Perception of Synthesized Audible and Visible Speech

Dominic W. Massaro and Michael M. Cohen

Psychological Science 1990 1: 55
DOI: 10.1111/j.1467-9280.1990.tb00068.x

Research Article

PERCEPTION OF SYNTHESIZED AUDIBLE AND VISIBLE SPEECH

Dominic W. Massaro and Michael M. Cohen

Program in Experimental Psychology, University of California, Santa Cruz

Abstract—*The research reported in this paper uses novel stimuli to study how speech perception is influenced by information presented to ear and eye. Auditory and visual sources of information (syllables) were synthesized and presented in isolation or in factorial combination. A five-step continuum between the syllables /ba/ and /da/ was synthesized along both auditory and visual dimensions, by varying properties of the syllable at its onset. The onsets of the second and third formants were manipulated in the audible speech. For the visible speech, the shape of the lips and the jaw position at the onset of the syllable were manipulated. Subjects' identification judgments of the test syllables presented on videotape were influenced by both auditory and visual information. The results were used to test between a fuzzy logical model of speech perception (FLMP) and a categorical model of perception (CMP). These tests indicate that evaluation and integration of the two sources of information makes available continuous as opposed to just categorical information. In addition, the integration of the two sources appears to be nonadditive in that the least ambiguous source has the largest impact on the judgment. The two sources of information appear to be evaluated, integrated, and identified as described by the FLMP—an optimal algorithm for combining information from multiple sources. The research provides a theoretical framework for understanding the improvement in speech perception by hearing-impaired listeners when auditory speech is supplemented with other sources of information.*

INTRODUCTION

Speech perception is a human skill that rivals our other impressive achievements. People manage to communicate under the most adverse conditions imaginable. The robustness of human communication contrasts sharply with that of machines. Even after decades of intense effort, speech recognition by machine remains far inferior to human performance. One assumption underlying the present research is that human recognition of speech is robust because there are usually multiple sources of information supporting speech perception, which the perceiver evaluates and integrates to achieve perceptual recognition.

The goal of the present research is to determine how auditory and visual information together influence speech perception. We use the experimental method of manipulating these sources of information using synthetic speech and observing corresponding changes in perceptual judgments. In previous studies, the inability to synthesize visible speech continua has limited the power of experimental contrasts. In the present study, synthetic auditory and visual speech are manipulated in an expanded factorial design. Different auditory syllables are combined factorially with different visual syllables, and the design is expanded to include the unimodal presentation of each syllable. The same number of different auditory and visual syllables is used because the strongest tests of mathematical models of performance consist of expanded factorial designs in which the independent variables have the same number of levels. These designs have the largest number of independent observations to be predicted relative to the number of free model parameters that must be estimated from the data. In a successful theoretical account of results from an expanded factorial design we must describe how the identification of each bimodel syllable occurs as a function of the identifications of the unimodal syllables that compose it. Before presenting the empirical study, we will adduce evidence that visible speech is an influential source of information in face-to-face communication, and review how visible speech can compensate for hearing loss. We also describe how visual information can influence normal speech perception even when the auditory information is well-specified, and summarize the types of information provided by visible speech.

Correspondence and reprint requests to Dominic W. Massaro, Program in Experimental Psychology, University of California, Santa Cruz, CA 95064.

SPEECH PERCEPTION BY EYE

Hearing impairment is a more prevalent problem than most of us realize, with an estimated 21.1 million individuals (9% of the population) in the United States affected to some degree (Hotchkiss, 1987). Those affected (as of a decade ago) include 841,000 youth 6 to 17 years of age (Ries, 1987). Both absolute sensitivity to undegraded speech and the ability to hear speech in noisy environments decline systematically with aging (National Research Council, 1987). Some, but not all, of this decline can be accounted for by a loss of hearing with age. Unfortunately, there are limitations in the degree to which hearing can be restored with hearing aids, primarily because hearing aids amplify all sound, not just the critical speech sounds.

Luckily, loss of hearing can be compensated for by other sources of information in speech perception. Watching a speaker's face and lips provides powerful information in speech perception and language understanding. Experiments have shown that this visible speech is particularly important when the auditory speech is degraded by noise, bandwidth filtering, or hearing-impairment. Summerfield (1979), for example, found that subjects recognized only 23% of the sentences presented in a

noisy environment consisting of a continuous prose background, while their accuracy increased to 65% with a view of the speaker's face. In another study, Breeuwer and Plomp (1984) found that the perception of bandpass filtered short sentences improved from 23% to 79% correct when subjects were also permitted a view of the speaker. For hearing-impaired adults, lipreading the speaker improved consonant recognition from 55% to 80% correct (Walden, Prosek, & Worthington, 1974). Thus, visual information from the speaker's face can compensate for a lack of auditory information in language processing.

The strong influence of visible speech is not limited to situations with degraded auditory input, however, and has an important influence even when paired with perfectly intelligible speech sounds. We have all noticed the discrepancy of sight and sound in dubbed movies, but a modification of this situation by McGurk and MacDonald (1976) illustrates the power of visible speech. A videotape of a person making a visible labial articulation /pa-pa/ was dubbed with the alveolar nasal speech sounds /na-na/. This dubbed speech event gives a situation in which intelligible auditory speech is paired with a contradictory visual articulation. Even though subjects were asked to indicate what they heard, a strong effect of the visual source of information was observed, with subjects often reporting hearing the labial nasal /ma-ma/. A perceiver's recognition of an auditory-visual syllable reflects the contribution of both sound and sight. This surprising perceptual experience has come to be known as the McGurk effect.

Analogous to the perception of auditory speech by hearing-impaired individuals, visible speech alone is not sufficient for comprehensive perception. Although sound is potentially adequate for discriminating all distinctions in speech, only a subset of these distinctions is carried by visible speech. Information about place of articulation and duration are visible to some extent, whereas voicing is not.[1] Considerable research effort has uncovered the contrasts in speech that can be conveyed by visual information (MacDonald & McGurk, 1978; McGurk & MacDonald, 1976; Summerfield, 1979, 1983; Walden, Prosek, Montgomery, Scherr, & Jones, 1977).

The classes of consonant speech gestures that can be discriminated are called *visemes*, analogous to the *phonemes* of audible speech. Adults without any previous training usually can distinguish six categories of consonants. In a typical study (Walden, Prosek, Montgomery, Scherr & Jones, 1977), the six viseme categories that could be discriminated were [θð, fv, pbm, sz$\int^{v}_{z}$, w, and l], based on a criterion of 72% judgments from within the viseme category. Training improves the discrimination and identification of visible speech. After 14 hours of training in lipreading, the number of discriminable viseme categories increased from six to nine. The nine viseme categories were {θð, fv, $\int^{v}_{z}$, sz, pbm, tdnkgj, w, r, and l].

Compared to consonants, it is easier to articulate the same vowel with different vocal tract configurations (Ladefoged, Harshman, Goldstein & Rice, 1978). Vowels have consistent visible information, however, and this information is used in speech perception by eye (Summerfield, 1983). In one study, observers lipread 15 vowels and diphthongs in the context /h-g/, produced by four female talkers (Montgomery & Jackson, 1983). All vowels were recognized at better than chance accuracy. An analysis of the recognition confusions was used to determine what visible characteristics are functional in lipreading. One property appeared to be the opening of the mouth and the vertical separation of the lips. This dimension provides information about the height of the tongue in the mouth. A second property was the degree of lip spreading/rounding, which is a visible dimension correlated with the part of the tongue that is raised or lowered. The vowels in English are fairly discriminable from one another on the basis of these two properties.

The value of visible speech in speech perception warrants its study, in the same way that auditory speech has been studied. Although progress has been made using natural speech as stimuli, the computer synthesis of a speaker's face permits better-controlled and more systematic analysis of the perceptual process. Synthetic speech has been valuable, if not indispensable, for the study of auditory speech perception; and the study of visible speech perception should be no different. In addition to achieving exact control over the speech stimulus, synthetic speech allows the creation of novel speech segments that are not easily produced naturally. Presenting these novel segments to subjects in psychophysical tasks provides important information about the processes involved in speech perception. Before further describing the new synthetic visual speech we will describe the synthetic audible speech used in the present study.

SYNTHETIC AUDIBLE SPEECH

Tokens of the first author's /ba/ and /da/ were analyzed using linear prediction to derive a set of parameters for driving a software formant serial resonator speech synthesizer (Klatt, 1980). By altering the parametric information specifying the first 80 msec of the consonant-vowel syllable, a set of five 400 msec syllables covering the range from /ba/ to /da/ was created. The center and lower panels of Figure 1 show how some of the acoustic synthesis parameters changed over time for the most /ba/-like and /da/-like of the 5 auditory syllables. During the first 80 msec, the first formant (F1) went from 250 Hz to 700 Hz following a negatively accelerated path. The F2 followed a negatively accelerated path to 1199 Hz, beginning with one of five values equally spaced between 1187 and 1437 Hz from most /ba/-like to most /da/-like, respectively. The F3 followed a linear transition to 2729 Hz from one of five values equally spaced between 2387 and 2637 Hz. All other stimulus characteristics were identical for the five auditory syllables. Figure 2 gives the spectrograms of the five syllables along the continuum.

SYNTHETIC VISIBLE SPEECH

Several forms of stimulated facial display have been used for speech studies. Relatively simple Lissajou's figures have been displayed on an oscilloscope to simulate lip movement, using

1. Loosely speaking, place of articulation refers to the point in the oral cavity that is the most constricted during articulation; duration is the temporal duration of the consonant segment; and voicing describes vibration of vocal cords.

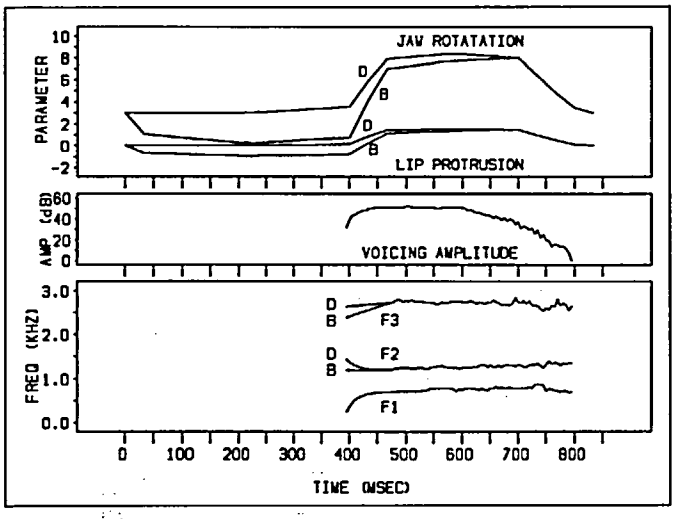

Fig. 1. Visual and auditory parameter values over time for visual /ba/ and /da/ stimuli and auditory /ba/ and /da/ stimuli. Bottom panel shows formants F1, F2, and F3, middle panel shows voicing amplitude, and top panel shows jaw rotation and lip protrusion. See text for details.

analog control voltages to vary the height and width of the simulated lips (Erber & De Filippo, 1978). A model for lip shape was developed which allowed computation of coarticulatory effects for CVCVC segments (Montgomery, 1980). The lip shape display was done on a vector graphic device using about 130 vectors at a rate of about 4 times real time. A real-time vector display system for displaying simple 2-dimensional faces has also been reported (Brooke & Summerfield, 1983).

To generate more realistic full facial displays, two general strategies have been employed: musculoskeletal models (Platt & Badler, 1981), and parametrically controlled polygon topology (Parke, 1974). The latter strategy was used in a fairly realistic animation by modeling the facial surface as a polyhedral object composed of about 900 small surfaces arranged in 3D, joined together at the edges (Parke, 1974, 1975, 1982). The left side of Figure 3 shows a framework rendering of this model. To achieve a natural appearance, the surface is smooth shaded using Gouraud's (1971) method (shown in the right panel of Figure 3). The face is animated by altering the location of various points in the grid under the control of 50 parameters, 11 of which are used for speech animation. Control parameters used for several demonstration sentences were selected and refined by the investigator by studying his own articulation frame by frame and estimating the control parameter values (Parke, 1974).

Recently, this software and facial topology has been translated from the original JOVIAL language to C and given new speech- and expression-control routines (Pearce, Wyvill, Wyvill & Hill, 1986). In this system, a user can type a string of phonemes which are then converted to control parameters which are changed over time to produce the desired animation sequence. Each phoneme is defined in a table according to target values for segment duration, segment type (stop, vowel, liquid, etc.) and 11 control parameters. The parameters that are used are 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, upper lip raise, and x and z teeth offset. The revised software of Pearce et al. (1986) was implemented by us on a Silicon Graphics IRIS 3030 computer. We have adapted the software to allow new intermediate test phonemes and written several output processors (pipes) for rendering the polygonal image information in different ways. One pipe produces wireframe images, a second produces Gouraud shaded images with a diffuse illumination model, a third also includes specular illumination (white highlights), and a fourth pipe uses tesselation (recursive polygon subdivision) for improved skin texture appearance as well as randomly determined hair. The diffuse pipe used in the present experiment now takes about 1 min to render and record each frame while the diffuse + specular rendering takes 3 min. This is a considerable improvement over a speed of about 15 min per frame for a diffuse illumination model previously reported by Pearce et al. (1986). To create an animation sequence, each frame was recorded

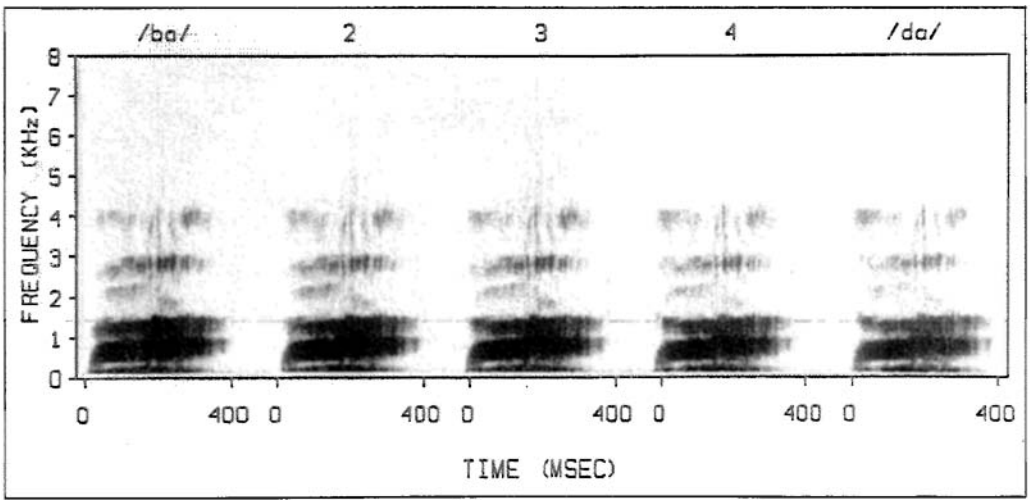

Fig. 2. Spectrograms for the five levels of auditory speech between /ba/ and /da/.

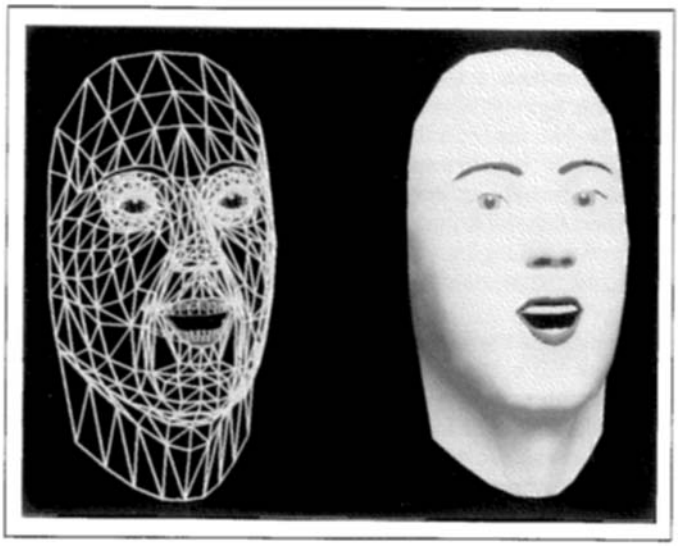

Fig. 3. Framework (left) and Gouraud shaded (right) renderings of polygon facial model.

using a broadcast quality Betacam video recorder under control of the IRIS.

In prototypical experiments on auditory-speech perception, some property of the speech stimulus is varied in small steps to give a continuum of speech sounds between two alternatives. For example, the onsets of the second and third formants can be varied to give an auditory continuum between the syllables /ba/ and /da/. In analogous fashion, we systematically varied parameters of the facial model to give a continuum between visual /ba/ and /da/. Figure 4 gives pictures of the facial model at the time of maximum stop closure for each of the five levels between /ba/ and /da/. Table 1 gives the parameter target values used in the visual synthesis for the consonant portion of each visual stimulus, the default resting parameter values, and the values for the vowel /a/. The top panel of Figure 1 shows how the visual synthesis parameters changed over time for the first (/ba/) and last (/da/) visual levels. For clarity, only two of the visual parameters are shown—jaw rotation (larger parameter means more open), and lip protrusion ("mouth z offset" in Table 1, smaller number means more protrusion). Not shown in the figure, the face with the default parameter values was recorded for 2000 msec preceding and 2000 msec following the time shown for a total visual stimulus of 4866 msec. An immobile face (for auditory alone trials) was also synthesized which kept the default values for the entire 4866 msec.

Following the synthesis the Betacam tape was dubbed to ¾" U-Matic for editing. Only the final 4766 msec of each video sequence was used for each trial. A tone marker was dubbed onto the audio channel of the tape at the start of each syllable to allow the playing of the 400 msec auditory speech stimulus just following consonant release of the visual stimulus. The marker tone on the video tape was sensed by a Schmidt trigger on a PDP-11/34A computer which presented the auditory stimuli from digitized representations on the computer's disk. Figure 1 shows the temporal relationship between the auditory and visual parts of the stimulus. As can be seen in the figure, the parameter transitions specifying the consonantal release occurred at about the same time for both modalities.

EXPERIMENTAL DESIGN

The ability to synthesize both audible and visible speech offers a potentially informative approach to the study of speech perception in face-to-face communication. Specifically, the question of whether and how multiple sources of information influence speech perception can be addressed. To study this question, we combined synthetic audible and visible speech in a novel expanded factorial design. Figure 5 illustrates a 5×5 expanded factorial design. Five levels of audible speech varying between /ba/ and /da/ were crossed with five levels of visible speech varying between the same alternatives. In addition, the audible and visible speech also were presented alone for a total of $25 + 5 + 5 = 35$ independent stimulus conditions. Six random sequences were determined by sampling the 35 conditions without replacement. Using this sequence, an edited ¾" videotape with six blocks of 35 trials was created from the original six tokens. This tape was then dubbed to ½" VHS videotape with the auditory syllables added by the computer.

We carried out an experiment using this design with seven college students as subjects. The subjects gave their informed consent after the nature of the experiment was explained to

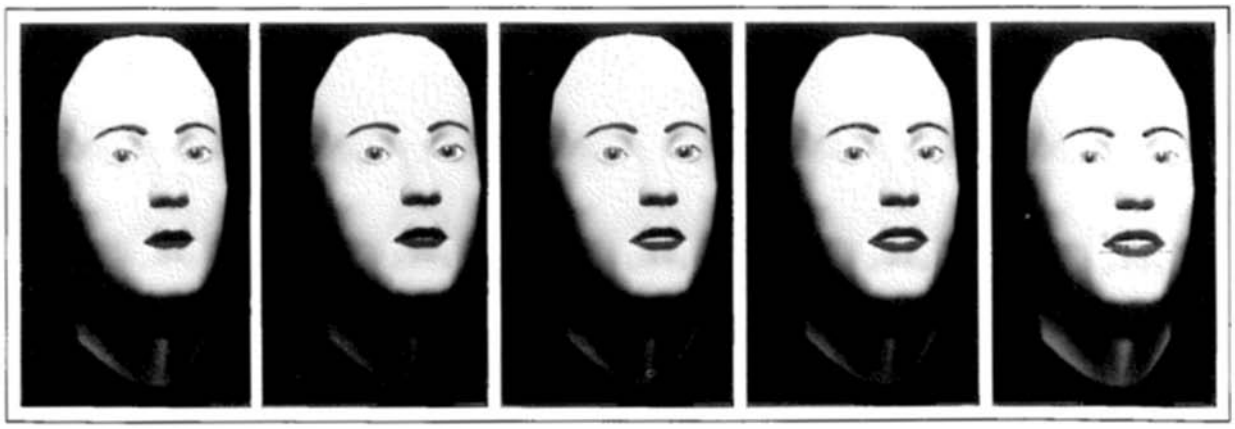

Fig. 4. The facial model at the onset of articulation for each of the five levels of visible speech between /ba/ and /da/.

Table 1. *Visual synthesis parameters for the default position, five stops, and /a/*

Parameter	Default	/b/	2	3	4	/d/	/a/
Jaw rotation	3.00	0.00	0.75	1.50	2.25	3.00	10.00
Mouth x scale	1.00	1.00	1.05	1.10	1.15	1.20	1.00
Mouth z offset	0.00	−1.00	−0.75	−0.50	−0.25	0.00	2.00
Lip corner x width	0.00	0.00	1.25	2.50	3.75	5.00	20.00
Mouth corner z offset	0.00	−15.00	−15.00	−15.00	−15.00	−15.00	0.00
Mouth corner x offset	0.00	2.00	4.50	7.00	9.50	12.00	0.00
Mouth corner y offset	0.00	0.00	0.75	1.50	2.25	3.00	−5.00
Lower lip 'f' tuck	0.00	−5.00	−5.00	−5.00	−5.00	−5.00	0.00
Upper lip raise	0.00	2.00	4.75	7.50	10.25	13.00	2.00

them. Subjects were instructed to listen and to watch the speaker, and to identify the syllable as /ba/, /da/, /bda/, /dba/, /ð/, /va/, /ga/, or "other." These response alternatives were determined from pilot studies in which the responses were not constrained. Each of the 35 possible stimuli were presented a total of 12 times during two sessions and the subject identified each stimulus during a 3 sec response interval.

RESULTS

The mean observed proportion of identifications for each of the possible responses across subjects is shown as the points in Figure 6. The results for only five responses are shown because the responses /dba/, /ga/, and "other" each accounted for no more than 5% of the judgments in all of the three presentation conditions. Both the auditory and visual sources of information had a strong impact on the identification judgments. As can be seen in Figure 6, the proportion of responses changed systematically across the visual continuum, both for the unimodal, $F(28, 168) = 18.16$, $p < .001$, and bimodal, $F(28, 168) = 17.00$, $p < .001$, conditions. Similarly, the pattern of responses changed in an orderly fashion across the auditory continuum, for both the unimodal, $F(28, 168) = 9.63$, $p < .001$, and bimodal, $F(28, 168) = 29.73$, $p < .001$, conditions. Finally, the auditory and visual effects were *not* additive, as indicated by the significant auditory-visual interaction on response probability, $F(112, 672) = 6.23$, $p < .001$, in the bimodal condition. The results will now be used to test between two contrasting models of speech perception.

FUZZY LOGICAL MODEL OF PERCEPTION (FLMP)

Within the present framework, speech perception is viewed as having available multiple sources of information supporting the identification and interpretation of the language input. The results from a wide variety of experiments can be described within a framework of a fuzzy logical model of perception (FLMP). The assumptions central to the model are (a) each source of information is evaluated to give the degree to which that source specifies various alternatives; (b) the sources of information are evaluated independently of one another; (c) the sources are integrated to provide an overall degree of support for each alternative; and (d) perceptual identification follows the relative degree of support among the alternatives (Massaro, 1987). The FLMP was tested against the results of the present experiment. Following the research strategy of strong inference (Platt, 1964), an alternative categorical model of speech perception (CMP) is also tested against the results.

According to the FLMP, well-learned patterns are recognized in accordance with a general algorithm, regardless of the modality or particular nature of the patterns. Three operations assumed by the model are illustrated in Figure 7. Continuously-valued features are evaluated, integrated, and matched against prototype descriptions in memory, and an identification decision is made on the basis of the relative goodness of match of the stimulus information with the relevant prototype descriptions.

Applying the FLMP to the present task using auditory and visual speech, both sources are assumed to provide continuous and independent evidence for the each of the response alternatives. Defining the onsets of the second (F2) and third (F3) formants as the important auditory feature and the degree of initial opening of the lips as the important visual feature, the prototype for /da/ might be something like:

/da/ : Slightly falling F2-F3 & Open lips.

The prototypes for /ba/ and /ðə/ would be defined in an analogous fashion,

/ba/ : Rising F2-F3 & Closed lips,

/ðə/ : Somewhat rising F2-F3 & Somewhat closed lips.

Given a prototype's independent specifications for the auditory and visual sources, the value of one source cannot change the value of the other source. The integration of the features defining each prototype is evaluated according to the product of the feature values. If a_{Di} represents the degree to which the auditory stimulus A_i supports the alternative /da/, that is, has Slightly falling F2-F3; and v_{Dj} represents the degree to which the visual stimulus V_j supports the alternative /da/, that is, has Open lips, then the outcome of prototype matching for /da/ would be:

/da/ : $a_{Di}\, v_{Dj}$

where the subscripts i and j index the levels of the auditory and visual modalities, respectively. Analogously, if a_{Bi} represents

Synthesized Audible and Visible Speech

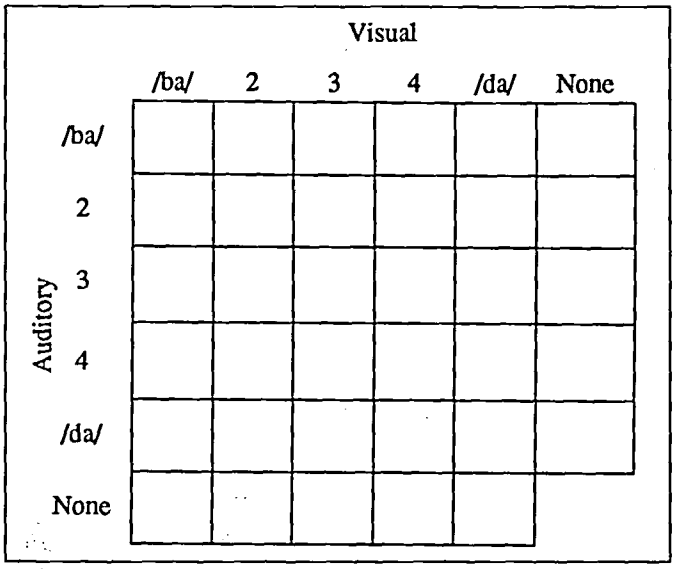

Fig. 5. Expansion of a typical factorial design to include auditory and visual conditions presented alone. The five levels along the auditory and visible continua represent auditory and visible speech syllables varying in equal physical steps between /ba/ and /da/.

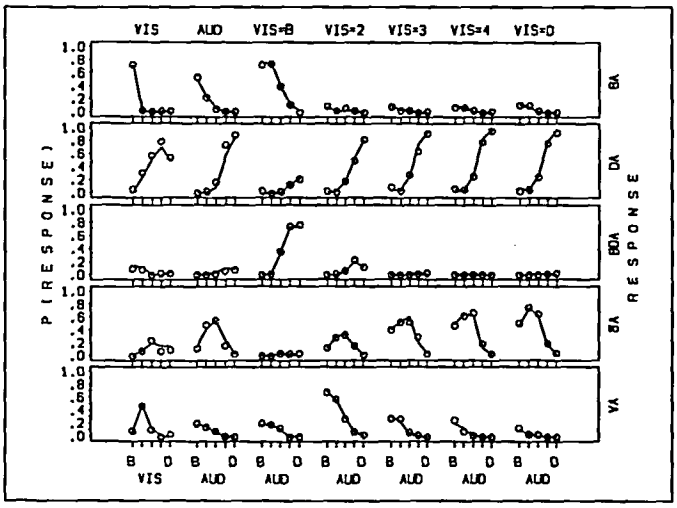

Fig. 6. Observed (points) and predicted (lines) proportion of /ba/, /da/ /bda/, /ða/, and /va/ identifications for the visual alone (left panel), auditory alone (second panel) and bimodal (remaining panels) conditions as a function of the five levels of the synthetic auditory (AUD) and visual (VIS) speech varying between /ba/ (B) and /da/ (D). The lines give the predictions for the FLMP.

the degree to which the auditory stimulus A_i has Rising F2-F3 and v_{Bj} represents the degree to which the visual stimulus V_j has Closed lips, the outcome of prototype matching for /ba/ would be:

$$\text{/ba/} \; : \; a_{Bi}\, v_{Bj}$$

and so on for the other alternatives.

The decision operation determines the merit of one alternative relative to the sum of the merits of all the alternative responses. With just a single source of information, such as the auditory one A_i, the probability of a /da/ response, P(/da/), is predicted to be:

$$P(/da/|A_i) = \frac{a_{Di}}{\sum_k a_{ki}} \tag{1}$$

where the denominator is equal to the sum of the merits of all k response alternatives. Given two sources of information A_i and V_j, P(/da/) is predicted to be:

$$P(/da/|A_i \text{ and } V_j) = \frac{a_{Di} \times v_{Dj}}{\sum_k a_{ki} \times v_{kj}} \tag{2}$$

where the denominator is equal to the sum of the merit of all k relevant alternatives.

It can be proven that the FLMP is mathematically equivalent to Bayes' theorem—an optimal algorithm for combining multiple sources of information (Massaro, 1987). In simplified form with equal a priori probabilities of the hypotheses, Bayes' theorem states that

$$P(H_1|E) = \frac{P(E|H_1)}{\sum_k P(E|H_k)} \tag{3}$$

where $P(H_k|E)$ is the probability that some hypothesis H_k is true given that some evidence E is observed; $P(E|H_k)$ is the probability of the evidence E, given that the hypothesis H_k is true. Bayes' theorem also specifies how different sources of evidence are combined. Given two independent pieces of evidence E_1 and E_2, the probability of a hypothesis H_1 is equal to

$$P(H_1|E_1 \text{ and } E_2) = \frac{P(E_1 \text{ and } E_2|H_1)}{\sum_k P(E_1 \text{ and } E_2|H_k)} = \frac{P(E_1|H_1) \times P(E_2|H_1)}{\sum_k P(E_1|H_k) \times P(E_2|H_k)} \tag{4}$$

Equations 3 and 4 have a direct correspondence with the FLMP, as formulated in Equations 1 and 2. Equation 4 gives the outcome for integrating two sources of information, where $P(E_1|H_1)$ represents evaluation of the first source and $P(E_2|H_1)$ represents separate evaluation of the second source. Equation 4 describes optimal information integration in the currency of probability under two assumptions. First, the prior probabilities of all relevant response alternatives are equal. Second, it is assumed that the sources of evidence are evaluated independently of one another. Under these assumptions, Equation 4

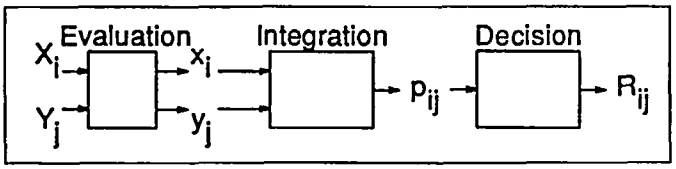

Fig. 7. Schematic representation of the three operations involved in perceptual recognition. The evaluation of a source of information X_i produces a truth value x_i, indicating the degree of support for alternative R. Integration of the truth values gives an overall goodness of match p_{ij}. The response R_{ij} is equal to the value p_{ij} relative to the goodness of match of all response alternatives.

follows from probability theory in which the probability of the joint occurrence of two independent events is the multiplicative combination of the probabilities of the separate events. The probability of two heads in two tosses of a coin, for example, is the multiplicative combination of the probability of a head on each toss. If auditory and visual speech are viewed as two sources of information, then Equation 4 is mathematically equivalent to Equation 2. Analogous to Equation 4, the evidence from audible speech is evaluated independently of the evidence from visible speech.

One important assumption of the FLMP is that the auditory source supports each alternative to some degree and analogously for the visual source. Each alternative is defined by ideal values of the auditory and visual information. Each level of source supports each alternative to a differing degree represented by feature values. Since we cannot predict the degree to which a particular auditory or visible syllable supports a response alternative, a free parameter is necessary for each unique syllable and each unique response. An auditory parameter is forced to remain invariant across variation in the different visual conditions and, analogously, for a visual parameter. The model, therefore, requires 5 parameters for the visual feature values and 5 parameters for the auditory feature values for each of eight response alternatives. Eighty free parameters are used to predict $35 \times 8 = 280$ data points. The feature values representing the degree of support from the auditory and visual information for a given alternative are integrated following the multiplicative rule given by the FLMP. The outcome of integration gives an overall goodness-of-match for each response alternative. Finally, the probability of choosing a given response alternative is predicted to be equal to its overall goodness-of-match divided by the sum of these values across all response alternatives.

The FLMP was fit to the individual results of each of the seven subjects. The quantitative predictions of the model are 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. We report the goodness-of-fit of a model by the root mean square deviation (RMSD)—the square root of the average squared deviation between the predicted and observed values. The lines in Figure 6 give the average predictions of FLMP. The model provides a good description of the identifications of both the unimodal and bimodal syllables (an average RMSD of .0557 across the individual subject fits).

Table 2. *Average parameter values supporting each response alternative as a function of the five levels of the visible (v1–v5) and audible (a1–a5) speech for the FLMP*

	Response Alternative				
Level	/ba/	/da/	/bda/	/ð a/	/va/
v1	.9390	.0210	.2535	.0006	.1423
v2	.0247	.4751	.2357	.1790	.7150
v3	.0281	.8443	.0222	.4160	.2874
v4	.0176	.9671	.0029	.2819	.0334
v5	.0101	.7725	.0996	.2858	.0169
a1	.7608	.0095	.0078	.2244	.3033
a2	.4890	.0362	.0165	.6320	.3247
a3	.1582	.1639	.0691	.9584	.1635
a4	.0229	.8088	.1963	.4724	.0216
a5	.0001	.9955	.1553	.0385	.0051

Table 2 gives the average best fitting parameters of the FLMP. These parameter values index the degree of support for each response alternative by each level of the audible and visible stimuli. As can be seen in the table, the parameter values change in a systematic fashion across the five levels of the audible and visible synthetic speech. For example, the support for the alternative /ð a/ is an inverted u-shape function of both the audible and visible continua. That is, creating a continuum between /ba/ and /da/ alternatives actually creates stimuli that are fairly representative of /ð a/ in the middle of the continuum.[2]

CATEGORICAL MODEL OF PERCEPTION (CMP)

In the categorical model of perception (CMP), it is assumed that only categorical information is available from the auditory and visual sources and that the identification judgment is based on separate categorizations of the auditory and visual sources. Considering a /ba/ identification judgment, the visual and auditory categorizations could be /ba/-/ba/, /ba/-/not ba/, /not ba/-/ba/, or /not ba/-/not ba/. If the two categorizations to a given speech event agree, the identification response can follow either source. When the two categorizations disagree, it is assumed that the subject will respond with the categorization to the auditory source on some proportion p of the trials, and with the categorization to the visual source on the remainder $(1 - p)$

2. It should be noted that the predicted points shown in Figure 5 cannot be recovered from the parameter values shown in Table 2. The figure and table give values averaged across the model fits of the seven subjects. Given that the FLMP is nonlinear, the average predictions cannot be computed from the average parameter values.

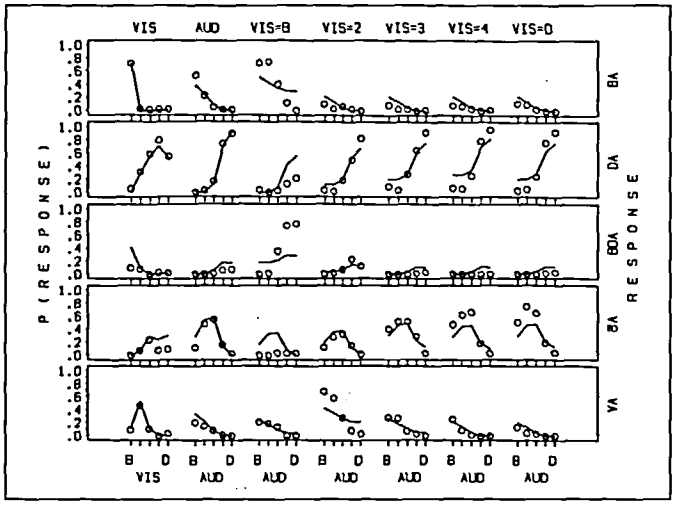

Fig. 8. Observed (points) and predicted (lines) proportion of /ba/, /da/ /bda/, /ðа/, and /va/ identifications for the visual alone (left panel), auditory alone (second panel) and bimodal (remaining panels) conditions as a function of the five levels of the synthetic auditory (AUD) and visual (VIS) speech varying between /ba/ (B) and /da/ (D). The lines give the predictions for the CMP.

of the trials. The weight p reflects the relative dominance of the auditory source.

The probability of a /ba/ identification response, P(/ba/), given a particular auditory/visual speech event, A_iV_j, is predicted to be:

$$P(/ba/|A_iV_j) = (1)\, a_{Bi}\, v_{Bj} + (p)\, a_{Bi}\,(1 - v_{Bj}) + (1 - p)(1 - a_{Bi})v_{Bj} + (0)(1 - a_{Bi})(1 - v_{Bj}) \quad (5)$$

where i and j index the levels of the auditory and visual modalities, respectively. The a_{Bi} value represents the probability of a /ba/ categorization given the auditory level i, and v_{Bj} is the probability of a /ba/ categorization given the visual level j. The value p reflects the amount of bias to follow the categorization of the auditory source. Each of the four terms in Equation 5 represents the likelihood of one of the four possible outcomes multiplied by the probability of a /ba/ identification response given that outcome. Note that Equation 5 reduces to:

$$P(/ba/|A_iV_j) = (p)(a_{Bi}) + (1 - p)(1 - v_{Bj}) \quad (6)$$

To fit this model to the results, each unique level of the auditory stimulus requires a unique parameter a_{Bi}, and analogously for v_{Bj}. The modeling of /ba/ responses thus requires 5 auditory parameters plus 5 visual parameters. Each of the eight response alternatives requires 10 free parameters. The p value would be fixed across all conditions for a total of 81 parameters. Thus, we have a fair comparison to the FLMP which requires 80 parameters.

The CMP was fit to the individual results in the same manner as in the fit of the FLMP. Figure 8 gives the average observed results and the average predicted results of the CMP. As can be seen in the figure, the CMP gave a poor description of the observed results. The RMSD was .1321—more than twice as large as the average RMSD of .0557 for the FLMP. An analysis of variance on the RMSD values showed that the FLMP gave a significantly better description of the results than did the CMP: $F(1, 6) = 123.843$, $p < .001$.

The CMP is mathematically identical to a weighted adding or a weighted averaging model (Massaro, 1987). Thus, a test of the CMP also allows a test of whether the inputs are added or combined in a nonadditive manner. The good fit of the FLMP relative to the CMP is evidence against additive integration. The integration of the multiple sources appears to result in the least ambiguous sources having the most impact on processing. Given that the FLMP is mathematically equivalent to Bayes' theorem—an optimal algorithm for integrating multiple sources of information, the good fit of the model to the present results is evidence for near optimal speech recognition by humans.

CONCLUSION

The present framework provides a valuable approach to the study of speech perception. We have learned how multiple sources of information are used in speech perception. In addition, we appear to have uncovered some of the fundamental stages of processing involved in speech perception by ear and eye. Given the potential for evaluating and integrating multiple sources of information in speech perception and understanding, no single source should be considered necessary. There is now good evidence that perceivers have continuous information about the various sources of information, each source is evaluated, and all sources are integrated in speech perception. Our ability to perceive speech appears to be so good, in part, because of the skillful use of multiple sources of information. Future work would continue to address the nature of the variety of sources of information, and how they function in recovering the speaker's message. We conclude with an observation of Sherlock Holmes speaking of two pieces of evidence: "Each is suggestive. Together they are conclusive."

Acknowledgments—The research reported in this paper and the writing of the paper were supported, in part, by grants from the Public Health Service (PHS R01 NS 20314), the National Science Foundation (BNS 8812728), the graduate division of the University of California, Santa Cruz, and a James McKeen Cattell Fellowship to the first author. The authors thank Joe Bunnett, Bill Ladusaw, Ray Gibbs, Art Samuel, Bill Estes, and an anonymous reviewer for their comments on the research presented in this article.

REFERENCES

Breeuwer, M., & Plomp, R. (1984). Speechreading supplemented with frequency-selective sound-pressure information. *Journal of the Acoustical Society of America, 76,* 686-691.

Brooke, N.M, & Summerfield, A.Q. (1983). Analysis, synthesis, and perception of visible articulatory movements. *Journal of Phonetics, 11,* 63-76.

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

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

Erber, N.P., & De Filippo, C.L. (1978). Voice-mouth synthesis of /pa, ba, ma/. *Journal of the Acoustical Society of America, 64,* 1015-1019.

Gouraud, H. (1971). Continuous shading of curved surfaces, *IEEE Transactions on Computers, C-20(6),* 623-628.

Hotchkiss, D. (1987). *Demographic aspects of hearing impairment: Questions and answers.* Washington, DC: Center for Assessment and Demographic Studies, Gallaudet Research Institute.

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

Ladefoged, P., Harshman, R., Goldstein, L., & Rice, L. (1978). Generating vocal tract shapes from formant frequencies. *Journal of the Acoustical Society of America, 64,* 253-257.

MacDonald, J., & McGurk, H. (1978). Visual influences on speech perception process. *Perception & Psychophysics, 24,* 253-257.

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

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

Montgomery, A.A. (1980). Development of a model for generating synthetic animated lip shapes. *Journal of the Acoustical Society of America, 68,* S58. (Abstract)

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

National Research Council (1987). *Speech understanding and aging.* Report of Committee on Hearing, Bioacoustics, and Biomechanics. Washington, DC: Commission on Behavioral and Social Sciences and Education.

Parke, F.I. (1974). *A parametric model for human faces* (Tech. Rep. UTEC-CSc-75-047). Salt Lake City: University of Utah.

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

Parke, F.I. (1982). Parameterized models for facial animation. *IEEE Computer Graphics, 2(9),* 61-68.

Pearce, A, Wyvill, B., Wyvill, G., & Hill, D. (1986). Speech and expression: A computer solution to face animation. *Graphics Interface '86,* 136-140.

Platt, J.R. (1964). Strong inference. *Science, 146,* 347-353.

Platt, S.M., & Badler, N.I. (1981). Animating facial expressions. *Computer Graphics, 15(3),* 245-252.

Ries, P. (1987). Characteristics of hearing impaired youth in the general population and of students in special education programs for the hearing impaired. In A.N. Schildroth & M.A. Karchmer (Eds.), *Deaf children in America* (pp. 1-31). Hillsdale, NJ: Lawrence Erlbaum.

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

Summerfield, A.Q. (1983). Audio-visual speech perception. In M.E. Lutman & M.P. Haggard (Eds.), *Hearing science and hearing disorders.* London: Academic.

Walden, B.E., Prosek, R.A., Montgomery, A., Scherr, C.K., & Jones, C.J. (1977). Effects of training on the visual recognition of consonants. *Journal of Speech and Hearing Research, 20,* 130-145.

Walden, B.E., Prosek, R.A., & Worthington, D.W. (1974). Predicting audiovisual consonant recognition performance of hearing-impaired adults. *Journal of Speech and Hearing Research, 18,* 272-280.

(RECEIVED 4/18/89; ACCEPTED 7/25/89)