MULTIMODAL EMOTION PERCEPTION: ANALOGOUS TO SPEECH PROCESSES

Dominic W. Massaro

Perceptual Science Laboratory, University of California, Santa Cruz, Santa Cruz, CA 95060 U.S.A.

ABSTRACT

The fuzzy logical model of perception (FLMP) has been successful in accounting for a variety of unimodal and multimodal aspects of speech perception. This same framework has been extended to account for the perception of emotion from the face and the voice. The FLMP accurately describes how perceivers evaluate and integrate these sources of information to determine the affect signaled by the talker. This same research falsifies emotion processing as following a specialized analysis such as holistic or categorical perception.

1. PERCEIVING EMOTION IN FACES

Face recognition and the perception of facial expression are now being studied intensively by cognitive and neuro scientists. This field might be viewed as being at a stage similar to where the study of speech perception was about 2 decades ago. Only a handful of hard-nosed experimental psychologists have brought the phenomenon into the laboratory and subjected it to the disinterested scrutiny of empirical inquiry. Much of the previous literature has also been overburdened by a casual and less than well-informed application of evolutionary theory.

In spite of our belief in universal rather than domain-specific processes, we acknowledge that the information sources available for emotion perception belong to a different family than those available for speech perception. Nonetheless, our successes (Massaro, 1998, Chapters 4 and 6) demonstrating common principles of information processing across various domains encourages us to expect that these principles will hold up in the emotion domain

We operate under the assumption that multiple sources of information are also used to perceive a person's emotion. These consist of a variety of paralinguistic signals co-occurring with the verbal content of the speech. They may be aspects of voice quality, facial expression, and body language. In order to study how multiple paralinguistic sources of information are used, it is important first to define these sources. In our research, two sources of paralinguistic information, facial expressions and vocal cues, are chosen, to be analogous to the situation of bimodal speech.

Baldi, our computer-animated talking head (see Figure 1), makes possible a set of quite realistic faces for research that are standardized and replicable, as well as controllable over a wide range of feature dimensions. Thus, it quickly became apparent that we could initiate a cottage industry in the study of facial and vocal cues to emotion. There was no shortage of literature on facial cues to emotion but we found a tremendous void in the domain of vocal cues. We learned that Baldi had to be given

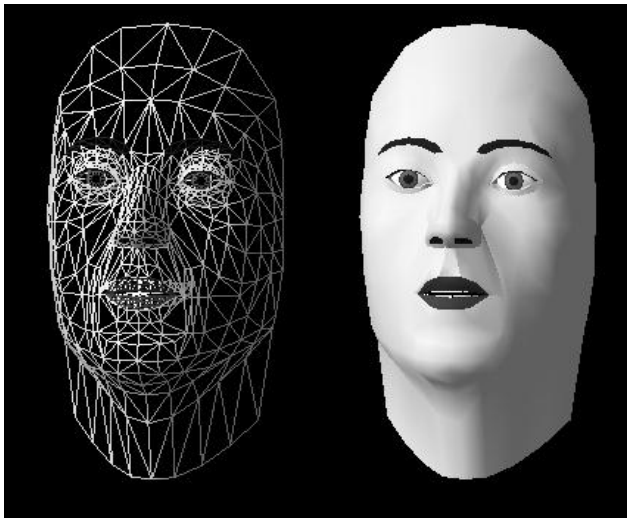

Figure 1 [http://mambo.ucsc.edu/psl/pela/wg.jpg] shows the talking head, called Baldi. As can be seen by the underlying wireframe model, there is not much behind his attractive exterior.

increased resolution in certain parts of the face, as well as additional controls over these parts (see Massaro, Chapter 12). Figure 2 illustrates the outcome of Baldi's six basic emotions.

We use the expanded-factorial design to study the pattern recognition of emotion (Ellison & Massaro, 1997). The affective categories happy and angry were chosen because they represent two of the basic categories of emotion. Of course, happy and angry faces are not discrete nonoverlapping emotional displays but a face can vary in the degree to which it represents one emotion as opposed to the other. To implement the expanded factorial design, it was necessary to choose two features to vary systematically to create a range of emotions between happy and angry. We chose two features that seem to differ somewhat in happy and angry faces. The features varied were brow displacement (BD) and mouth corner displacement (MD). BD was varied from slightly elevated and arched for a prototypically happy emotion to fully depressed and flattened for a prototypically angry emotion. MD was varied from fully curled up at corners for a prototypically happy emotion to fully curled down at corners for a prototypically angry emotion.

Our task was a two-alternative forced choice between HAPPY and ANGRY. There were 35 different test faces. Participants were not shown any exemplar faces, nor were they given any feedback. After 10 practice trials, each stimulus face was randomly presented 16 times to each of 26 participants for identification.

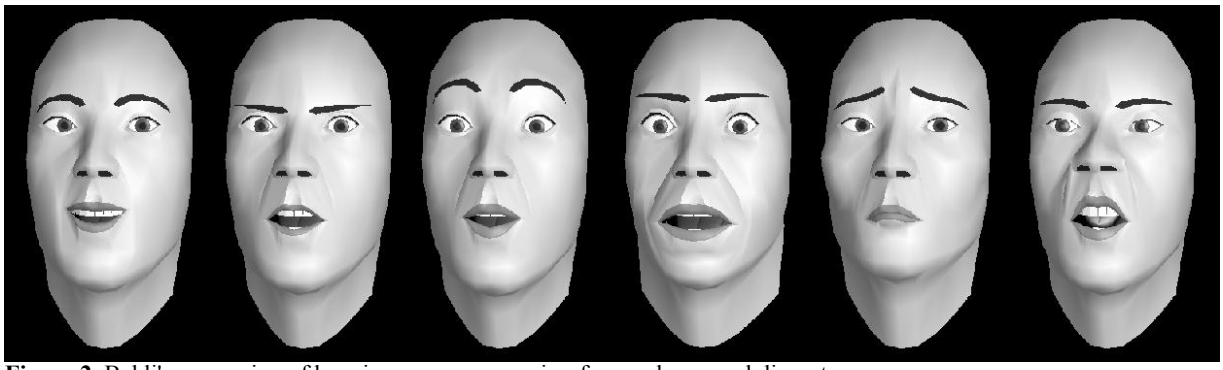

Figure 2. Baldi's expression of happiness, anger, surprise, fear, sadness, and disgust.

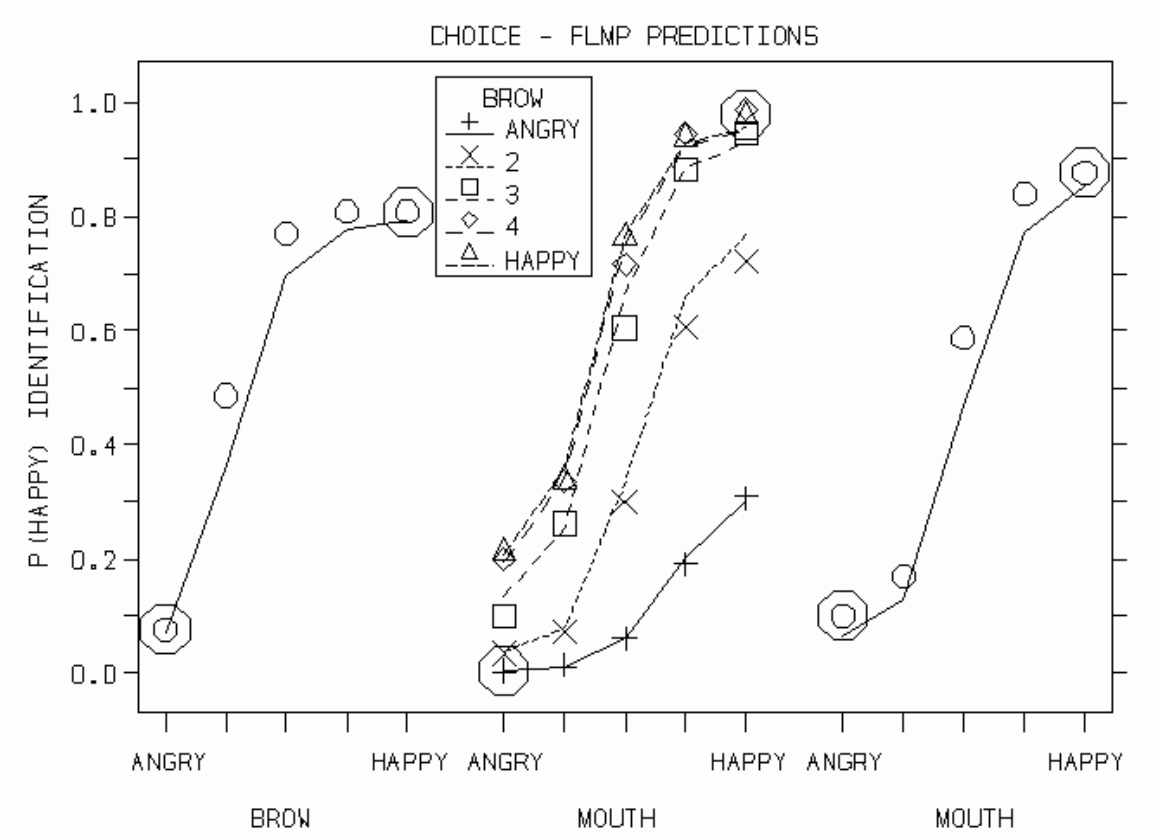

Figure 3. Predicted (lines) and observed (points) proportion of happy judgments as a function of the levels of the brow and mouth variables. The left panel shows performance for just the upper-half of the face and the right panel for just the lower-half. The middle panel gives performance for the factorial combination of the two halves. Average results across 26 subjects. The circled points show the superadditivity predicted by the FLMP. Predictions are for the FLMP. (From Ellison & Massaro, Experiment 1, 1998.)

The points in Figure 3 gives the average observed results as a function of the mouth and brow variables. The left panel shows performance when just the upper-half of the face was presented. Changes in the displacement of the brow were effective in changing the identified emotion in the expected direction. Similarly, the lower-half of the face influenced the number of "happy" judgments in the anticipated way. The steeper curve for the mouth variable illustrates that it was somewhat more influential than the brow variable. The middle panel gives the factorial combination of the two halves of the face. As can be seen in the figure, each of the two variables continued to be influential even when paired with the other variable.

The average results show most conclusively how two sources of information are more informative than just one. The probability of a happy judgment was about .80 when just the most upward deflection of the brow was presented and was about .88 for the most upward deflection of the mouth. However, when then two features were presented together in the whole face, the probability of a happy judgment was near 1. An analogous result was found for the most downward deflection of these two variables. These outcomes, called superadditivity, are consistent with our general view of pattern recognition. We now derive the predictions of the FLMP in order to test the model quantitatively against all of the results.

2. FUZZY LOGICAL MODEL OF PERCEPTION (FLMP)

Because of the close analogy to speech, we give only a short implementation of the FLMP's description of emotion perception. Participants are assumed to have prototypes corresponding to happy and angry faces. A happy face is characterized by the eyebrows slightly elevated and arched and the mouth corners fully curled up. An angry face is represented as having the eyebrows fully depressed and flattened and the mouth corners fully curled down. Of course, there are other sources of information described in the prototypes, but these do not require our attention because they should not be influenced systematically by the two independent variables. If B_i represents the brow information, then feature evaluation would transform B_i to b_i, the degree to which the brow supports the alternative happy. With just two response alternatives, happy (H) and angry (A), we can assume that the degree to which the evaluation of the brow supports the alternative A is $1 - b_i$ (Massaro & Friedman, 1990). The mouth information M_j is evaluated analogously: its support for H is m_j and its support for A is $1 - m_j$.

Feature integration consists of a multiplicative combination of the feature values supporting a given alternative. Given that b_i and m_j are the values of support for alternative H, then the total support for this alternative, s(H), would be given by their product

$$s(H)= b_i m_j .$$

The support for the alternative A would be .

$$s(H)= (1 - b_i) (1 - m_j) .$$

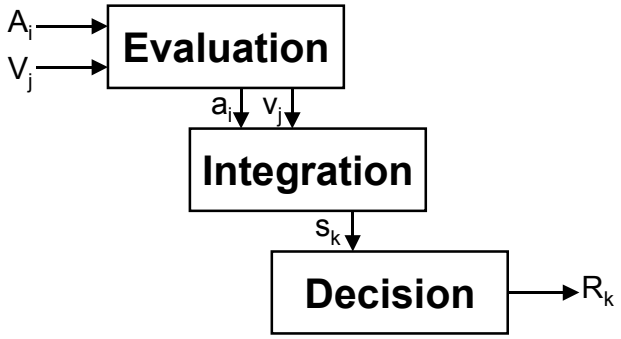

Figure 4. Schematic representation of the three processes involved in perceptual recognition (see text for details). The three processes are shown to proceed left to right in time to illustrate their necessarily successive but overlapping processing. These processes make use of prototypes stored in long-term memory. 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 or (fuzzy truth, Zadeh, 1965) values (indicated by lowercase letters a_i and v_j) These sources are then integrated to give an overall degree of support, s_k, for each speech alternative k. The decision operation maps the outputs of integration into some response alternative, R_k. The response can take the form of a discrete decision or a rating of the degree to which the alternative is likely.

The third operation is decision, which uses a relative goodness rule to give the relative degree of support for each of the test alternatives. In the two-alternative choice task, the probability of a happy choice given stimulus B_i M_j is given by

$$P(H \mid B_i \ M_j) = \{s(H)\} / \{s(H) + s(A)\}.$$

where $P(H \mid B_i M_j)$ is the predicted choice given the stimulus B_i M_j.

Multiplicative integration yields a measure of total support for a given category identification. This operation, implemented in the model, allows the combination of two imperfect sources of information to yield better performance than would be possible using either source by itself. However, the output of integration is an absolute measure of support; it must be relativized, due to the observed factor of relative influence (the influence of one source increases as other sources become less influential, i.e. more ambiguous). Relativization is effected through a decision stage, which divides the support for one category by the summed support for all other categories. An important empirical claim about this algorithm is that while information may vary from one perceptual situation to the next, the manner of combining this information--information processing--is invariant. With our algorithm, we thus propose an invariant law of pattern recognition describing how continuously perceived (fuzzy) information is processed to achieve perception of a category.

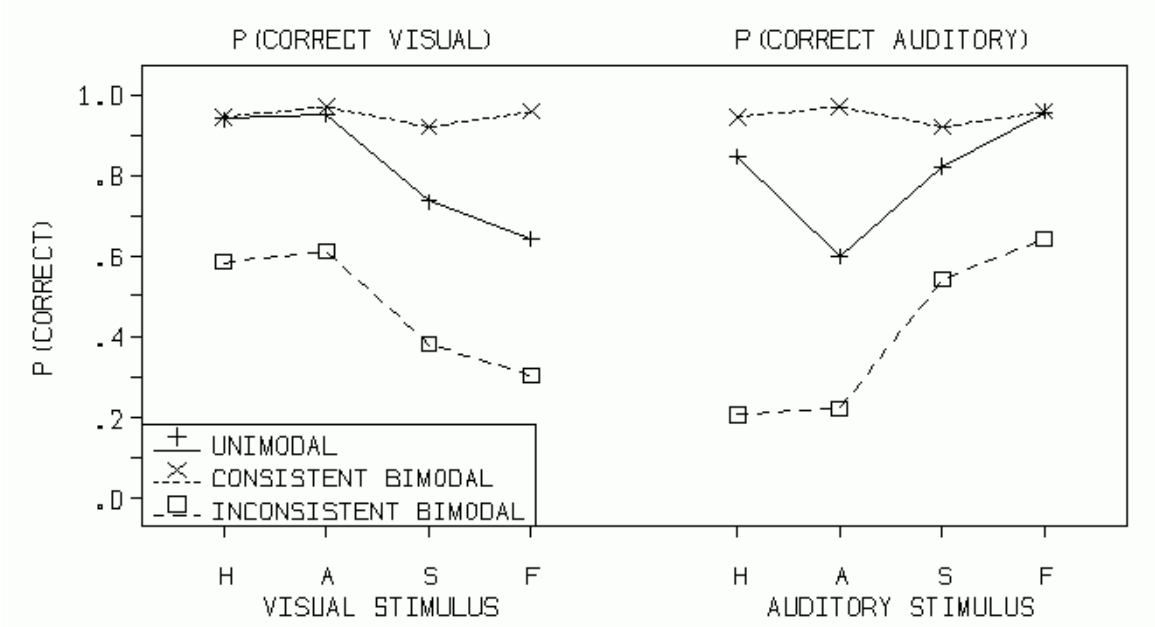

Figure 5. Probability correct visual (left panel) and auditory (right panel) responses as a function of visual or auditory level of happy (H), angry (A), surprised (S), and fearful (F) stimuli for unimodal, consistent, and inconsistent trials. The consistent condition is necessarily identical in the two panels because it represents the same results (from Massaro, 1998).

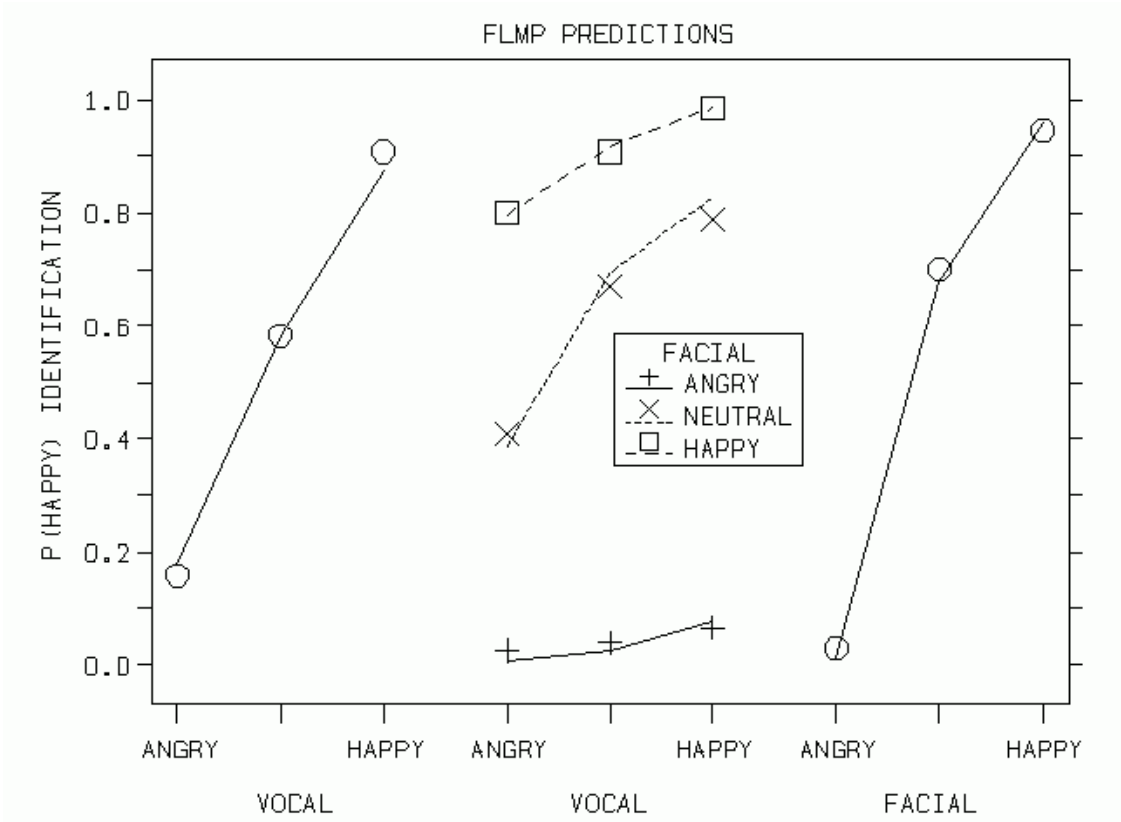

Figure 6. The observed (points) and predicted (lines) proportion of Happy judgments as a function of the visual and auditory variables. The left panel shows performance for just the face and the right panel for just the voice. The middle panel gives performance for the factorial combination of the two modalities. Predictions are for the FLMP. (From Massaro & Egan, 1996).

As in the case of bimodal speech, the FLMP requires 10 free parameters for the 5 levels of brow displacement and the 5 levels of mouth displacement. In the two-choice identification task, the model gave a very good description of the results. We conclude that perceivers have continuous information about each of the two facial characteristics being varied. Both the identification results and also continuous rating judgments are well-described by the FLMP, which not only assumes continuous information about each of the two features, but also an optimally efficient integration algorithm (see Massaro, 1998). We now explore the perception of emotion given information from the face and the voice.

3. VARYING THE FACE AND THE VOICE

The multimodal nature of emotion perception allows us to carry out empirical research exactly parallel to our studies of bimodal speech perception. As in our studies of speech perception, a particularly valuable experimental paradigm is to independently vary the two modalities in an expanded factorial design. Using an expanded factorial design, the four emotions were presented auditorily, visually, and bimodally. For the bimodal presentation, each audible word was presented with each visible word for a total of 4 times 4 or 16 unique conditions. Twelve of the bimodal words had inconsistent auditory and visual information. These conditions are necessary to achieve an informative picture of how these two modalities are processed.

A set of four stimuli was constructed from our synthetic face to portray affectual expressions representing happy, angry, surprised, and fearful states. The face maintained its emotional expression during the articulation of the test word *please*, and lasted about one second. The participants were instructed to watch the talking head and listen to the voice on each trial and to indicate which of the four emotion categories was being communicated. As observed in Figure 5, the participants made errors on the unimodal trials. When measured relative to the unimodal results, the bimodal results show a large influence of both modalities on performance. Overall performance was more accurate with two sources of consistent information than with either source of information alone. These results are consistent with the principle that the influence of one source of information when combined with another source is related to their relative ambiguity when presented in isolation.

An expanded factorial design with unimodal, consistent and inconsistent bimodal stimuli illuminated how emotion is communicated via the face and the voice. Emotion communicated along two modalities is more influential than just one. Furthermore, these sources appear to be processed in accordance with our universal law, the FLMP algorithm. These conclusions about emotion derived from both the face and the voice are consistent with those reached about emotion derived from two features of the face (see Massaro, Chapter 7).

4. AMBIGUITY OF VISIBLE AND AUDIBLE SOURCES

Massaro and Egan (1996) created happy, neutral, and angry expressions in both the face and the voice and used these in an expanded-factorial design. This gives 3 unimodal auditory, 3 unimodal visual, and 9 factorial test stimuli, for a total of 15 conditions. Participants judged the faces, the voices, and the combinations of the face and the voice as happy or angry. This study also allows us to address the question of how facial expression and vocal cues are evaluated and integrated in the judgment of two specific emotions, happiness and anger.

For facial expression, two features were changed together (eyebrow displacement and mouth corner displacement) to create the three levels. The brows were lowered and flattened and the corners of the mouth downturned for angry, the brows were raised and arched and the mouth corners upturned for happy, and these features were intermediate for the neutral expression. The face maintained its emotional expression during the articulation of the test word please, lasting about one second altogether. For vocal expression, the three types of emotion were simply recorded from a human speaker. In the experiment, the 15 participants were instructed to watch the face and to listen to the word and to identify the emotion as happy or angry.

The two independent variables influenced performance as expected. As can be seen in Figure 6, the face had a larger effect on the judgments than the voice. This is evident from the fact that the probability of a happy judgment changed a larger amount across the three unimodal levels of facial emotion than the three unimodal levels of vocal emotion. In addition, the bimodal results show that the voice had a minor effect when paired with the angry or happy face. A substantial influence of the voice occurred only when the face was neutral. As expected, Figure 6 also reveals a significant interaction between the variables, in that the influence of one variable was larger to the extent that the other variable was ambiguous.

These observed data were used to test the FLMP and competing models. The FLMP required 6 free parameters and a single channel model and a weighted averaging model each required 7 to fit the 15 independent data points. Figure 6 shows the good fit of the FLMP to the average results. The FLMP provided a significantly better overall fit than the single channel model and additive model.

An average reaction time (RT) was also computed for each of the 15 stimulus conditions. To assess the influence of ambiguity on RT, we computed a measure of ambiguity between 0 and 1, as in Chapter 7. These RTs show that participants were significantly faster in making a choice when the stimulus was unambiguous. When the emotional cues are contradictory or ambiguous, more time is required before a sufficient degree of support accumulates and a response is emitted. Average RT correlated .88 with the average ambiguity of the 15 test conditions. 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.

As noted in Massaro (1998, Chapter 7), 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.

5. INTENTIONALITY AND INSTRUCTIONS

In the bimodal speech perception task, instructions appear to play only a relatively minor role. One observes a large influence of visible speech even when the observers are instructed to report only what they hear (Massaro, 1987, pp. 66-83). One of the characteristics of our integration law is its naturalness or automaticity. When multiple sources of information are present, we cannot help but integrate them. The consequence of this automaticity of integration is the difficulty of ignoring a source of information. This lack of control over integration has been previously demonstrated in speech (Massaro, 1987), making it valuable to question whether the same is the case in the emotion domain.

An extension of this research was, therefore, implemented to assess how easily people could filter out one of the sources of emotion information. We replicated the three by three expanded factorial design under different sets of instructions. Three types of instructions were given to direct the intentional set or goal of the participant. For all instructions, participants were instructed to watch the face and to listen to the word and to identify the emotion as happy or angry. For the auditory instructions, they were instructed to "make the judgment on the basis of what you hear the voice to be expressing." Of course, it was necessary to warn the participants that sometimes only the face would be presented and, therefore, they would have to make their judgments on this basis. For the visual instructions, they were instructed to "make the judgment on the basis of what you see the face to be expressing."

The same task was carried out in auditory/visual speech to allow a direct comparison of instructions across the speech and emotion domains. One of the themes of our framework is that there are analogous processes across a broad range of domains. Thus, we expect to find similar results for emotion and speech. In addition to testing for analogous processes in the two domains, the instruction experiment was also aimed at whether instructions and intention would affect the two domains differently. As far as we can tell from the results of this research, instructions and intention modulate the impact of a given source of information but cannot preclude its influence completely. Although the influence of the to-be-ignored modality is significantly decreased, there is still a substantial influence. Thus, people are not able to completely filter out the influence of a to-be-ignored modality. On the other hand, they can attenuate its influence so that some degree of control is possible.

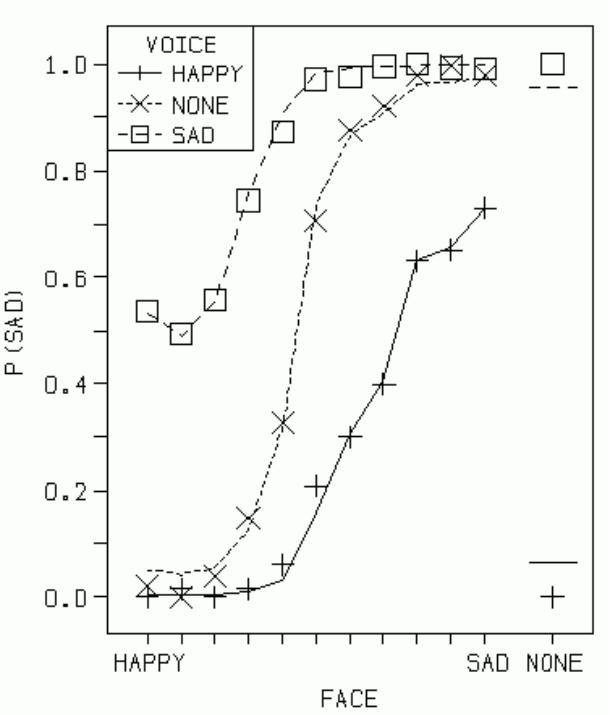

Figure 7. The points give the observed proportion of sad identifications in the auditory-alone, the factorial auditory-visual, and the visual-alone conditions as a function of the auditory and visual stimuli. The lines are the predictions of the FLMP.

5.1 Independent Confirmation

Results from other laboratories are particularly valuable to us because they represent an independent test of our theoretical framework by researchers other than ourselves. De Gelder and Vroomen (2000) asked participants to identify an emotion (e.g., happy or sad) given a photograph and/or an auditory spoken sentence. They found that their identification judgments were influenced by both sources of information, even when they were instructed to base their judgment on just one of the sources. In their first experiment, de Gelder and Vroomen asked participants to identify the person as happy or sad. The stimuli were manipulated in an expanded factorial design with an 11-step visual continuum between happy and sad and an auditory sentence that was read in either a happy or sad voice. Thus, there were 11 x 2 bimodal conditions, 11 visual-alone conditions, and 2 auditory-alone conditions, for a total of 35 unique stimulus conditions. The participants were instructed to watch the screen and to listen to the voice on each trial.

The FLMP was fit to the average results by estimating free parameters for the 11 levels of visual information and 2 levels of auditory information. Figure 7 gives the observed and predicted results. As can be seen in the figure, the FLMP gives a good description of the average results (Massaro & Cohen, 2000).

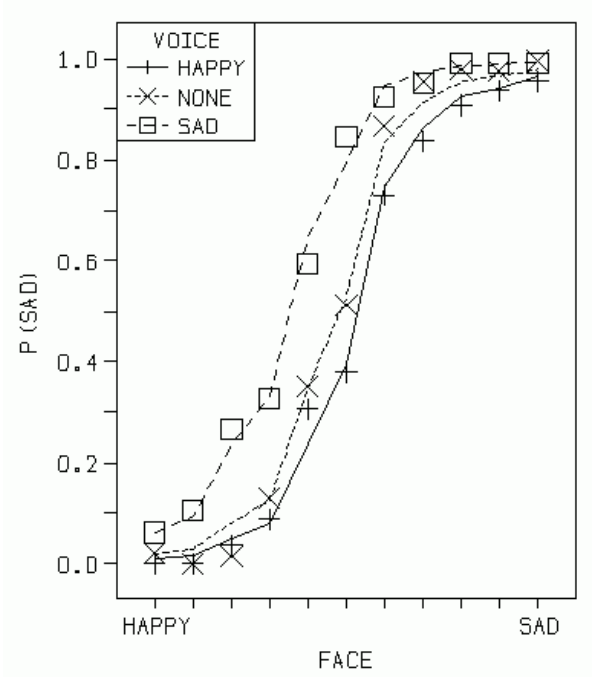

Figure 8. The points give the observed proportion of sad identifications in the factorial auditory-visual and the visual-alone conditions as a function of the auditory and visual stimuli. The lines are the predictions of the FLMP.

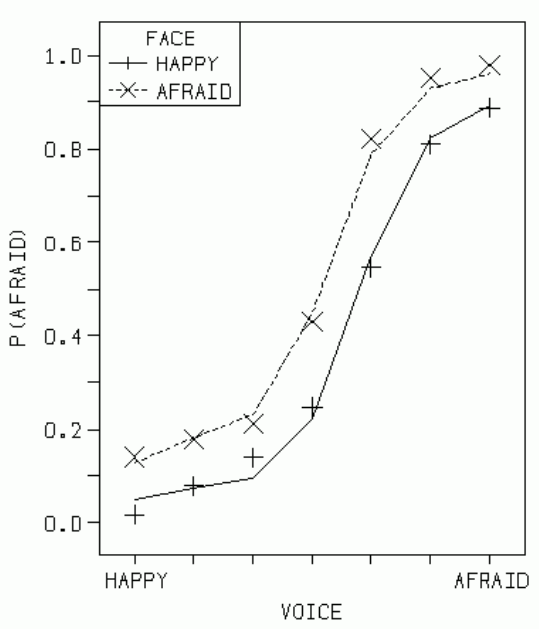

Figure 9. The points give the observed proportion of afraid identifications in factorial auditory-visual conditions as a function of the auditory and visual stimuli. The lines are the predictions of the FLMP.

The same design was used in the second experiment except that the two auditory-alone trials were omitted. Observers were told to judge the face and to ignore the voice. The FLMP was fit to these new results by estimating a new set of free parameters for the 11 levels of visual information and 2 levels of auditory information. Figure 8 gives the observed and predicted results. As can be seen in the figure, the FLMP gives a good description of the average results.

Comparison of the parameter values across the two experiments allows us to test the hypothesis that there are information differences in the two different instruction conditions. The parameter values for the happy and sad voice are made much more neutral (closer to .5) in the situation in which participants were instructed to ignore the voice than in the situation in which they were told to use both modalities. The parameter values for the face were mostly similar across the two conditions. Thus, the FLMP is capable of describing the results by simply assuming that the information from the voice was attenuated when participants were instructed to ignore it. The good fit of the FLMP in both instruction conditions, however, indicates that the two sources are integrated in the same manner regardless of instructions.

In the third experiment aimed at having observers judge the voice and to ignore the face, a 7-step auditory sentence continuum was made between happy and afraid. The visual stimuli were happy and fearful photographs of the speaker of the sentences. For some reason the auditory-alone stimuli were not presented. Thus, there were only 7 x 2 = 14 experimental conditions. The FLMP was fit to the average results by estimating free parameters for the 7 levels of auditory information and 2 levels of visual information. Figure 9 gives the observed and predicted results. As can be seen in the figure, the FLMP gives a good description of the average results. Although a direct comparison between this experiment and Experiment 1 is not justified because of the different stimuli that were used, we can observe that the influence of the face was much smaller when participants were instructed to ignore it. The parameter values representing the face for the prototypical emotions were much attenuated in Experiment 3 relative to Experiment 1.

In conclusion, we were successful in testing the FLMP against a new set of data from a new set of investigators. The framework and model provide a parsimonious account of several experimental manipulations. The distinction between information and information processing is a powerful concept and reveals how instructional differences can modulate performance in the task. This outcome replicates the findings in Massaro (1998, Chapter 8) and adds to the body of results supporting a universal principle for pattern recognition.

6. CONCLUDING COMMENTS

The paradigm that we have developed permits us to determine how one source of information is processed and integrated with other sources of information. The results also inform us about which of the many potentially functional cues are actually used by human observers (Campbell & Massaro, 1997; Massaro,

1987, Chapter 1; Massaro & Friedman, 1990). The systematic variation of properties of the signal combined with the quantitative test of models of speech perception enables the investigator to test the psychological validity of different cues. This paradigm has already proven to be effective in the study of audible, visible, and bimodal speech perception (Massaro, 1987, 1998). Thus, our research strategy not only addresses how different sources of information are evaluated and integrated, but can uncover what sources of information are actually used. We believe that the research paradigm confronts both the important psychophysical question of the nature of information and the process question of how the information is transformed and mapped into behavior. Many independent tests point to the viability of the FLMP as a general description of pattern recognition. The FLMP is centered around a universal law of how people integrate multiple sources of information. This law and its relationship to other laws are presented in detail in Massaro (1998).

The assumptions of the FLMP are testable because they are expressed in quantitative form. One is the idea that sources of information are evaluated independently of one another. Independence of sources is motivated by the principle of category-conditional independence (Massaro & Stork, 1998): it isn't possible to predict the evaluation of one source on the basis of the evaluation of another, so the independent evaluation of both sources is necessary to make an optimal category judgment. While sources are thus kept separate at evaluation, they are then integrated to achieve perception and interpretation.

Given this framework, one emerging feature of the FLMP is the division of perception into the twin levels of information and information processing. The sources of information from the auditory and visual channels make contact with the perceiver at the evaluation stage of processing. The reduction in uncertainty effected by each source is defined as information. In the fit of the FLMP, for example, the parameter values indicating the degree of support from each modality correspond to information. These parameter values represent how informative each source of information is. Information processing refers to how the sources of information are processed. In the FLMP, this processing is described by the evaluation, integration, and decision stages. Within this framework, we can ask what information differences exist among individuals and across different pattern-recognition situations. Similarly, we can ask whether differences in information processing occur. For example, we can look for differences in both information and information processing when participants are given different instructions in a pattern-recognition task.

7. REFERENCES

1. Campbell, C. S., & Massaro, D. W. (1997).Visible Speech Perception: Influence of Spatial Quantization. Perception, 26, 627-644.
2. De Gelder, b., & Vroomen, J. (2000). The perception of emotions by ear and by eye. *Cognition and Emotion, 14*, 289-311.
3. Ellison, J.W., & Massaro, D.W., (1997) Featural evaluation, integration, and judgement of facial affect, *Journal of Experimental Psychology: Human Perception and Performance, 23*, 213-226.
4. Massaro, D. W. (1987). *Speech perception by ear and eye: A paradigm for psychological inquiry.* Hillsdale, NJ: Lawrence Erlbaum Associates.
5. Massaro, D. W. (1998). Perceiving talking faces: From speech perception to a behavioral principle. Cambridge, MA: MIT Press.
6. Massaro, D.W., & Cohen, M. M. (2000). Fuzzy logical model of emotion perception: Comments on The perception of emotions by ear and by eye by de Gelder & Vroomen, *Cognition and Emotion, 14*, 313-320.
7. Massaro, D.W., & Egan, P.B., (1996) Perceiving Affect from the Voice and the Face, *Psychonomic Bulletin and Review, 3*, 215-221.
8. Massaro, D. W., & Friedman, D. (1990). Models of integration given multiple sources of information. Psychological Review, 97, 225-252.
9. Massaro, D.W., & Stork, D.G. (1998). Speech recognition and sensory integration. American Scientist, 86, 236-244.
10. Zadeh, L. A. (1965). Fuzzy sets. *Information and Control, 8*, 338-353

8. ACKNOWLEGEMENTS

Portions of this paper were adapted from Massaro (1998) and Massaro and Cohen (2000). The research reported in this paper was supported by grants from National Science Foundation (NSF grant ECS-9726645, NSF CHALLENGE grant CDA-9726363, NSF Grant 23818), the National Institutes of Health (PHS R01 DC00236), Intel Corporation, and the University of California Digital Media Program. I gratefully acknowledge the contributions of Christopher Campbell, Rashid Clark, Michael Cohen, David Jones, and Goh Kawai.