant alternatives.

With respect to the temporal alignment of vision and audition, the FLMP predicts integration by implicitly assuming that the visible and audible speech are synchronous. It predicts integration across different asynchronies as long as the two modalities are perceived as belonging to the same perceptual event (Massaro, 1985, 1987). In this framework, it is only natural to integrate the two sources of information when they represent the same perceptual event. If the integration predicted by the FLMP does not occur given a large asynchrony, the model would necessarily give a poor description of the results.

I. METHOD

A. Subjects

The participants were 28 native talkers of American English and were students from the University of California, Santa Cruz. In experiment 1, five subjects participated for a course requirement and five were paid 6 dollars per hour. Their ages ranged from 19 to 22 years (average 20.4 years). In experiment 2, 11 subjects participated for a course requirement and six were paid 25 dollars. One subject received some course credit and 12 dollars. Their ages ranged from 18 to 22 years (average 19.5 years). All subjects reported having normal hearing, and normal or corrected-to-normal vision. It is not necessary to screen the participants for deficits in hearing and vision because (1) both auditory and visual unimodal trials are presented during the experiment and (2) the model tests are robust across individual differences (see Massaro, 1992).

1. Synthetic and natural audible speech

The synthetic syllables, with approximately equal vowel duration, were produced by a software formant serial resonator speech synthesizer (Klatt, 1980). The durations of the

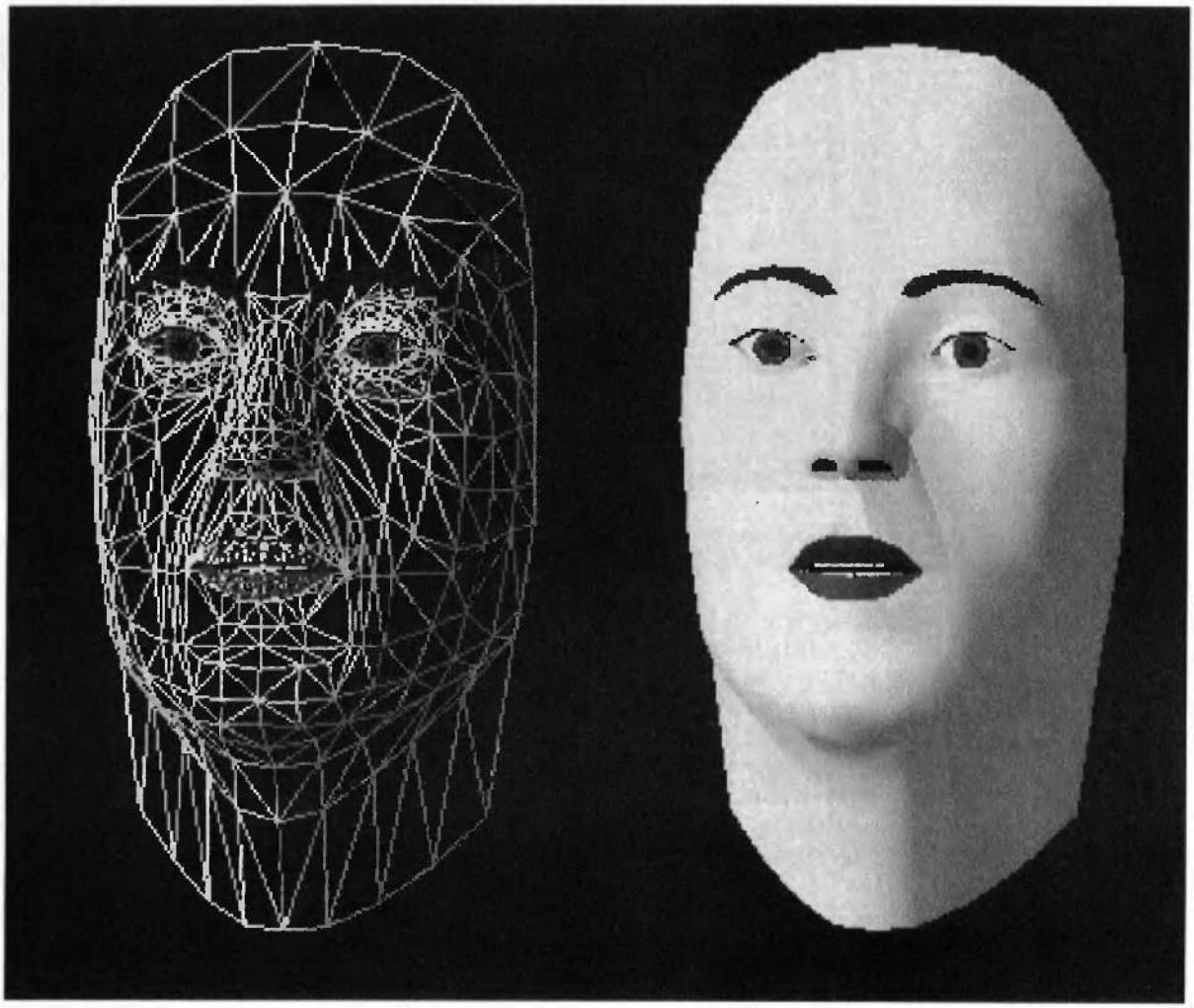

FIG. 1. Framework (left) and Gouraud shaded (right) renderings of polygon facial model.

complete syllables were 354, 414, 482, and 328 ms for /ba/, /va/, /ða/, and /da/, respectively. The natural auditory speech were the same syllables spoken by a male talker taken from the Bernstein and Eberhardt (1986) videodisk database. Their durations were 396, 470, 506, and 422 ms for /ba/, /va/, /ða/, and /da/.

2. Synthetic visible speech

A parametrically controlled polygon topology was used to generate the visible speech syllables (Cohen and Massaro, 1990, 1993, 1994). The animated display was created by modeling the facial surface as a polyhedral object composed of about 900 small surfaces arranged in 3D, joined together at the edges (Parke, 1975, 1982). The left panel of Fig. 1 shows a framework rendering of this model. To achieve a natural appearance, the surface was smooth shaded using Gouraud's (1971) method (shown in the right panel of Fig. 1). The face was animated by altering the location of various points in the grid under the control of 50 parameters, 11 of which were used for speech animation. Each phoneme is defined in a table according to target values for 18 control parameters and segment duration. The control parameters include jaw rotation, mouth x scale, mouth z offset, lip corner x width, mouth corner z offset, mouth corner x offset, mouth corner y offset, lower lip "f" tuck, and upper and lower lip raise, tongue angle and length, and jaw thrust. Parke's software, revised by Pearce *et al.* (1986) and ourselves (Cohen and Massaro, 1990, 1993) was implemented on a Silicon Graphics Inc. Crimson-VGX computer. To achieve a more realistic synthesis, a tongue was added, with control parameters specifying its angle, length, width, and thickness.

The synthetic face was programmed to pronounce the CV syllables /ba/, /va/, /ða/, and /da/. Figure 2 shows the face at the onset of the articulation of the four syllables.

Audiovisual stimuli were created by combining the auditory speech of the four syllables with the visual speech of each of these syllables. The durations of the dynamic portions of the visible speech were approximately 730 ms for /ba/, 730 ms for /va/, 900 ms for /ða/, and 667 ms for /da/. Thus the dynamic portion of the visual syllable began before and finished after the co-occurring auditory syllable. The synthetic visible speech was modeled after natural syllables, and the beginning of the auditory speech was placed exactly where the auditory speech would have begun in the natural

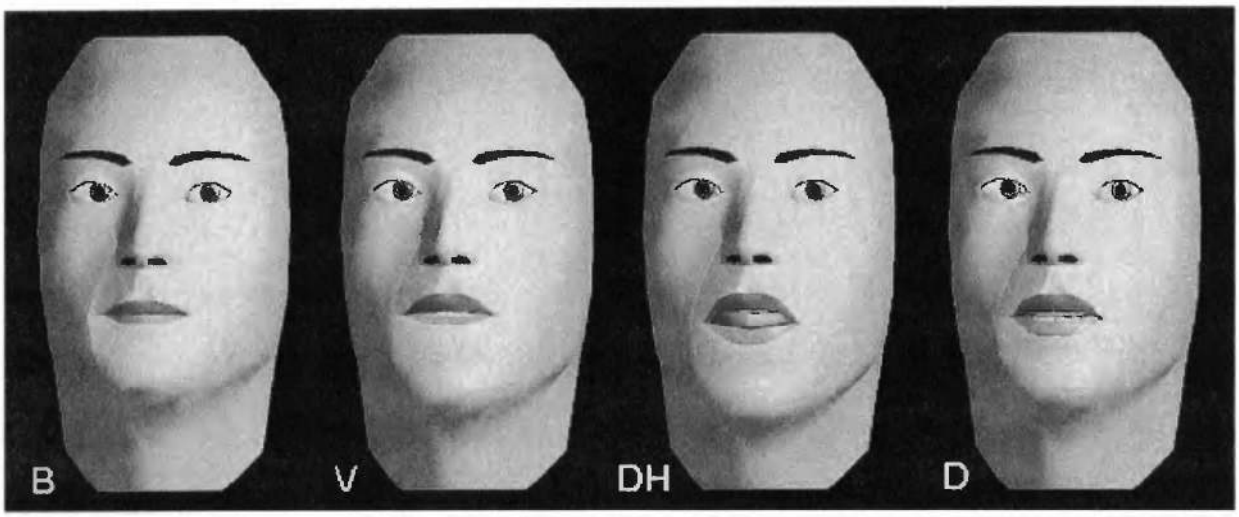

FIG. 2. The facial model at the onset of the syllable for each of the four consonants. The lips are closed at the onset of /ba/, much of the lower lip is hidden by the teeth in /va/, the tongue is between the teeth in /ða/, and the mouth is slightly open at the onset of /da/.

syllable (at 400, 300, 500, and 300 ms into the synthetic visual speech syllables /ba/, /va/, /ða/, and /da/, respectively). The asynchronous SOAs were created by offsetting the auditory and visual syllables by the specified duration relative to their normal co-occurrence at the simultaneous SOA condition. The synthetic face with default parameter values (neutral face) was presented for 700 ms preceding the visual syllable and occurred again at the offset of the syllable and remained on until the beginning of the next trial.

3. Apparatus and materials

The experimental stimuli were presented to the subjects over individual NEC model C12-202A 12-in. color monitors. The synthetic face was displayed in the center of the screen and subtended a visual angle of about 10 deg. The loudness level of the auditory stimuli was 67 dB A (B & K 2231). For visual-alone and bimodal conditions, a 100-ms, 1000-Hz tone was presented with a 700-ms neutral face. For auditory-alone conditions, each trial started with the same tone presented with a 700-ms black screen.

In both experiments 1 and 2, the auditory and visual syllables were presented at seven SOAs. In experiment 1, the SOAs were −267 ms (8 frames), −167 ms (5 frames), −67 ms (2 frames) where audition preceded vision, 0 ms (synchronous), 67 ms (2 frames), 167 ms (5 frames), and 267 ms (8 frames) where vision preceded audition. In experiment 2, the intervals were −533 ms (16 frames), −267 ms (8 frames), −133 ms (4 frames), 0 ms (synchronous), 133 ms (4 frames), 267 ms (8 frames), and 533 ms (16 frames). The visual information was presented at a rate of 30 frames/s. For the unimodal auditory condition, the monitor was blank during the trial. There was no sound during the unimodal visual condition.

4. Design and procedure

In each experiment, natural and synthetic auditory and synthetic visual speech were manipulated in an expanded factorial design, illustrated in Fig. 3. Each of the syllables was presented alone, as well as paired with every syllable from the other modality. Thus there were 12 unimodal trials consisting of four visible syllables and eight auditory syllables (four synthetic and four natural). For bimodal audiovisual speech, there were eight (four synthetic+four natural) auditory stimuli times four synthetic visual stimuli times seven desynchronizations=224 independent conditions. To equate the number of unimodal and bimodal trials, there were seven times as many observations for each of the 12 unimodal trials. Thus the total number of trials randomized within a trial block was 224+84=308. Six random blocks were generated yielding 6 times 308=1848 test trials for each subject.

Subjects were instructed to listen to and watch the talker, and to identify what was presented as /b/, /v/, /ð/, or /d/, or as a combination of two of these alternatives (a consonant cluster). The subjects made their responses by pressing a key on the terminal keyboard labeled as "b," "v," "th," or "d." For consonant clusters, the subjects hit two of these keys in the appropriate order. In either case, the re-

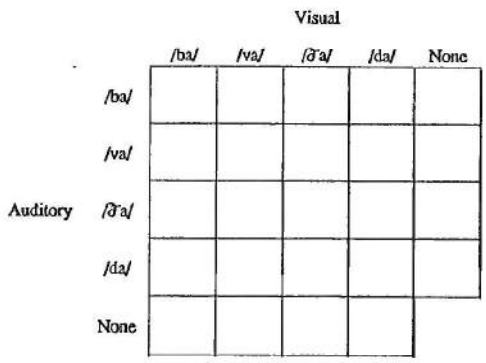

FIG. 3. Expanded factorial design with four auditory syllables crossed with four visual syllables.

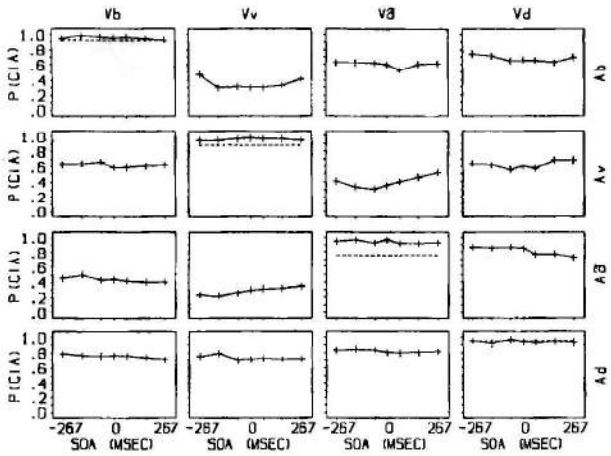

FIG. 4. The average correct performance scored with respect to the auditory stimulus. Individual panels corresponding to the results for the 16 different bimodal stimuli. The visual stimulus (V) is indicated at the top, the auditory stimulus (A) is indicated to the right. The abscissa is the desynchronization between the auditory and visible syllables in ms. Negative values indicate audition leads. The percentage of responses that matched the presented auditory stimulus is given on the ordinate. The results are pooled over natural and synthetic auditory speech stimuli. In the graphs for the matching audiovisual conditions the performance in the auditory-alone conditions is indicated by the dashed straight lines. Averaged results from ten subjects in experiment 1.

sponse was completed by hitting the return button.

Subjects were tested on two blocks of 308 trials each per day with a short break after the first block. Subjects were tested on three successive days, except for eight subjects who had 3 blocks per day for two successive days. Before the recorded trials of a session, subjects responded to ten unscored practice trials. There was never any feedback in the task.

Up to four subjects could be tested simultaneously in individual sound-attenuated rooms. The experiment was subject-driven, e.g., a next trial would only occur when all of the simultaneously tested subjects had responded. After the last subject had responded there was a 1-s intertrial interval.

II. RESULTS: EXPERIMENT 1

The identification judgments were recorded and the mean observed proportion of identifications was computed for each subject at each stimulus condition. The overall proportion of cluster responses was relatively small (8%) and therefore we settled on five response categories of the four syllables plus a cluster category. Previous research using these syllables has shown that it is informative to analyze the results as a function of accuracy with respect to each of the two modalities (Massaro and Cohen, 1995). Figure 4 presents accuracy performance scored with respect to the auditory speech stimulus. The figure gives individual graphs corresponding to the 16 bimodal conditions. In the four graphs (along the negative diagonal) for the matching audiovisual conditions, the performance based on audition alone is also indicated by the straight dashed lines. To simplify the graphs, the plotted results are pooled over the natural and synthetic auditory syllables. However, the data analyses and tests of the FLMP included natural versus synthetic auditory syllable as a factor. The seven levels of desynchronization are plotted on the x axis of each individual graph. Negative values correspond to the case in which auditory speech occurred earlier than the visual.

As can be seen in the four graphs along the negative diagonal, auditory accuracy with consistent audiovisual speech was better than with the unimodal auditory syllables, $F(1,9)=15.5$, $p=0.004$. This advantage differed for the four different syllables; for /b/ (0.950 vs 0.922), /v/ (0.954 vs 0.880), and /ða/ (0.937 vs 0.758), and for /d/ (0.942 vs 0.948), $F(3,27)=8.446$, $p<0.001$.

Figure 4 also shows that inconsistent visual information decreased auditory accuracy relative to the unimodal auditory condition. This result is most easily seen by comparing the curves across each row. The three curves from the inconsistent conditions can be compared to the straight line representing the unimodal auditory condition. For example, the first row in Fig. 4 shows that auditory accuracy is lower in the three inconsistent conditions relative to the unimodal auditory condition. Although the decrease in auditory accuracy with inconsistent visual information occurs for all syllables, the effect is larger for some combinations than others. Auditory /da/ (bottom row) is least susceptible to inconsistent visual information and visual /da/ (right column) has the least impact on auditory speech.

Not shown in the figure, performance on the unimodal auditory speech was somewhat better for natural than for synthetic speech (93% vs 83%), $F(1,9)=18.7$, $p=0.002$. This superiority of natural speech is largely due to the relatively poor identification of synthetic /ða/, with an average of about 62% correct. Consistent with the poorer quality of synthetic speech, subjects showed only a 3% improvement in accuracy when the visual syllable matched the natural auditory syllable and a 10% improvement when it matched the synthetic auditory syllable, $F(1,9)=9.01$, $p=0.014$.

A complementary analysis is to examine performance with respect to accuracy on the visual speech syllable. Figure 5 gives these results plotted in the same manner as Fig. 4. The four panels for the consistent bimodal conditions also give performance based on the unimodal visual condition (straight dashed lines). Identification of the visible speech was relatively good (average 87%). Performance on visual /d/ was somewhat poorer than on the other three syllables, which could be due to a less significant facial movement relative to the following vowel (/a/).

As can be seen in the four graphs along the negative diagonal in Fig. 5, performance with consistent audiovisual speech was somewhat better than the unimodal visual syllables, and this advantage differed for the four different syllables.

Figure 5 also shows that inconsistent auditory information decreased visual accuracy relative to the unimodal visual condition. This result occurred for all syllables and for all combinations.

Comparing Fig. 5 with Fig. 4 indicates that overall performance was more influenced by the auditory than the visual information. Most of the curves in Fig. 4 depicting auditory accuracy are higher than the corresponding curves in Fig. 5 depicting visual accuracy. Visual speech had a larger

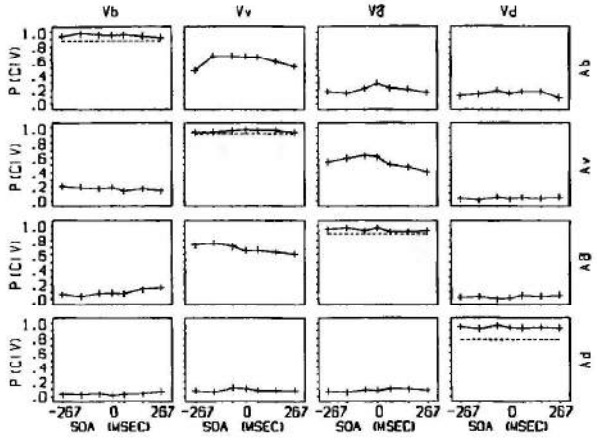

FIG. 5. The average correct performance scored with respect to the visual stimulus. Individual panels corresponding to the results for the 16 different bimodal stimuli. The visual stimulus (V) is indicated at the top, the auditory stimulus (A) is indicated to the right. The abscissa is the desynchronization between the auditory and visible syllables in ms. Negative values indicate audition leads. The percentage of responses that matched the presented visual stimulus is given on the ordinate. The results are pooled over natural and synthetic auditory speech stimuli. In the graphs for the matching audiovisual conditions the performance in the visual-alone conditions is indicated by the dashed straight lines. Averaged results from ten subjects in experiment 1.

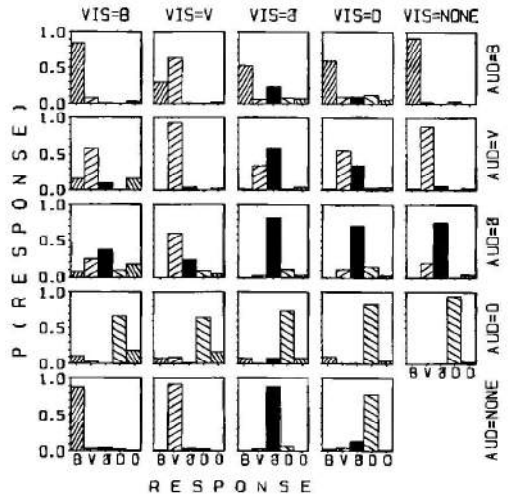

FIG. 6. Graphs of the proportion of responses as a function of the auditory and visual conditions for bimodal, auditory alone, and visual alone conditions. Proportions are given for the responses /b/, /v/, /ð/, /d/, and cluster responses (O). The columns correspond to the visual conditions and the rows to the auditory. Averaged results from ten subjects in experiment 1.

effect than the auditory speech in a few conflicting conditions. For example, visual /va/ paired with auditory /ba/ produced more responses that matched the visual than matched the auditory syllable. A visual advantage also occurred when visual /va/ was paired with auditory /ða/ and when visual /ða/ was paired with auditory /va/.

It appears that the range of asynchrony used in the first experiment did not have much effect on identification performance. Looking at the separate graphs in Figs. 4 and 5 for both the matching and conflicting audiovisual conditions, we see that the proportions of responses did not change much as a function of SOA. Although there seems to be a hint of an effect in the cases where visual /va/ was paired with auditory /ba/ and where visual /ða/ was paired with auditory /va/, there is no significant main effect of desynchronization conditions, $F(6,54)=1.45$, $p=0.21$.

In addition to the two types of accuracy scores, the response confusions are also of interest. Figure 6 gives the proportion of responses across the 24 conditions. The plots are again pooled over natural and synthetic auditory speech and also pooled across SOA. The cluster responses were grouped into a single category (labeled O in Fig. 6). These results show that responses do not simply agree with one of the syllables or the other. For example, a visual /da/ paired with an auditory /va/ gave 29% /ða/ judgments.

As noted earlier, 8% of the responses were consonant clusters. Consonant clusters appeared mainly when visual /ba/ was paired with either auditory /va/, /ða/, or /da/ (53% of the consonant clusters), and when visual /va/ or /ða/ was paired with auditory /da/ (26%). For the cluster responses that matched the auditory and visual syllables, the visual consonant nearly always came first (81%). The small number of clusters might appear unusual but Repp *et al.* (1983) and Massaro and Cohen (in press) have found similar results with these 4 stimulus alternatives.

III. RESULTS: EXPERIMENT 2

The analysis of experiment 2 was similar to that done for experiment 1. Figure 7 presents accuracy performance scored with respect to the auditory speech stimulus. As can be seen in the four graphs along the negative diagonal, auditory accuracy with consistent audiovisual speech was better

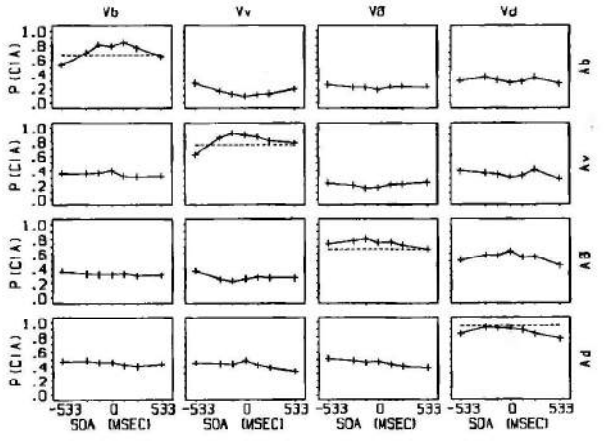

FIG. 7. The average correct performance scored with respect to the auditory stimulus. Individual panels corresponding to the results for the 16 different bimodal stimuli. The visual stimulus (V) is indicated at the top, the auditory stimulus (A) is indicated to the right. The abscissa is the desynchronization between the auditory and visible syllables in ms. Negative values indicate audition leads. The percentage of responses that matched the presented auditory stimulus is given on the ordinate. The results are pooled over natural and synthetic auditory speech stimuli. In the graphs for the matching audiovisual conditions the performance in the auditory-alone conditions is indicated by the dashed straight lines. Averaged results from 18 subjects in experiment 2.

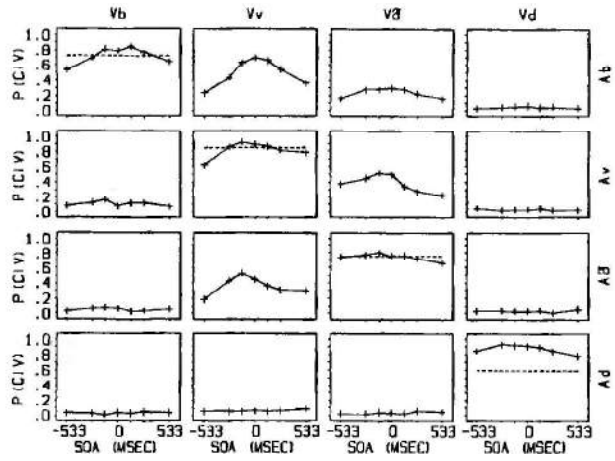

FIG. 8. The average correct performance scored with respect to the visual stimulus. Individual panels corresponding to the results for the 16 different bimodal stimuli. The visual stimulus (V) is indicated at the top, the auditory stimulus (A) is indicated to the right. The abscissa is the desynchronization between the auditory and visible syllables in ms. Negative values indicate audition leads. The percentage of responses that matched the presented visual stimulus is given on the ordinate. The results are pooled over natural and synthetic auditory speech stimuli. In the graphs for the matching audiovisual conditions the performance in the visual-alone conditions is indicated by the dashed straight lines. Averaged results from 18 subjects in experiment 2.

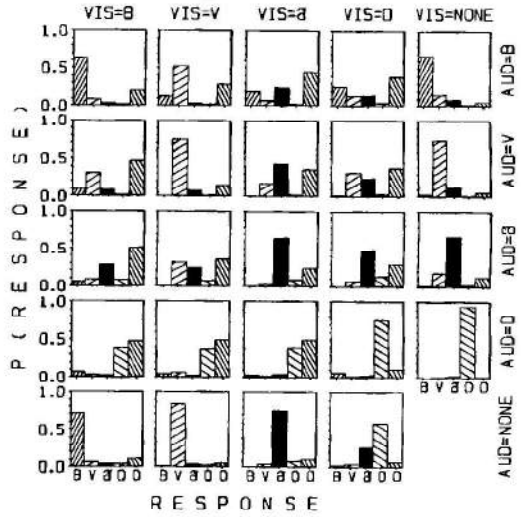

FIG. 9. Graphs of the proportion of responses as a function of the auditory and visual conditions for bimodal, auditory alone, and visual alone conditions. Proportions are given for the responses /b/, /v/, /ð/, /d/, and cluster responses (O). The columns correspond to the visual conditions and the rows to the auditory. Averaged results from 18 subjects in experiment 2.

than with the unimodal auditory syllables, and this advantage differed for the four different syllables; for /b/ (0.726 vs 0.668), /v/ (0.820 vs 0.748), and /ða/ (0.744 vs 0.664), and for /d/ (0.874 vs 0.949), $F(3,51)=8.32$, $p<0.001$.

As in experiment 1, Fig. 7 also shows that inconsistent visual information decreased auditory accuracy relative to the unimodal auditory condition. Comparing the curves across each row, auditory accuracy is lower in the three inconsistent conditions relative to the unimodal auditory condition.

As in experiment 1, performance on the unimodal auditory speech was somewhat better for natural than for synthetic speech (83% vs 68%), $F(1,17)=20.23$, $p=0.001$.

Figure 8 gives performance with respect to accuracy on the visual speech syllable. Identification of the visible speech averaged 73% correct, with the poorest performance (59%) on /da/. As can be seen in the four graphs along the negative diagonal in Fig. 8, performance with consistent audiovisual speech was better than the unimodal visual syllable only for /da/.

Figure 8 also shows that inconsistent auditory information decreased visual accuracy relative to the unimodal visual condition. As in experiment 1, this result occurred for all syllables and for all combinations.

In contrast to experiment 1, SOA had a large effect on identification performance. Looking at the separate graphs in Figs. 7 and 8 for both the matching and conflicting audiovisual conditions, we see that accuracy decreased significantly as SOA became positive or negative, $F(6,102)=2.31$, $p=0.039$ for auditory accuracy, and $F(6,102)=19.04$, $p<0.001$ for visual accuracy.

Figure 9 gives the proportion of responses across the 24 conditions. In experiment 2, 35% of the responses were consonant clusters. Consonant clusters occurred in all of the bimodal conditions, even when the two modalities were consistent. Cluster responses increased to the extent that the two syllables were asynchronous, $F(6,102)=8.97$, $p<0.001$.

A. Tests of the FLMP

The FLMP was tested against the individual subject data of each experiment for five responses: the four syllables plus the cluster category. The quantitative predictions of the model were determined by using the program STEPIT (Chandler, 1969). A model is represented to the program in terms of a set of prediction equations and a set of unknown parameters. By iteratively adjusting the parameters of the model, the program minimizes the squared deviations between the observed and predicted points. The outcome of the program STEPIT is a set of parameter values which, when put into the model, come closest to predicting the observed results. Thus STEPIT maximizes the accuracy of the description of a given model. The goodness-of-fit of a model is given by the root mean square deviation (rmsd), the square root of the average squared deviation between the predicted and observed values.

The FLMP was fit to each desynchronization condition separately. The same unimodal trials were used in the fit at each of the seven unique SOA conditions. Thus each fit for each SOA and each subject involved 32 bimodal and 12 unimodal trials, with five response categories for a total of 220 data points. The fit of the FLMP requires four visual and eight auditory parameter values for each of the five response categories (Massaro *et al.*, 1993, 1995). These values indicate the degree to which each unimodal source of information supports each response alternative. Thus the FLMP requires a total of 12 times 5 or 60 parameters.

Given our previous results (e.g., Massaro and Cohen, 1995), we expect the FLMP to give a good description at short SOAs. If integration breaks down at longer SOAs, the

FLMP should correspondingly give a poorer description. Because we are using the fit of the FLMP as a measure of integration, it is necessary to determine how good the fit is in an absolute sense. A benchmark measure has been developed to provide this index of goodness of fit of a model (Massaro and Cohen, 1993). Even if a model is perfectly correct, we cannot expect it to fit results perfectly because a probabilistic prediction is necessarily associated with some expected variability. This expected variability depends on the number of response alternatives, the number of observations and the response probability. Given that we grouped all of the cluster responses into a single category, the present task can be analyzed as having five response alternatives. There were 42 observations on unimodal trials and six on bimodal trials. Given these values, the expected variability can be computed. We can therefore ask if the fit of a model is poorer than this expected variability, which is called a benchmark RMSD and is is the expected RMSD if the model is correct.

One question of central interest is whether the goodness of fit of the FLMP differed across the different SOAs. A straightforward evaluation of the goodness of fit of the FLMP to the different SOA conditions would be to compare their respective RMSD values. However, the different SOA conditions led to different proportions of identification, and therefore different benchmark RMSDs. Therefore, each observed RMSD must be compared to its corresponding benchmark.

A benchmark RMSD was computed for each participant by the following procedure. The observations for each subject were used to generate a simulated subject by replicating the conditions of the present experiment. For example, the response proportions for a given subject to the natural auditory /ba/ might be 0.823 /ba/ responses and 0.177 /va/ responses. The value for this condition for a simulated subject would be determined by sampling with these proportions on 56 trials (the number of occurrences of this condition in the experiment). A random number between 0 and 1 would be chosen. If the number was 0.823 or lower, a /ba/ response would be counted; 0.824 to 1.000 would be counted as /va/. This procedure was carried out for each condition to create a set of simulated results. The original observations of the subject were then compared to the results of this simulated subject to compute a simulated RMSD. This exact procedure was repeated 20 times to create 20 simulated subjects for each real subject. Twenty subjects were simulated to give a reliable estimate of the simulated RMSD. The average of these 20 simulated RMSDs gives a average simulated RMSD for the original empirical subject. Finally, this simulated RMSD has to be adjusted to take into account the fact that the FLMP has 60 free parameters to predict the 220 independent data points at each SOA. For every free parameter, one of the data points could be predicted exactly. Thus 60 data points could in principle have a zero RMSD, and the remaining ones should have only the simulated variability. Following this statistical reasoning, the benchmark RMSD should be (220-60)/220 or 8/11ths of the simulated RMSD. This procedure was carried out for each empirical subject to give a benchmark RMSD to compare with the original observed RMSD.

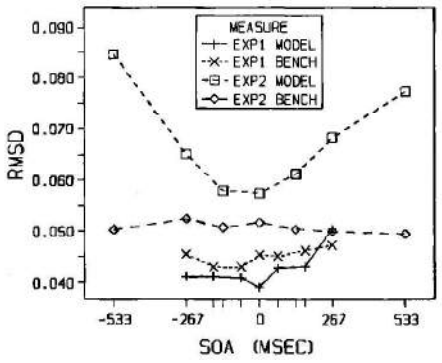

FIG. 10. The average RMSD for the fit of the FLMP and the average benchmark RMSD as a function of SOA in experiments 1 and 2.

Figure 10 gives the average observed and average benchmark RMSDs for experiments 1 and 2. First, note that the benchmark RMSDs are smaller for experiment 1 than for experiment 2. This difference reflects the fact that the responses were more variable and less extreme in experiment 1 than in experiment 2. The benchmark RMSDs make it easier to compare results across experimental conditions with necessarily different levels of performance.

For each experiment, an analysis of variance was carried out comparing the observed to the the benchmark RMSD with SOA and subjects as factors. In experiment 1 with short SOAs, there was no difference between the observed and benchmark RMSDs and no effect of SOA. The bottom two curves in Fig. 10 show that perhaps there was some minor disruption with integration at the 267 ms SOA. Given that the goodness of fit changed very little across SOA, we would conclude that integration of audible and visible speech can occur even with SOAs as large as 1/4 s. Testing longer SOAs in experiment 2, however, not only indicates when the good fit can break down, it also shows that the results might be somewhat context dependent. In experiment 2 with longer SOAs, there was a difference between the observed and benchmark RMSDs and this effect interacted with SOA. As can be seen in the top two curves of Fig. 10, the FLMP gave a poorer fit at the more extreme SOAs. The analysis of variance showed a significant difference between the observed and benchmark RMSDs, $F(1,17)=26.09$, $p<0.001$; and a significant effect of SOA, $F(6,102)=6.35$, $p<0.001$. Most importantly, the significant interaction between these two variables, $F(6,102)=9.01$, $p<0.001$, documents that the goodness-of-fit of the FLMP deteriorated with more extreme SOAs.

One potentially troubling result is that the overall performance on both auditory and visual unimodal trials was significantly poorer in experiment 2 than in experiment 1. In principle, the longer SOAs in experiment 2 should not have changed performance on unimodal trials. However, there are very large individual differences in these speech perception tasks, and we are reluctant to make much of an overall difference in performance accuracy across the two experiments. To assure that this overall performance difference was not responsible for the different effects of SOA in the two experiments, we separated the 18 subjects in experiment 2 into high and low accuracy groups. The high accuracy group

achieved the same level of performance accuracy as the subjects in experiment 1. The statistical analyses and tests of the FLMP led to the same conclusions with the high accuracy group as was reached for all 18 subjects in experiment 2.

The combined results of experiments 1 and 2 provide a reasonably coherent answer to the temporal window of integrating auditory and visual speech. One-half of a second is clearly disruptive, whereas integration does not appear to be disrupted at around 150 ms.

Finally, there might be an influence of the test context in that repeated tests with extreme SOAs might lead to some disruption of integration even at SOAs that would normally produce integration. Figure 10 shows that observed RMSDs at the short SOAs were slightly smaller than their benchmarks in experiment 1 whereas they were somewhat larger than their benchmarks in experiment 2. It appears that subjects are able to integrate audible and visible speech more easily if most of the bimodal speech events have fairly short SOAs. On the other hand, having bimodal speech events with very long SOAs might disrupt integration to some extent regardless of the SOA.

IV. DISCUSSION

The present study was carried out to determine to what extent asynchrony between auditory and visual speech would disrupt their integration. Variations in the temporal onset of a visual and an auditory CV syllable, be it congruent or incongruent, can dramatically influence their integration. On the other hand, the integration process appears to be somewhat flexible. Embedding very long SOAs in the task can lead to a disruption of integration at shorter SOAs that would produce integration in the context of short SOAs. In experiment 1 with SOAs within a quarter of a second, the FLMP gave a good description of the results across all SOAs. One might conclude that integration appears to take place even if the two sources of information are offset by a quarter of a second. In experiment 2 with some SOAs over a half a second, however, integration was disrupted at some of the SOAs that gave integration in experiment 1.

The breakdown of integration at an SOA of 500 ms is consistent with some findings of Reisberg *et al.* (1987). Their participants shadowed a difficult philosophical passage taken from Kant. The percentage of words correctly shadowed was somewhat larger when the full face of the talker was in view than when only sound was available. This advantage of having visible speech disappeared when the audible and visible speech were offset by about 0.5 s.

One might view temporal asynchronies between the auditory and visual speech as an unusual situation. However, even when these two inputs are perfectly correlated in the natural world, they cannot be simultaneous in terms of their perceptual processing. Given the physics of auditory and visual stimuli and the physiology of their respective sensory systems, they cannot arrive at the relevant processing sites at the same time. Light travels faster than sound and will arrive at the sensory surface sooner. For stimuli at 10 m, the light arrives about 30 ms before the sound. Because the chemical process of the retina transducing light is slower than the basilar membrane transducing sound, we can expect a faster neural reaction for sound than light. Also it is well known that the physiological travel time to neural centers varies inversely with stimulus intensity (Kohfeld, 1971). For example, McGill (1961) observed that increasing the amplitude of a test tone from 30 to 100 dB SPL decreased simple reaction time by about 100 ms.

We might ask how information from the two modalities can be integrated when the inputs can have relatively large differences in arrival times at the appropriate neural sites. The integration of auditory and visual sources seems to occur even though their relative onset times of neural activation can differ by tens or even hundreds of milliseconds. The sensory systems seemed to have solved the problem of integrating multisensory stimuli arriving at different times by extending the neural activation resulting from stimulation beyond the stimulation period. Thus stimuli that co-occur at roughly the same time in the world will tend to have overlapping activation patterns. This observation meshes with the concept of a sensory storage that has played an important role in information processing theory for the last three decades (Massaro and Loftus, in press). This storage, estimated at roughly 250 ms, allows both auditory and visual processing to continue after the relevant stimulation is removed. Given this storage, information inputs from several modalities can be integrated even though the two inputs have different neural arrival times. This integration occurs with fairly discrepant onset asynchronies of up to roughly 150–250 ms.

ACKNOWLEDGMENTS

This research was supported, in part, by grants from the Public Health Service (PHS R01 DC 00236), the National Science Foundation (BNS 8812728), the University of California, Santa Cruz, and from the Netherlands Organization for Scientific Research (NWO) and the Delft University Fund (DUF). The authors are grateful for the conscientious review of three anonymous reviewers and their convincing argument for the second study with longer SOAs.

Bernstein, L. E. and Eberhardt, S. P. (**1986**). *Johns Hopkins Lipreading Corpus I-II: Disc I* (The Johns Hopkins University, Baltimore).

Binnie, C. A., Montgomery, A. A., and Jackson, P. L. (**1974**). "Auditory and visual contributions to the perception of selected English consonants for normally hearing and hearing-impaired listeners," in *Visual and Audio-visual Perception of Speech*, edited by H. Birk Nielsen and E. Kampp, Scand. Audiol. Suppl. **4**, 181–209.

Braida, L. D. (**1991**). "Cross-modal integration in the identification of consonant segments," T. Q. J. Exp. Psychol. **43A**, 647–677.

Breeuwer, M., and Plomp, R. (**1984**). "Speechreading supplemented with frequency-selective sound-pressure information," J. Acoust. Soc. Am. **76**, 686–691.

Campbell, R., and Dodd, B. (**1980**). "Hearing by eye," Q. J. Exp. Psychol. **32**, 85–99.

Chandler, J. P. (**1969**). "Subroutine STEPIT—Finds local minima of a smooth function of several parameters," Behav. Sci. **14**, 81–82.

Cohen, M. M., and Massaro, D. W. (**1990**). "Synthesis of visible speech," Behav. Res. Methods, Instrum. Comput. **22**, 260–263.

Cohen, M. M., and Massaro, D. W. (**1993**). "Modeling coarticulation in synthetic visual speech," in *Models and Techniques in Computer Animation*, edited by N. M. Thalmann and D. Thalmann (Springer-Verlag, Tokyo), pp. 139–156.

Cohen, M. M., and Massaro, D. W. (**1994**). "Development and experimentation with synthetic visible speech," Behav. Res. Methods Instrum. Comput. **26**, 260–265.

Erber, N. P. (1972). "Auditory, visual, and auditory-visual recognition of consonants by children with normal and impaired hearing," J. Speech Hear. Res. **15**, 423–422.

Gouraud, H. (1971). "Continuous shading of curved surfaces," IEEE Trans. Comput. **C-20**(6), 623–628.

Green, K. P., and Kuhl, P. K. (1989). "The role of visual information in the processing of place and manner features in speech perception," Percept. Psychophys. **45**, 34–42.

Klatt, D. H. (1980). "Software for a cascade/parallel formant synthesizer," J. Acoust. Soc. Am. **67**, 971–995.

Koenig, E. (1965). (Cited by Pandey *et al.*)

Kohfeld, D. L. (1971). "Simple reaction time as a function of stimulus intensity in decibels of light and sound," J. Exp. Psychol. **88**, 251–257.

Massaro, D. W. (1985). "Attention and preception: An information-integration perspective," Acta Psychol. **60**, 211–241.

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

Massaro, D. W. (1989). "Multiple Book Review of *Speech Perception by Ear and Eye: A Paradigm for Psychological Inquiry*," Behav. Brain Sci. **12**, 741–794.

Massaro, D. W. (1990). "A Fuzzy logical Model of Speech Perception," *Proceedings of the XXIV International Congress of Psychology*, in *Human Information Processing: Measures, Mechanisms, and Models*, edited by D. Vickers and P. L. Smith (North-Holland, Amsterdam), pp. 367–379.

Massaro, D. W. (1992). "Broadening the domain of the fuzzy logical model of perception," in *Cognition, Conceptual, and Methodological Issues*, edited by H. L. Pick, Jr., P. Van den Broek, and D. C. Knill (American Psychological Association, Washington, DC), pp. 51–84.

Massaro, D. W., and Cohen, M. M. (1983). "Evaluation and integration of visual and auditory information in speech perception," J. Exp. Psychol. Human Percept. Perform. **9**, 753–771.

Massaro, D. W., and Cohen, M. M. (1990). "Perception of synthesized audible and visible speech," Psychol. Sci. **1**, 55–63.

Massaro, D. W., and Cohen, M. M. (1993). "Perceiving asynchronous bimodal speech in consonant-vowel and vowel syllables," Speech Commun. **13**, 127–134.

Massaro, D. W., and Cohen, M. M. (1995). "Perceiving talking faces," Curr. Directions Psychol. Sci. **3**, 104–109.

Massaro, D. W., and Cohen, M. M. (**in press**). "Perceiving speech from inverted faces," Percept. Psychophys.

Massaro, D. W., and Loftus, G. (**in press**). "Sensory and perceptual storage: Data and theory," in *Handbook of Perception and Cognition*, edited by E. L. Bjork and R. A. Bjork (Academic, San Diego), Vol. 10.

Massaro, D. W., Tsuzaki, M., Cohen, M. M., Gesi, A., and Heredia, R. (1993). "Bimodal speech perception: An examination across languages," J. Phon. **21**, 445–478.

Massaro, D. W., Cohen, M. M., and Smeele, P. M. T. (1995). "Cross-linguistic comparisons in the integration of visual and auditory speech," Mem. Cognit. **23**, 113–131.

McGill, W. J. (1961). "Loudness and reaction time," Acta Psychol. **19**, 193–199.

McGrath, M., and Summerfield, Q. (1985). "Intermodal timing relations and audio-visual speech recognition by normal-hearing adults," J. Acoust. Soc. Am. **77**, 678–685.

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

Mills, A. E., and Thiem, R. (1980). "Auditory-visual fusions and illusions in speech perception," Linguistische Berichte **68/80**, 85–108.

Montgomery, A. A., and Jackson, P. L. (1983). "Physical characteristics of the lips underlying vowel lipreading performance," J. Acoust. Soc. Am. **73**, 2134–2144.

Munhall, K. G., Gribble, P., Sacco, L. and Ward, M. (1996). "Temporal constraints on the McGurk effect," Percept. Psychophys. **58**, 351–362.

Pandey, P. C., Kunov, H., and Abel, S. M. (1986), "Disruptive effects of auditory signal delay on speech perception with lipreading," J. Aud. Res. **26**, 27–41.

Parke, F. I. (1975). "A model for human faces that allows speech synchronized animation," Comput. Graphics J. **1**(1), 1–4.

Parke, F. I. (1982). "Parameterized models for facial animation," IEEE Comput. Graphics **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 Graphics Interface '86*, pp. 136–140.

Reisberg, D., McLean, J., and Goldfield, A. (1987). "Easy to hear but hard to understand: A lip-reading advantage with intact auditory stimuli," in *Hearing by Eye: The Psychology of Lip-Reading*, edited by B. Dodd and R. Campbell (Erlbaum, London), pp. 97–113.

Repp, B. H., Manuel, S. Y., Liberman, A. M., and Studdert-Kennedy, M. (1983). "Exploring the 'McGurk Effect'," paper presented at the 24th meeting of the Psychonomic Society, San Diego, California, November 1983, Bull. Psychonom. Soc., 358 (abstract).

Risberg, A., and Lubker, J. L. (1978). "Prosody and speechreading," Report STL-OPSR 4/78, Dept. of Speech Communication and Musical Acoustics, Royal Institute of Technology, Stockholm, 1–16.

Sekiyama, K., and Tohkura, Y. (1991). "McGurk effect in non-English listeners: Few visual effects for Japanese subjects hearing Japanese syllables of high auditory intelligibility," J. Acoust. Soc. Am. **90**, 1797–1805.

Smeele, P. M. T., and Sittig, A. C. (1990). "The contribution of vision to speech perception," Proceedings of the 13th International Symposium on Human Factors in Telecommunications, Torino, 525.

Smeele, P. M. T., and Sittig, A. C. (1991). "The contribution of vision to speech perception," Proceedings of the 2nd European Conference on Speech Communication and Technology, Eurospeech 91, Genova, 1495–1497.

Sumby, W. H., and Pollack, I. (1954). "Visual contribution to speech intelligibility in noise," J. Acoust. Soc. Am. **26**, 212–215.

Summerfield, A. Q. (1979). "Use of visual information in phonetic perception," Phonetica **36**, 314–331.

Thompson, L. A., and Massaro, D. W. (1989). "Before you see it, you see its parts: Evidence for feature encoding and integration in preschool children and adults," Cognitive Psychol. **21**, 334–362.

Vroomen, J. H. M. (1992). "Hearing voices and seeing lips: Investigations in the psychology of lipreading," Dissertation, Katholieke Universiteit Brabant.