Psychonomic Bulletin & Review
1996, 3 (2), 215-221

Perceiving affect from the voice and the face

DOMINIC W. MASSARO and PETER B. EGAN
University of California, Santa Cruz, California

This experiment examines how emotion is perceived by using facial and vocal cues of a speaker. Three levels of facial affect were presented using a computer-generated face. Three levels of vocal affect were obtained by recording the voice of a male amateur actor who spoke a semantically neutral word in different simulated emotional states. These two independent variables were presented to subjects in all possible permutations—visual cues alone, vocal cues alone, and visual and vocal cues together—which gave a total set of 15 stimuli. The subjects were asked to judge the emotion of the stimuli in a two-alternative forced choice task (either HAPPY or ANGRY). The results indicate that subjects evaluate and integrate information from both modalities to perceive emotion. The influence of one modality was greater to the extent that the other was ambiguous (neutral). The fuzzy logical model of perception (FLMP) fit the judgments significantly better than an additive model, which weakens theories based on an additive combination of modalities, categorical perception, and influence from only a single modality.

Research has shown that we use multiple sources of information when we comprehend speech (Massaro, 1987b, 1989; Massaro & Cohen, 1990). Visual information from a speaker's face, for example, can strongly influence speech perception, especially when the auditory information is degraded: in one study, recognition of auditory sentences in noisy environments improved from 23% to 65% when the perceivers could also see the speaker's face (Summerfield, 1979). We also use multiple sources of information when we perceive a speaker's emotion. These sources include a variety of paralinguistic signals, as well as the speech's verbal content. The emotion may be interpreted in different ways, depending on the voice quality, facial expression, and body language of the speaker. To study the degree to which paralinguistic sources of information are used, it is important that one first define these sources and then determine how they are evaluated and integrated. In the present study, in order to investigate the perception of a speaker's emotion, two sources of paralinguistic information were varied: facial expressions and vocal cues.

Facial expressions are an effective means of communicating emotion. Darwin (1872) argued that facial expressions originate in basic acts of self-preservation common to human beings and other animals, and that these expressions are related to the emotional states that they convey. Research by Meltzoff and Moore (1977) suggests that we are biologically prepared from birth to respond to facial expressions. They produced evidence which showed that infants as young as 12 days old are able to imitate adult facial expressions. Although many possible facial movements are possible, only a few such combinations are used to communicate emotions (Ekman, 1984). In other research, Ekman (1993) concluded that the processing of affect is particularly well developed in humans, who appear to be able to recognize and characterize facial expressions of emotional affect in other humans with a high degree of accuracy and consistency. Studies of animals, children born deaf and blind, and preliterate and isolated societies have provided evidence that some facial expressions, such as rage, startle, fear, and pleasure, may be universal. But these findings are not conclusive, and other research (Fridlund, 1991, 1994) suggests that behavioral and ecological factors have been the main influence in the development of facial expressions. As an example, we might ask why smiling is usually interpreted as a friendly gesture. Smiling while speaking raises the fundamental frequency (F0) and formant frequencies of speech, which can be interpreted as coming from a smaller and less threatening organism (Ohala, 1984).

Tanaka and Farah (1993) have found that individual facial features are recognized more easily when displayed as part of a whole face than when displayed in isolation. They concluded that facial recognition is a holistic process. However, *holistic processing* is a loaded term that is difficult to quantify. Etcoff and Magee (1992) tested subjects on computer-averaged faces, which differed by constant increments along a dimension of emotional affect. Given the similarity of their results to previous findings of "categorical perception," they concluded that these facial expressions were perceived categorically. However, this type of result is not necessarily evidence for categorical perception (Massaro, 1987a). Recently Ellison and Massaro (1995) tested models of recognition of facial affect in an attempt to overcome the limitations of verbal theories. They used a set of stimuli that would be standardized and replicable,

This research was supported, in part, by grants from the the National Institute on Deafness and Other Communication Disorders, the National Institutes of Health (2 R01 DC00236-13A1), the National Science Foundation (BNS 8812728), and the University of California, Santa Cruz. The authors thank Michael M. Cohen for help at all stages of this research. Correspondence should be addressed to D. W. Massaro, Program in Experimental Psychology, Clark Kerr Hall, University of California, Santa Cruz, CA 95064 (e-mail: massaro@fuzzy.ucsc.edu).

as well as controllable over a wide range of feature dimensions. A fully controllable synthetic talking face developed in our laboratory (Cohen & Massaro, 1993, 1994) was used to control and display the individual features within the face without the "cutting and splicing" required by the use of real faces. This simplifies the process of displaying features in a controllable manner, while simultaneously maintaining a highly realistic facial image. Thus meeting one of the requirements of our paradigm for inquiry (Massaro, 1987b), ambiguous, contradictory, and partial feature presentations can be tested very easily.

Ellison and Massaro (1995) performed two experiments with the same expanded-factorial design, with five levels of brow deflection crossed with five levels of mouth deflection, as well as their corresponding half-face conditions, for a total stimulus set of 35 faces. In one experiment, the task was a two-alternative forced choice between HAPPY and ANGRY; in another, nine rating steps from HAPPY to ANGRY were used. The results indicated that participants evaluate and integrate information from both features to perceive affective expressions. Both choice probabilities and ratings showed that the influence of one feature was greater to the extent that the other feature was ambiguous. By *ambiguous*, we mean that the information is relatively neutral with respect to the response alternatives available to the participants. The fuzzy logical model of perception (FLMP) fit the judgments from both experiments significantly better than did an additive model. Given the good fit of the FLMP with its assumptions of continuous and independent features, this research questions previous claims of categorical and holistic perception of affect.

The goal of the present study was to extend this paradigm to the independent manipulation of the face and the voice in the conveyance of affect. We also used an expanded-factorial design, with the advantages that single features as well as all feature combinations were tested. This design provides a stronger test of models of perceptual recognition and judgment than do single-factor or factorial designs (Massaro, 1987b).

With respect to voice quality, studies have shown that among other attributes, a person's age, intelligence, and emotional state can be abstracted from the voice alone (see review of literature by Kramer, 1963; Archer, 1991). The change in voice quality in emotional situations appears in part to be due to physiological changes. Ohala (1981) suggested three physical changes that can affect the voice: dryness in the mouth or larynx, accelerated breathing rate, and muscle tension. Pollack, Rubenstein, and Horowitz (1960) showed that emotion can be conveyed in speech segments as short as 60 msec, which indicates that emotion is present throughout an utterance. An experiment by Johnson, Emde, Scherer, and Klinnert (1986) demonstrated that the voice is a powerful source of information about emotion. In a forced choice experiment, subjects were almost perfect at recognizing joy, sadness, anger, and fear when listening to a semantically neutral sentence spoken in different simulated emotional states. Tartter and Braun (1994) asked speakers to produce syllables while smiling, frowning, or in a neutral manner. The auditory utterances were then played to listeners, who identified the happier one of a pair. Listeners were able to select the appropriate utterance at (only slightly) better than chance accuracy in both normal and whispered speech.

Williams and Stevens (1972) concluded that the pitch contour is the best indicator of the emotional content of an utterance. In their review of the literature, Murray and Arnott (1993) noted that the most commonly referenced vocal parameters are pitch (i.e., both the average value and range of the fundamental frequency), duration, intensity, and the undefined term *voice quality*.

Research on whether facial or vocal cues are more effective in the communication of emotion is inconclusive. In her literature review, Noller (1985) concluded that the issue was far too complex for general claims to be made about the relative importance of verbal and nonverbal channels. Relative importance is affected by too many variables, which include the type of emotion, the sex of the encoder, the sex of the decoder, and the age of the decoder. In the present framework, one modality is not more effective than the other. Rather, the relative effectiveness of the two modalities will fluctuate as a function of context. In general, however, we can expect the effectiveness of one modality to increase to the extent that the other modality is ambiguous.

The present study will focus on two independent variables—facial expression and vocal cues—and will examine how these variables are evaluated and integrated in the judgment of two specific emotions, happiness and anger. With respect to facial expression, two features are manipulated together—eyebrow deflection and mouth corner deflection. Ekman (1993) used photographs of persons who were instructed to hold a specific facial pose, and these photographs were dissected to create various stimuli. However, pictures of actual human faces will always be confounded by factors such as familiarity, attractiveness, and covariation, which can make the stimuli difficult to standardize. The synthetic talking head used in the present experiment enables the facial and audio cues to be combined with ease and also allows better control and more systematic analysis of the perceptual process.

Because of the current inability of synthetic voice programs, such as DECtalk, to adequately portray emotion in a single word, the vocal stimuli are produced by recording a male amateur actor speaking a semantically neutral stimulus word in three different simulated emotional states: happy, neutral, and angry.

It should be noted that our goal is simply to vary the amount of information supporting happiness and/or anger in the face and in the voice. We do not attempt to imply that we create a stimulus continuum between happy and angry, because we do not assume that happiness and anger are end points of a natural continuum. We simply created unimodal stimulus events that we believed varied in the degree to which they signified anger or happiness. In the end, it is really the subject's responses that tell us how well we did. There is no need to assume that these two emotions are end points of a single continuum. Similarly, there is no need to assume that the midpoint stimulus is an emo-

tionally ambiguous face. (In fact, the midpoints turned out to be somewhat more happy.)

Fuzzy Logical Model of Perception (FLMP)

The results from the experiment will be used to test among quantitative models of perceptual recognition. The FLMP has been shown to provide an impressive predictive model for similar type perception tasks (Massaro & Cohen, 1990, 1993; Massaro & Ferguson, 1993). The model assumes three basic stages of processing shown in Figure 1: (1) each source of continuous information is evaluated to ascertain the degree to which it matches various stored prototypes; (2) the sources are integrated according to a multiplicative formula to provide an overall degree to which they support each alternative; and (3) a decision is made on the basis of the relative goodness of match with each prototype. Within this framework, it is predicted that both modalities will influence the perception of emotion, and that the influence of one modality will be greater to the extent that the other modality is ambiguous. It is also predicted that reaction times (RTs) will be greater when the stimulus is ambiguous.

Within the context of the FLMP, it is assumed that participants generate prototypes corresponding to happy and angry affects. The prototype corresponding to a happy affect would include a description of a happy face and a happy voice, whereas an angry affect would have angry descriptions. At the feature evaluation stage, the face and voice are evaluated to provide the degree to which each source supports each of the two alternatives. If A_i represents the voice information, then A_i would be transformed to a_i, the degree to which the voice supports the alternative *angry*, A. An important assumption is that the evaluation of a particular feature occurs independently of the presence or absence of other features and their information value. The evaluation of the voice would produce the same result when presented with and without a face. With just two alternatives, happy (H) and angry (A), we can make the simplifying assumption that the degree to which the voice supports the alternative happy H is $1 - a_i$ (Massaro & Friedman, 1990). Feature evaluation occurs analogously for the face: V_j is transformed to v_j, the degree of support for *angry*.

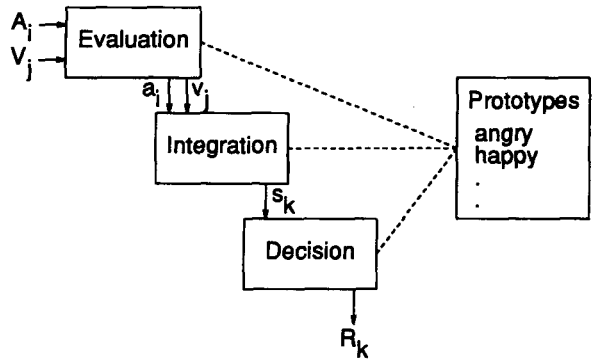

Figure 1. Schematic representation of the three stages involved in perceptual recognition. The three stages are shown to proceed left to right in time to illustrate their necessarily successive but overlapping processing. These processes make use of prototypes, which are shown for the alternatives in the present task. The sources of information are represented by uppercase letters. Auditory information is represented by A_i, and visual information, by V_j. The evaluation process transforms these sources of information into psychological values (indicated by lowercase letters a_i and v_j). These sources are then integrated to give an overall degree of support for a given alternative s_k. The decision operation maps this value into some response, R_k, such as a discrete decision or a rating.

Feature integration consists of a multiplicative combination of the feature values supporting a given alternative. If a_i and v_j are the values supporting alternative A, then the total support, $M(A)$, for the alternative A would be given by the product of a_i and v_j:

$$M(A) = a_i v_j. \tag{1}$$

With just two response alternatives, the auditory and visual support for alternative happy are $1 - a_i$ and $1 - v_j$, respectively. The total support, $M(H)$ is therefore

$$M(A) = (1 - a_i)(1 - v_j). \tag{2}$$

The third operation is decision, which uses a relative goodness rule (Massaro & Friedman, 1990) to give the relative degree of support for each of the test alternatives. In the two-alternative choice task, the probability of an angry choice, $P(A)$, is equal to

$$P(A \mid A_i, V_j) = \frac{M(A)}{M(A) + M(H)} \tag{3}$$

where $P(A \mid A_i, V_j)$ is the predicted choice given stimulus A_i, V_j.

Additive Model of Perception (AMP)

An additive model of perception (AMP) is also tested against the results (Cutting, Bruno, Brady, & Moore, 1992; Massaro, 1988; Massaro & Cohen, 1993). The evaluation stage is similar to the FLMP, but the values are added at the integration stage. Additive integration with a relative goodness of match at decision reduces to an averaging model (Massaro, 1987b, chap. 7). In addition, this model can be made more general by allowing one featural dimension to have more influence than the other. In this case, the probability of an angry identification is predicted to be

$$P(A \mid A_i, V_j) = wa_i + (1 - w)v_j, \tag{4}$$

where w is the weight given to the voice and $(1 - w)$ is the weight given to the face.

Some evidence for an AMP has been claimed by Huber and Lenz (1993), who trained pigeons to discriminate between features in a schematic human face. Although they were able to predict the decisions of their pigeons fairly well with an AMP, they did not test these results against other models. It should be noted that the AMP is mathematically equivalent to a single-channel model in which the participants attend to information from just one modality on a particular trial (Thompson & Massaro, 1989). The AMP is also equivalent to a categorical model in which

the participant categorizes an emotion from each modality and responds with the outcome from one of these categorizations with a certain probability (Massaro, 1987a).

METHOD

Participants

Fifteen students from the undergraduate psychology subject pool at the University of Santa Cruz participated in the experiment. The subjects were involved in the subject pool as part of their required undergraduate psychology work. They ranged in age from 18 to 23 (mean age, 20.1); 4 were male and 11 were female.

Apparatus

The visual stimuli used in the experiment were generated by facial synthesis software utilizing a parametrically controlled polygon topology synthesis technique with texture-mapped skin surfaces and ray-tracing lighting simulation. This program, Face39, is capable of producing animated visible speech at 60 frames per second which can be synchronized with audio-speech stimuli. A complete description of the technology is given by Cohen and Massaro (1993, 1994).

A set of three Face39 stimuli was constructed to portray affectual expressions representing happy, angry, and neutral emotional states. Brow deflection and mouth corner deflection were conjointly varied because of their effectiveness in changing the judgment of emotion from happy to angry (Ellison & Massaro, 1995). Figure 2 shows that both of the features were lowered for *angry*, were intermediate for the neutral expression, and were raised for *happy*. The face maintained its emotional expression during the articulation of the test word *please*, and lasted 1.07 sec.

To create the audio stimuli, a male amateur actor spoke the word *please* in a simulated happy, angry, and neutral emotional state. The stimuli were recorded on a SGI Crimson VGX workstation with a Vigra MMI-210 sound card. Some noticeable differences across the three different emotions were the pitch contour during the vocalic portion, the duration, and the amount of final frication. Using the spectrogram facility, the words were standardized for length (771 msec) and intensity (−15.3 dB) during the vocalic portion across the three emotions. In the bimodal condition, the auditory /p/ release was synchronized with the mouth opening of the /p/ articulation.

The experiment used an expanded factorial design with three audio stimuli and three visual stimuli. Together with the unimodal conditions, this gave a total of 15 conditions.

The Face39 program and the experimental control programs used to run the experiment and to collect data were implemented on a Silicon Graphics 4D/Crimson VGX workstation running under the IRIX operating system. The stimuli were presented to the subjects on 12-in. NEC model C12-202A color monitors. The subjects' responses and RTs were collected on TV1 950 VDT terminals and their associated keyboards. The stimulus face was sized to fill the vertical dimension of the screen, with a height of 18.5 cm and a width of 12 cm. The face was viewed at a distance of approximately 45 cm. No visual fixation point was provided. The auditory test word was presented at a comfortable listening intensity.

Procedure

Participants were instructed to watch the face and to listen to the word and to identify the emotion as happy or angry. They were required to respond to each stimulus with a two-alternative forced choice (2AFC) response, either HAPPY or ANGRY, by pressing a correspondingly labeled key on the VDT keyboard.

After a short title sequence, the control program displayed the stimuli on the monitor in the subjects' cubicles. The control program collected all subjects' responses (maximum of 4 subjects per session) before displaying the next stimulus. Therefore, there was a short but variable time between trials (about 3–4 sec).

Each experimental session included 10 practice trials and 180 test trials. The test trials were selected from the stimulus set according to a random selection without replacement protocol, which resulted in each stimulus' being displayed 12 times per session. Each participant took part in two experimental sessions, separated by a 5-min rest period. Each participant's results were therefore based on 24 observations for each of the 15 test trials.

RESULTS

Given the 2AFC task, the dependent measure, the probability of an angry judgment, completely represents the choice behavior. This probability can be expressed as P(Angry), and the probability of identifying the stimulus as happy is $1 - P$(Angry). The average P(Angry) judgments are shown in Figure 3. The two independent variables influenced performance as predicted. Figure 3 also reveals a significant interaction, because the influence of one

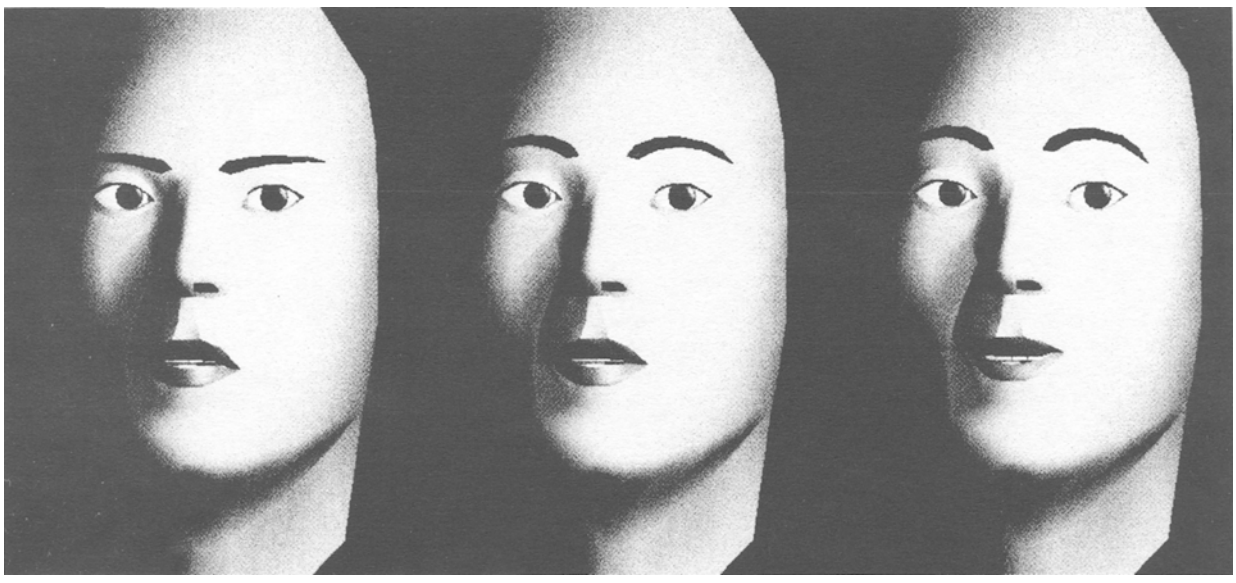

Figure 2. The three faces representing (from left to right) *angry*, *neutral*, and *happy*.

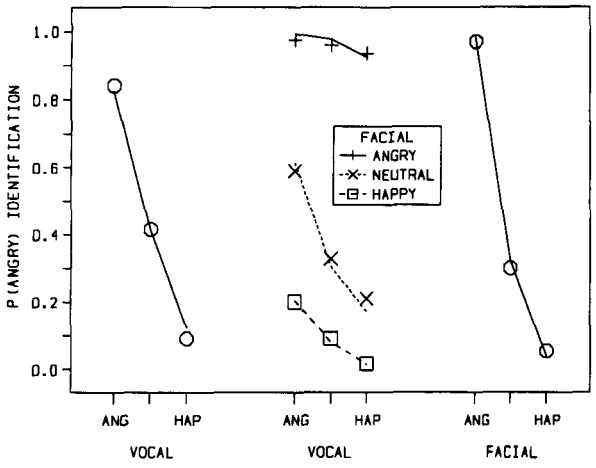

Figure 3. The observed (points) and predicted (lines) *P*(Angry) judgments as a function of the facial and vocal variables. Left graph gives performance for vocal only; right graph, for facial only; and the middle panel, for the bimodal condition. Predictions are for the FLMP.

variable was larger to the extent that the other variable was ambiguous (for all Fs, $p < .001$). As can be seen in Figure 3, the face had a larger effect on the judgments than did the voice. The probability of an angry judgment changed a larger amount across the three unimodal levels of facial affect than across the three unimodal levels of vocal affect. Furthermore, the face was very influential across all three levels of vocal affect.

These observed data were used to test the FLMP and AMP. Model fitting was accomplished through the use of STEPIT (Chandler, 1969). The FLMP and AMP were fit to the results of each of the 15 participants individually. The fit of the FLMP requires the estimation of three a_i values for the three levels of the auditory information and three v_j values for the three levels of the visual information. The fit of the AMP requires the estimation of these values plus a weight w value. The goodness of fit was calculated by using root mean square deviation (RMSD), which is the square root of the averaged squared deviation between predicted and observed values.

Figure 3 shows the average fit of the FLMP, along with the average results. As can be seen in the figure, the FLMP gave a good description of the results. The RMSDs for the FLMP fit of individual participants ranged from .0081 to .1200, with an average RMSD of .0509.

Table 1 presents the parameters for the fit of the FLMP to the results of each of the 15 participants. As can be seen in the table, the parameter values are meaningfully related to the levels of the auditory and visual stimuli. In general, the facial information was more influential than the voice information, as noted by the more extreme parameter values given the happy and angry faces than the happy and angry voices. In addition, the neutral stimuli were actually more happy than angry.

As noted in the presentation of the AMP, it is mathematically equivalent to a single-channel model, and a few of the observed points seem consistent with this model. The large proportion of angry judgments given the angry face and happy voice might be seen as evidence for the possibility that only a single modality (the face) is influencing the judgment in this condition. However, the good fit of the FLMP and the parameter values in Table 1 indicate that both modalities are influential, even though the more extreme values for the angry face lead to a predominance of angry judgments.

The fit of the AMP produced larger RMSDs with a range of .0070 to .1923, and an average RMSD of .0864. Table 2 gives the parameter values that were determined in the fit of the AMP. An ANOVA was carried out on the RMSDs for the fits of these two models; the FLMP provided a significantly better overall fit than the AMP did [$F(1,14) = 6.394, p = .023$].

A benchmark measure has been developed to provide an index of goodness of fit of a model (Massaro & Cohen, 1993). Even if a model is perfectly correct, it cannot be expected to predict the results exactly. A benchmark RMSD for each participant was determined by computing the binomial variance for each of the 15 experimental conditions, averaging these values, and taking the square root. These values can be compared with the RMSD values from the fit of the FLMP. The average benchmark value was .0473, very close to the average fit of the FLMP (.0509). The benchmark RMSDs were not significantly different from the observed FLMP RMSD values [$F(1,14) = .372, p = .558$]. This result shows that the FLMP describes the experimental results as well as can be expected from an accurate model.

The RMSDs for the AMP were also compared with the benchmark RMSDs. These values were significantly different from one another [$F(1,14) = 8.427, p = .011$]. This result shows that the description of the results by the AMP is significantly poorer than can be expected from an accurate model.

An average RT was also computed for each subject for each of the 15 stimulus conditions. One prediction of the

Table 1
Parameter Values Indicating the Degree of *Angry* for the Visual and Auditory Stimuli for the Fit of FLMP to the 15 Subjects

	Visual			Auditory		
Subject	Angry	Neutral	Happy	Angry	Neutral	Happy
1	0.9784	0.7229	0.1283	0.8616	0.7873	0.5368
2	0.9846	0.0134	0.0062	0.8754	0.5933	0.0871
3	0.9987	0.4348	0.0792	0.9869	0.1256	0.0443
4	0.9938	0.6140	0.0566	0.8674	0.4262	0.1811
5	0.9932	0.2891	0.0122	0.8418	0.2558	0.1108
6	0.9172	0.0890	0.0318	0.5612	0.2497	0.1571
7	0.9637	0.6169	0.0491	0.7498	0.6716	0.3838
8	0.9971	0.1284	0.0061	0.9145	0.1921	0.0693
9	0.9999	0.0001	0.0001	0.7187	0.7500	0.0626
10	0.9930	0.5059	0.1311	0.3729	0.9367	0.0175
11	0.9999	0.0285	0.0105	0.9825	0.1260	0.0031
12	0.9994	0.9879	0.0079	0.8411	0.4961	0.1135
13	0.9999	0.1542	0.0081	0.9760	0.0605	0.0542
14	0.9998	0.1240	0.0045	0.9359	0.4797	0.0055
15	0.9938	0.1462	0.0645	0.8584	0.1115	0.0595
M	0.9875	0.3237	0.0398	0.8229	0.4175	0.1257

Table 2
Parameter Values Indicating the Degree of *Angry* for the Visual and Auditory Stimuli and the Weight Given the Auditory Stimulus for the Fit of AMP to the 15 Subjects

	Visual			Auditory			Weight
Subject	Angry	Neutral	Happy	Angry	Neutral	Happy	Auditory
1	0.9841	0.3325	0.0692	0.9099	0.8946	0.5473	0.1919
2	0.9405	0.0001	0.0001	0.8780	0.5961	0.0613	0.0611
3	0.9999	0.4251	0.1779	0.9999	0.0648	0.0083	0.3573
4	0.9999	0.5802	0.0538	0.9776	0.3853	0.0286	0.1566
5	0.9999	0.2917	0.0001	0.9574	0.1916	0.0781	0.1422
6	0.8663	0.0215	0.0001	0.5648	0.2711	0.0342	0.1597
7	0.9531	0.6484	0.1093	0.9687	0.7812	0.3125	0.0001
8	0.9999	0.1854	0.0001	0.9990	0.1788	0.0569	0.1436
9	0.9922	0.0001	0.0001	0.7188	0.7500	0.0625	0.0001
10	0.9999	0.4669	0.1431	0.3613	0.9999	0.0001	0.4669
11	0.9999	0.1045	0.0349	0.9999	0.1405	0.0276	0.2404
12	0.9961	0.9480	0.0128	0.8445	0.4989	0.0947	0.0301
13	0.9999	0.3448	0.0127	0.9999	0.0380	0.0312	0.1903
14	0.9999	0.2031	0.0001	0.9999	0.4327	0.0001	0.1578
15	0.9999	0.1305	0.0328	0.8145	0.1135	0.0620	0.2280
M	0.9832	0.3428	0.0517	0.8663	0.4225	0.0937	0.1684

FLMP is that decision time is influenced by the relative goodness-of-match of the test stimulus and the response alternatives; therefore, RT should increase as the overall ambiguity of the stimulus increases (Massaro & Cohen, 1994). The RTs in Figure 4 showed that participants were significantly faster in making a choice when the stimulus was unambiguous. We define a measure of ambiguity,

$$A = .5 - |P(\mathrm{A}) - .5|, \qquad (5)$$

where A is ambiguity varying between 0 and .5, $P(\mathrm{A})$ the probability of an angry response, and $|x|$ is the positive value of x. Equation 1 simply states that ambiguity A increases as $P(\mathrm{A})$ approaches .5. Overall average RT correlated .881 with A. This high correlation provides additional strong support for the FLMP, which assumes that decision time increases as the degree of support for one alternative becomes more similar to the degree of support for the other alternative. Categorical perception cannot easily explain any change in RT, because perceivers putatively have information about only the discrete category (angry or happy), not the degree of category membership.

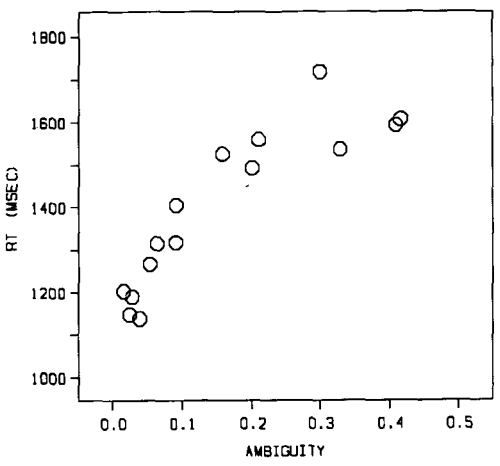

Figure 4. Average reaction time (RT) as a function of ambiguity *A* for the 15 conditions.

DISCUSSION

The present experiment is useful in helping us understand how facial and vocal cues are evaluated and integrated in the perception of emotion. Both variables (facial and vocal cues) were effective in changing the judgments of emotions from happy to angry. Each modality was more influential to the extent that the other variable was ambiguous. RTs of the perceptual judgments also increased as the stimuli became more ambiguous.

The ability to perceive emotion appears to be very good, owing to the use of multiple sources of information. The FLMP assumes continuous independent features for each source. The good fit of the FLMP to the results, and the poor fit of the AMP, is evidence against categorical models of perception, as well as additive models. Also, categorical models cannot easily explain why RTs increased when the stimuli were ambiguous (Massaro, 1987b, pp. 110–114). The FLMP, however, can explain this phenomenon. When the emotional cues are contradictory or ambiguous, more time is required before a sufficient degree of support accumulates and a response is emitted.

The recognition of emotion on the basis of facial and vocal cues is analogous to findings on pattern recognition in a wide variety of other domains. As an example, both facial and vocal information contribute to speech perception. Pursuing the analogy with speech perception, much research has shown that observers integrate auditory and visual information in an optimal manner, as described by the FLMP. For example, they can arrive at a relatively unambiguous percept when both modalities are somewhat ambiguous (Massaro, 1987b, p. 65). When the two sources of information conflict, on the other hand, an ambiguous percept is obtained. In addition, instructions and intention do not appear to be sufficient to preclude the influence of both sources of information in the bimodal speech perception task. One observes a large influence of visible speech even when the observers are instructed to report only what they hear or only what was presented on the auditory channel (Massaro, 1987b, pp. 66–83). We are currently studying the influence of instructions in the current emotion recognition task. It remains to be seen to what extent the facial information will actually change what the participants report about the voice.

The facial synthesis program provides a standardized set of faces that can be controlled precisely. However, as noted by Scherer, Banse, Wallbott, and Goldbeck (1991), research on vocal expression of emotion lags behind the study of facial affect expression, and this has resulted in vocal cues' being difficult to define and manipulate accurately. The recent in-

terest in the development of speech synthesis programs to signal attitudes and emotions holds promise that we will someday be able to manipulate the acoustic signal directly to convey affect (Cahn, 1990; Carlson, Granström, & Nord, 1993).

Future research using expanded factorial designs, which include other sources of emotional affect (such as hand movements and body posture), would help us to further understand how various features are combined in the perception of emotion.

REFERENCES

ARCHER, D. (PRODUCER) (1991). *A world of gestures: Culture and nonverbal communication* [Videorecording]. (Available from University of California Extension Media Center, Berkeley)

CAHN, J. E. (1990). The generation of affect in synthesized speech. *Journal of the American Voice I/O Society*, **8**, 1-19.

CARLSON, R., GRANSTRÖM, B., & NORD, L. (1993). Synthesis experiments with mixed feelings—A progress report. In *Fonetik-93: Papers from the Seventh Swedish Phonetics Conference* (Uppsala University Linguistics No. 23, pp. 65-68).Uppsala, Sweden.

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

COHEN, M. M., & MASSARO, D. W. (1993). Modeling coarticulation in synthetic visual speech. In N. M. Thalmann & D. Thalmann (Eds.), *Models and techniques in computer animation* (pp. 139-156). Tokyo: Springer-Verlag.

COHEN, M. M., & MASSARO, D. W. (1994). Development and experimentation with synthetic visible speech. *Behavior Research Methods, Instruments, & Computers*, **26**, 260-265.

CUTTING, J. E., BRUNO, N., BRADY, N. P., & MOORE, C. (1992). Selectivity, scope, and simplicity of models: A lesson from fitting judgments of perceived depth. *Journal of Experimental Psychology: General*, **121**, 364-381.

DARWIN, C. (1872). *The expressions of emotion in man and animals*. London: John Murray.

EKMAN, P. (1984). Expression and the nature of emotion. In K. Scherer & P. Ekman (Eds.), *Approaches to emotion* (pp. 319-343). Hillsdale, NJ: Erlbaum.

EKMAN, P. (1993). Facial expression and emotion. *American Psychologist*, **48**, 384-392.

ELLISON, J. W., & MASSARO, D. W. (1995). *Featural evaluation, integration, and judgment of facial affect*. Unpublished manuscript.

ETCOFF, N. L., & MAGEE, J. J. (1992). Categorical perception of facial expressions. *Cognition*, **44**, 227-240.

FRIDLUND, A. J. (1991). Evolution and facial action in reflex, social motive, and paralanguage. *Biological Psychology*, **32**, 3-100.

FRIDLUND, A. J. (1994). *Human facial expression: An evolutionary view*. San Diego, CA: Academic Press.

HUBER, L., & LENZ, R. (1993). A test of the linear feature model of polymorphous concept discrimination with pigeons. *Quarterly Journal of Experimental Psychology*, **46B**, 1-18.

JOHNSON, W. F., EMDE, R. N., SCHERER, K. R., & KLINNERT, M. D. (1986). Recognition of emotion from vocal cues. *Archives of General Psychiatry*, **43**, 280-283.

KRAMER, E. (1963). Judgment of personal characteristics and emotions from nonverbal properties of speech. *Psychological Bulletin*, **60**, 408-420.

MASSARO, D. W. (1987a). Categorical partition: A fuzzy logical model of categorization behavior. In S. Harnad (Ed.), *Categorical perception: The groundwork of cognition* (pp. 254-283). New York: Cambridge University Press.

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

MASSARO, D. W. (1988). Ambiguity in perception and experimentation. *Journal of Experimental Psychology: General*, **117**, 417-421.

MASSARO, D. W. (1989). *Experimental psychology: An information processing approach*. San Diego, CA: Harcourt Brace Jovanovich.

MASSARO, D. W., & COHEN, M. M. (1990). Perception of synthesized audible and visible speech. *Psychological Science*, **1**, 55-63.

MASSARO, D. W., & COHEN, M. M. (1993). The paradigm and the fuzzy logical model of perception are alive and well. *Journal of Experimental Psychology: General*, **122**, 115-124.

MASSARO, D. W., & COHEN, M. M. (1994). Visual, orthographic, phonological, and lexical influences in reading. *Journal of Experimental Psychology: Human Perception & Performance*, **20**, 1107-1128.

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

MASSARO, D. W., & FRIEDMAN, D. (1990). Models of integration given multiple sources of information. *Psychological Review*, **97**, 225-252.

MELTZOFF, A. N., & MOORE, M. K. (1977). Imitation of facial and manual gestures in human neonates. *Science*, **198**, 75-78.

MURRAY, I. R., & ARNOTT, J. L. (1993). Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. *Journal of the Acoustical Society of America*, **93**, 1097-1108.

NOLLER, P. (1985). Video primacy—a further look. *Journal of Nonverbal Behavior*, **9**, 28-47.

OHALA, J. J. (1981). The nonlinguistic components of speech. In J. Darby (Ed.), *Speech evaluation in psychiatry* (pp. 39-49). New York: Grune & Stratton.

OHALA, J. J. (1984). An ethological perspective on common cross-language utilization of F0 of voice. *Phonetica*, **41**, 1-16.

POLLACK, I., RUBENSTEIN, H., & HOROWITZ, A. (1960). Communication of verbal modes of expression. *Language & Speech*, **3**, 121-130.

SCHERER, K. R., BANSE, R., WALLBOTT, H. G., & GOLDBECK, T. (1991). Vocal cues in emotion encoding and decoding. *Motivation & Emotion*, **15**, 123-148.

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

TANAKA, J. W., & FARAH, M. J. (1993). Parts and wholes in face recognition. *Quarterly Journal of Experimental Psychology*, **46A**, 225-245.

TARTTER, V. C., & BRAUN, D. (1994). Hearing smiles and frowns in normal and whisper registers. *Journal of the Acoustical Society of America*, **96**, 2101-2107.

WILLIAMS, C. E., & STEVENS, K. N. (1972). Emotions and speech: Some acoustic correlates. *Journal of the Acoustical Society of America*, **52**, 1238-1250.

(Manuscript received June 14, 1995;
revision accepted for publication October 4, 1995.)