Massaro, D.W. (2004). A Framework for Evaluating Multimodal integration by Humans and A Role for Embodied Conversational Agents. In Proceedings of the 6th International Conference on Multimodal Interfaces, (ICMI'04)(pp.24-31). State College, PA. New York: ACM Press.

A Framework for Evaluating Multimodal integration by Humans and A Role for Embodied Conversational Agents

Dominic W. Massaro
Department of Psychology
University of California, Santa Cruz
Santa Cruz, California 95064 U.S.A.
1-831-459-2330
massaro@fuzzy.ucsc.edu

ABSTRACT

One of the implicit assumptions of multi-modal interfaces is that human-computer interaction is significantly facilitated by providing multiple input and output modalities. Surprisingly, however, there is very little theoretical and empirical research testing this assumption in terms of the presentation of multimodal displays to the user. The goal of this paper is provide both a theoretical and empirical framework for addressing this important issue. Two contrasting models of human information processing are formulated and contrasted in experimental tests. According to integration models, multiple sensory influences are continuously combined during categorization, leading to perceptual experience and action. The Fuzzy Logical Model of Perception (FLMP) assumes that processing occurs in three successive but overlapping stages: evaluation, integration, and decision (Massaro, 1998). According to nonintegration models, any perceptual experience and action results from only a single sensory influence. These models are tested in expanded factorial designs in which two input modalities are varied independently of one another in a factorial design and each modality is also presented alone. Results from a variety of experiments on speech, emotion, and gesture support the predictions of the FLMP. Baldi, an embodied conversational agent, is described and implications for applications of multimodal interfaces are discussed.

Categories and Subject Descriptors

J.4 [**Psychology**]

General Terms

Performance, Design, Experimentation, Human Factors, Theory

Keywords

Speech, Emotion, Gesture, Multisensory Integration,

1. INTRODUCTION

Central to this conference on multimodal interfaces is that the human-computer interaction is significantly facilitated by providing multiple input and output modalities. The research reported at ICMI appears to have studied this problem in an asymmetrical manner. The multimodal aspect of this conference has addressed primarily the machine's processing several input channels from the human user (Bauckhage et al., 2002; Chai et al., 2002; Corradini et al., 2003; Oviatt et al., 2003). For example, illuminating research has been carried out in the machine's processing and reaction to speech and pen inputs or speech and gesture inputs. Very little research, however, has been directed at how humans process several channels of information from multiple modalities. This issue is also central to any consideration of the value of interactive media in human-machine interaction.

The goal of this paper is to provide both a theoretical and empirical framework for the processing of information from multiple modalities. We strongly endorse formal models and methods in human computer interaction (Harrison & Thimbleby, 1990; Horvitz et al., 2001). In a typical human-machine interaction, the user might view text and images and this visual input could be accompanied by sounds and speech. In other cases, the presentation might be governed by an embodied conversational agent (ECA) who would naturally include audible and visible speech, emotions deployed by the face and voice, and gestures conveyed by the eyes, head, and body. The paper will review empirical and theoretical research addressing the question of how information from these channels is processed. Although the primary focus is on the processing of speech, emotion, and gesture, the paradigm for inquiry can be easily applied to any situation in which the user is presented with multiple sources of information. We begin with a description of the Fuzzy Logical Model of Perception (FLMP).

2. FUZZY LOGICAL MODEL OF PERCEPTION (FLMP)

The FLMP assumes that the various speech signals specifying a single event are continuously integrated during categorization, leading to perceptual experience and action. Before integration, however, each source is evaluated (independently of the other source) to determine how much that source supports various alternatives. The integration process combines these support values to determine how much their combination supports the various alternatives. The perceptual outcome for the perceiver will be a function of the relative degree of support among the competing alternatives.

To explain pattern recognition, representations in memory are an essential component. The current stimulus input has to be compared to the pattern recognizer's memory of previous patterns. One type of memory is a set of summary descriptions of the meaningful patterns. These summary descriptions are called

ICMI'04, October 13–15, 2004, State College, Pennsylvania, USA.

prototypes and they contain a description of the functional properties or features of the pattern. The features of the prototype correspond to the ideal values that an exemplar should have if it is a member of that category. To recognize a speech segment, for example, the evaluation process assesses the input information relative to speech prototypes in memory.

Figure 1 illustrates three major operations during bimodal speech recognition in the FLMP. Features are first independently evaluated (as sources of information) in terms of the degrees to which they match specific prototypes in memory. Each feature match is represented by a common metric of fuzzy logic truth-values that range from 0 to 1 (Zadeh, 1965). In the second operation, the feature values assigned to a given prototype are multiplied to yield an overall (absolute) goodness of match for that alternative. The decision is based on the goodness of match for each alternative divided by the sum of the support for all relevant alternatives (the relative goodness rule, Massaro, 1998).

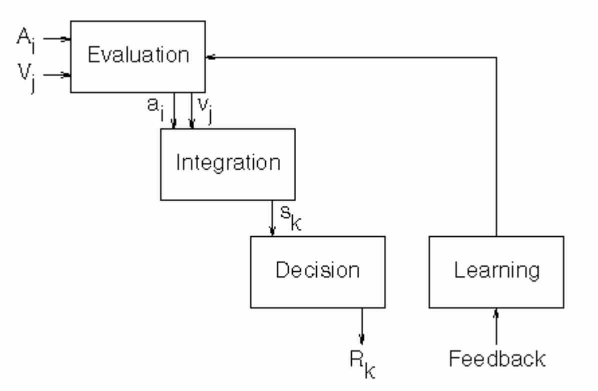


Figure 1. Schematic representation of the three processes involved in speech recognition. The three processes are shown to precede left to right in time to illustrate their necessarily successive but overlapping processing. These processes make use of prototypes stored in long-term memory. The sources 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 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 feedback is assumed to tune the prototypical values of the features used by evaluation.

The FLMP takes a strong stance on the question of discrete versus continuous information processing. Information input to a stage or output from a stage is continuous rather than discrete. Furthermore, the transmission of information from one stage to the next is assumed to occur continuously rather than discretely. The three processes shown in Figure 1 are offset to emphasize their temporal overlap. Evaluated information is passed continuously to integration while additional evaluation is taking place. Although it is logically the case that some evaluation must occur before integration can proceed, the processes are assumed to overlap in time. Similarly, integrated information is continuously made available to the decision process.

Given the FLMP framework, we are able to make an important distinction between "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 description of the FLMP, for example, the degree of support for each alternative from each modality corresponds to information. The predicted response probability in the unimodal condition is assumed to be a direct measure of the information given by that stimulus, which represents how informative or effective that 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.

More generally, multiple sources of information contribute to the identification and interpretation of the input. The assumptions central to the model are 1) each source of information is evaluated to give the continuous degree to which that source specifies various alternatives, 2) the sources of information are evaluated independently of one another, 3) the sources are integrated multiplicatively to provide an overall degree of support for each alternative, and 4) perceptual identification and interpretation follows the relative degree of support among the alternatives. The quantitative predictions of the FLMP have been derived and formalized in a number of different publications (e.g., Massaro, 1987, 1998). In a two-alternative bimodal speech perception task with /ba/ and /da/ alternatives, the degree of auditory support for /da/ can be represented by a_i, and the support for /ba/ by $(1 – a_i)$. The value i simply indexes the ith level along the auditory continuum and j indexes the level of the visual input. Similarly, the degree of visual support for /da/ can be represented by v_j, and the support for /ba/ by $(1 – v_j)$. The probability of a response to the unimodal stimulus is simply equal to its feature value. The predicted probability of a /da/ response given an auditory input, $P(/da/|A_i)$ is equal to

$$P(/da/\mid A_i)=\frac{a_i}{a_i+(1-a_i)}=a_i \tag{1}$$

In analogous fashion, the predicted probability of a /da/ response given an visual input, $P(/da/|V_j)$ is equal to

$$P(/da/\mid V_j)=\frac{v_j}{v_j+(1-v_j)}=v_j \tag{2}$$

For bimodal trials, the predicted probability of a /da/ response given auditory and visual inputs, $P(/da/|A_iV_j)$ is equal to

$$P(/da/\mid A_iV_j)=\frac{a_iv_j}{a_iv_j+(1-a_i)(1-v_j)} \tag{3}$$

Equations 1-3 assume independence between the auditory and visual sources of information. Independence of sources at the evaluation stage is motivated by the principle of category-conditional independence (Massaro, 1998; Massaro & Stork, 1998). Given that it isn't informative to predict the exact evaluation of one source on the basis of the evaluation of another, the independent evaluation of both sources is necessary to make an optimal category judgment. Although the sources are kept separate at evaluation, they are integrated to achieve perception, recognition, and interpretation. The multiplicative integration, implemented in the FLMP, allows the combination of two imperfect sources of information to yield better performance than would be possible using either source by itself. In addition, the FLMP predicts that one source of support will have more influence to the extent another source is ambiguous.

3. SINGLE CHANNEL MODEL (SCM)

According to nonintegration models, any perceptual experience and action results from only a single sensory influence. It follows that the pattern recognition of any multimodal event is determined by only one of the modalities, even though the influential modality might vary from one categorization event to the next. This idea is in the tradition of selective attention theories according to which only a single channel of information can be processed at any one time (Pashler 1998). According to the single channel model (SCM), only one of the available sources of information determines the response on any given trial.

Given a unimodal stimulus, it is assumed that the response is determined by the presented modality. A unimodal auditory stimulus will be identified as /da/ with probability a_i, and, analogously, the unimodal visual stimulus will be identified as /da/ with probability v_j.

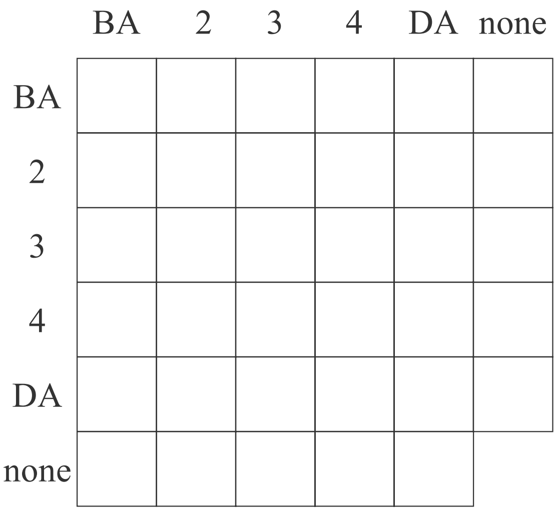
Because only one of the auditory and visual inputs can be used on any bimodal trial, it is assumed that the auditory modality is selected with some bias probability p, and the visual modality with bias 1 - p. If only one modality is used, it is reasonable to assume that it will be processed exactly as it is on unimodal trials. In this case, for a given bimodal stimulus, the auditory information will be identified as /da/ with probability a_i, and the visual information with probability v_j. Thus, the predicted probability of a /da/ response given the ith level of the auditory stimulus, a_i, and the jth level of the visual stimulus, v_j, is

$$P(/da/\,|\,A_iV_j) = pa_i + (1-p)v_j \qquad (4)$$

Equation 4 reveals that the contribution of the auditory and visual modalities are additive—that is, the absolute contribution given by a_i is independent of the value of v_j. Thus, it can be seen that the FLMP and SCM make very different predictions about how two sources of information are processed in the perception of a stimulus event.

4. EXPERIMENTAL TESTS

These models can be tested in expanded factorial designs in which two input modalities are varied independently of one another in a factorial design and each modality is also presented alone. For example, to evaluate the effectiveness of an embodied conversational agent, the properties of the auditory speech were varied to give an auditory continuum between the syllables /ba/ and /da/. In analogous fashion, properties of the animated face were varied to give a continuum between visual /ba/ and /da/. As shown in Figure 2, 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. This so-called expanded factorial design has been used with 82 participants who were repeatedly tested, giving 24 observations at each of the 35 stimulus conditions for each participant. These results have served as a database for testing models of pattern recognition (Massaro, 1998).



	BA	2	3	4	DA	none
BA						
2						
3						
4						
DA						
none						

Figure 2. 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/.

The proportion of /da/ responses for each of the stimulus conditions was computed for each participant. We present the results of just one representative participant to illustrate the nature of the data analysis and model testing. Figure 3 gives the observed (points) proportion of /da/ judgments as a function of the auditory and visual stimuli for a subset of the unimodal and bimodal conditions. Only two levels of visible speech are shown in the graph for pedagogical purposes. Notice that the columns of points are spread unevenly along the x-axis. The reason is that they are placed at values corresponding to the influence of the auditory speech (at a value equal to the marginal probability of a /da/ judgment for each level of the auditory independent variable). This spacing thus reflects relative influence of the successive levels of the auditory condition.

The single modality (unimodal) auditory curve (indicated by the open circles) shows that the auditory speech had a large influence on the judgments. The degree of influence of this modality when presented alone is indicated by the steepness of the response function across the auditory continuum. The two unimodal visual conditions are plotted at .5 on the auditory scale (which is considered to be completely neutral) on this scale. Their differential influence when presented alone is indexed by the vertical spread between these two levels at .5.

The other points give performance for the auditory-visual (bimodal) conditions. This graphical analysis shows that both the auditory and the visual sources of information had a strong impact on the identification judgments. The likelihood of a /da/ identification increased as the auditory speech changed from /ba/ to /da/, and analogously for the visible speech. The curves across

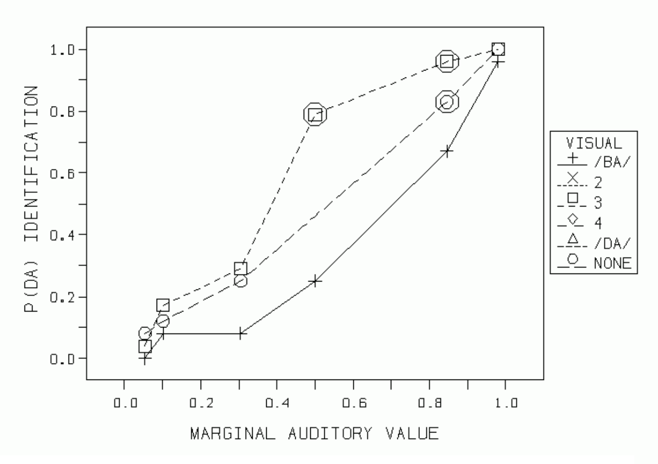


Figure 3. The points give the observed proportion of /da/ identifications in the unimodal and factorial auditory-visual conditions as a function of the five levels of synthetic auditory and visual speech varying between /ba/ and /da/. Only two levels of visible speech are shown in the graph for pedagogical purposes. The columns of points are placed at a value corresponding the marginal probability of a /da/ judgment for each auditory level on the independent variable. The auditory alone conditions are given by the open circles. The unimodal visual condition is plotted at .5 (completely neutral) on the auditory scale.

changes in the auditory variable are relatively steep and also spread out from on another and then converge again with changes in the visual variable. By these criteria, both sources had a large influence in the bimodal conditions.

Finally, the auditory and visual effects are not additive in the bimodal condition, as demonstrated by a significant auditory-visual interaction. The interaction is indexed by the change in the spread among the two bimodal curves across changes in the auditory variable. This vertical spread among the two curves is greater in the middle than at the ends of the auditory continuum. It means that the influence of one source of information is greatest when the other source is neutral or ambiguous. To understand multimodal speech perception, it is essential to understand how the two sources of information are used in perception. This question is addressed in the next section.

4.1 Evaluation of How Two Sources are Used

To address how the two sources of information are used, three points are circled in Figure 3 to highlight the conditions involving the fourth level of auditory information (A4) and the third level of visual information (V3). When presented alone, P(/da/| A4) and P(/da/| V3) are both about .8. When these two stimuli occur together, P(/da/| A4 V3) is about .95. This so-called synergistic result (the bimodal is more extreme than either unimodal response proportion) predicted by the FLMP does not seem to be easily explained by either the use of a single modality during a given bimodal presentation as assumed by the SCM or a simple averaging of the two sources on any trial.

In order to systematically evaluate theoretical alternatives, however, formal models must be tested against all of the results, not just selected conditions. Across a range of studies comparing specific mathematical predictions given by Equations 1-4 (Massaro, 1988, 1989, 1998), the FLMP has been more successful than the SCM (and other competitor models) in accounting for the experimental data (Massaro & Friedman, 1990; Massaro, 1989, 1998). To show that these results generalize to other domains, we describe an analogous study of the perception of emotion from the face and voice.

4.2 Perceiving Emotion

We assume that the FLMP is a general model of pattern recognition, and not limited to auditory-visual speech. To test this idea, we examined how emotion is perceived given facial and vocal cues of a speaker (Massaro, 1998; Massaro & Egan, 1996). Three levels of facial affect were presented using a computer-animated face. Three levels of vocal affect were obtained by recording the voice of a male amateur actor who spoke a semantically neutral word “please” in different simulated emotional states. These two independent variables were presented to participants in an expanded factorial design, i.e. visual cues alone, vocal cues alone and visual and vocal cues together, which gave a total set of 15 stimuli. The participants were asked to judge the emotion of the stimuli in a two-alternative forced choice task (either HAPPY or ANGRY).

The results indicate that participants 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 FLMP fit the judgments significantly better than the SCM, which also weakens theories based on an additive combination of modalities and categorical perception, as well as influence from only a single modality. Similar results have been found in other laboratories (de Gelder, & Vroomen, 2000; Massaro & Cohen, 2000). The perception of emotion appears to be well-described by our theoretical framework. Analogous to speech perception, we find a synergistic relationship between the face and the voice. Messages communicated by both of the modalities can be more informative than either one alone (Massaro, 1998).

4.3 Perceiving Speech and Gesture

It has been observed that manual gestures and speech are aspects of a single linguistic system (McNeill, 1985). If gestures and speech express the same meaning, then gestures and speech should function as two sources of information to be integrated by the perceiver. To test this hypothesis, we extended our framework to study the integration of a pointing gesture with audible speech (Thompson & Massaro, 1994). Using an expanded factorial design, preschool and fourth-grade children were presented with gesture, speech, and both sources of information together (Thompson & Massaro, 1994). An auditory speech continuum of five levels was made between the words ball and doll. The gestural information, also with five levels, was varied between pointing to the ball or doll objects. The child’s task was to indicate whether the talker meant the ball or the doll.

Both auditory speech and gesture influenced performance, and the form of the results were essentially identical to those found in experiments with audible and visible speech. Each source of information presented alone had some influence, and their joint influence followed the predictions of the FLMP. These results show that gestures from an ECA could possibly enhance a perceiver’s understanding of a human-machine interaction.

Having shown that individuals integrate multiple sources of information, it should be of interest to address several important issues related to this problem.

5. INTEGRATION OF MULTIPLE MODALITIES

One might question why perceivers integrate several sources of information when just one of them might be sufficient. Most of us do reasonably well in communicating over the telephone, for example. Part of the answer might be grounded in our ontogeny. Integration might be so natural for adults even when information from just one sense would be sufficient because, during development, there was much less information from each sense and therefore integration was all the more critical for accurate performance (Lewkowicz & Kraebel, 2004).

5.1 Underlying Neural Mechanism

A natural question concerns the neural mechanism underlying the integration algorithm specified in the FLMP. An important set of observations from single cell recordings in the cat's brain could be interpreted in terms integration of the form specified by the FLMP (Stein & Meredith, 1993). A single hissing sound or a light spot can activate neurons in the superior colliculus. A much more vigorous response is produced, however, when both signals are simultaneously presented from the same location. These results parallel the outcomes we have observed in unimodal and bimodal speech perception.

As proven elsewhere, the FLMP is mathematically equivalent to Bayes theorem (Massaro, 1998, Chapter 4), which is an optimal method for combining two sources of evidence to test among hypotheses. Anatasio and Patton (2004) propose that the brain can implement a computation analogous to Bayes theorem. They also show that the response of a neuron in the superior colliculus is proportional to the posterior probability that a target is present in its receptive fields, given its sensory input. This analysis assumes that the visual and auditory inputs are conditionally independent given the target, corresponding to our independence assumption at the evaluation stage. They observe that the target-present posterior probability computed from the impulses from the auditory and visual neurons is higher given sensory inputs of two modalities than it is given input of only one modality, analogous to the synergistic outcome of the FLMP.

5.2 A Universal Principle

The FLMP has proven to be a universal principle of pattern recognition (Campbell et al., 2001; Massaro, 1998; 2002; Massaro et al., 2001; Movellan & McClelland, 2001). In multisensory texture perception, for example, there appears to be no fixed sensory dominance by vision or haptics, and the bimodal presentation yields higher accuracy than either of the unimodal conditions (Lederman & Klatzky, 2004). In many cases, these sources of information are ambiguous and any particular source alone does not usually specify completely the appropriate interpretation. Parenthetically, it should be emphasized that these processes are not necessarily conscious or under deliberate control. We have found that typically developing children integrate information from the face and voice (Massaro, 1984; 1987; 1998) as well as do deaf and hard of hearing children (Massaro & Cohen, 1999) and autistic children (Massaro & Bosseler, 2003).

Erber (1972) tested three populations of children under auditory, visual, and bimodal conditions. The FLMP was applied to the results and gave an excellent description of the children's identification accuracy and confusion errors (Massaro, 1998, Chapter 14). Erber's results also reveal a strong complementarity between the audible and visible modalities in speech, which is discussed more fully in Section 6.3

Massaro and Bosseler (2003) tested whether autistic children integrate information in the identification of spoken syllables. An expanded factorial design was used in which information from the face and voice was presented either unimodally or bimodally, and either consistent with one another or not. After training the children in speechreading to enhance the influence of visible speech from the face, the identification task was repeated. Children behaved similarly in the two replications, except for a larger influence of the visible speech after training in speechreading. The FLMP gave a significantly better description of performance than the SCM, supporting the interpretation that autistic children integrate vocal and facial information in speech perception.

6. BALDI® AND THE VALUE OF AN ECA

Given the positive and eclectic evidence for the FLMP, one promising application area is the use of ECAs in human machine interaction. ECAs can mediate between the machine and the human in a variety of applications. One of the most valuable contributions would be the multimodal sources of information afforded by the ECA. We have seen that our perception and understanding are influenced by a speaker's face and accompanying gestures, as well as the actual sound of the speech (Massaro, 1987, 1998). We now review some of these valuable features of ECAs by illustrating their contribution to human-machine interfaces.

The value of visible speech in face-to-face communication was the primary motivation for the development of Baldi®, a 3-D computer-animated talking, emoting, and gesturing persona, shown in Figure 4. Baldi also has teeth, a tongue and a palate to simulate the inside of the mouth, and the tongue movements have been trained to mimic natural tongue movements (Cohen et al., 1998). Our software can animate Baldi in real time on a commodity PC, and Baldi is able to say anything at any time in our applications, using text-to-speech synthesis and facial animation. Baldi can be thought of as a puppet controlled by a set of strings that move and modify its appearance. In our algorithm for the synthesis of visible speech, each speech segment is specified with a target value or what we call a facial control parameter. The successive segments are blended together to implement coarticulation, which is defined as changes in the articulation of a speech segment due to the influence of neighboring segments (Cohen & Massaro, 1993; Massaro, 1998).

A central and somewhat unique quality of our work is the empirical evaluation of the visible speech synthesis, which is carried out hand-in-hand with its development. These tests show that Baldi now provides realistic visible speech that is almost as accurate as a natural speaker (Cohen et al., 2002; Massaro, 1998, Chapter 13). We have repeatedly modified the control values of Baldi in order to meet this criterion. We modify some of the control values by hand and also use data from measurements of real people talking (Cohen et al., 2002; Ouni et al., 2003). Baldi

Figure 4. Baldi, an embodied conversational agent with realistic speech, emotion, and gesture.

now speaks a variety of languages including Arabic (Badr, Ouni et al., 2003), Spanish (Baldero), Mandarin (Bao), Italian (Baldini, Cosi et al., 2002), German (Balthasar), Danish (Balder), and French (Baladin).

There are several reasons why Baldi is so successful, and why it holds so much promise for human-machine applications (Massaro, 1998). These include a) the information value of visible speech, b) the robustness of visual speech, c) the complementarity of auditory and visual speech, and d) the optimal integration of these two sources of information. We will review evidence for each of these properties and begin by describing an experiment illustrating how facial information increases recognition and memory for linguistic input.

6.1 Information Value of Visible Speech

Across a series of experiments, we asked 71 typical college students to report the words of sentences presented in noise (Jesse et al., 2000/2001). On some trials, only the auditory sentence was presented (unimodal condition). On some other trials, the auditory sentence was accompanied by Baldi, which was appropriately aligned with the auditory sentence (bimodal condition). The talking face facilitated performance for everyone. Performance was more than doubled for those participants performing particularly poorly given auditory speech alone. Although a unimodal visual condition was not included in the experiment, we know that participants would have performed significantly lower than the unimodal auditory condition. Thus, the combination of the auditory and visual sentences was synergistic because their combination led to accuracy that was significantly greater than accuracy on either modality alone. We have seen that our perception and understanding are influenced by a speaker's face and accompanying gestures, as well as the actual sound of the speech (Massaro, 1987, 1998). A similar synergistic result is found when noise-free speech is presented to persons with limited hearing (Erber, 1972).

6.2 Robustness of Visible Speech

Empirical findings indicate that speech reading, or the ability to obtain speech information from the face, is robust; that is, perceivers are fairly good at speech reading in a broad range of viewing conditions. To obtain information from the face, the perceiver does not have to fixate directly on the talker's lips but can be looking at other parts of the face or even somewhat away from the face (Smeele et al., 1998). Furthermore, accuracy is not dramatically reduced when the facial image is blurred (because of poor vision, for example), when the face is viewed from above, below, or in profile, or when there is a large distance between the talker and the viewer (Massaro, 1998; Munhall & Vatikiotis-Bateson, 2004). These findings indicate that speech reading is highly functional in a variety of nonoptimal situations. The robustness of the influence of visible speech is also illustrated by the fact that people naturally integrate visible speech with audible speech even when the temporal occurrence of the two sources is displaced by about a 1/5 of a second (Massaro & Cohen, 1993). Light and sound travel at different speeds and the dynamics of their corresponding sensory systems also differ (the retina transduces a visual stimulus much more slowly than the cochlea transduces an auditory one). Thus, a successful crossmodal integration should be relatively immune to small temporal asynchronies (Massaro, 1998, Chapter 3).

6.3 Complementarity of Auditory and Visual Information

A visual talking head allows for complementarity of auditory and visual information. Auditory and visual information are complementary when one of these sources is most informative in those cases in which the other is weakest. Because of this, a speech distinction between segments is differentially supported by the two sources of information. That is, two segments that are robustly conveyed in one modality tend to be relatively ambiguous in the other modality (Massaro & Cohen, 1999). For example, the difference between /ba/ and /va/ is easy to see but relatively difficult to hear. On the other hand, the difference between /ba/ and /pa/ is relatively easy to hear but very difficult to discriminate visually. The fact that two sources of information are complementary makes their combined use much more informative than would be the case if the two sources were non-complementary, or redundant (Massaro, 1998, Chapter 14).

6.4 Optimal Integration of Auditory and Visual Speech

The final value afforded by a visual talking head is that perceivers combine or integrate the auditory and visual sources of information in an optimally efficient manner (Massaro, 1987; Massaro & Cohen, 1999; Massaro & Stork, 1998). There are many possible ways to treat two sources of information: use only the most informative source, average the two sources together, or integrate them in such a fashion that both sources are used but that the least ambiguous source has the most influence. We have just seen in Section 4.1 that perceivers in fact integrate the information available from each modality to perform as efficiently as possible.

7. CONCLUSION AND IMPLICATIONS

We have provided a theoretical framework that illuminates how humans benefit from multiple sources of information from multiple modalities. People naturally evaluate and integrate the various input modalities to infer an intended communicative act. This pattern recognition function is nicely described by the FLMP, an optimal method for combining multiple input sources that vary continuously.

Our focus has been on how humans integrate information from multiple modalities. There has been considerable machine recognition research on the integration of auditory and visual speech (e.g., 42). Several distinctions in that field can be related to the present approach within the FLMP framework. One is the issue of early or late fusion or analogously feature-level fusion or decision-level fusion. Because information is continuously integrated in the FLMP, early versus late fusion is a non-issue. Similarly, integration can occur asynchronously or synchronously depending on the arrival and time course of evaluation of the auditory and visual sources of information. It is proposed that machine recognition systems should test the hypothesis that integration occurs continuously as the auditory and visual sources are processed and that algorithms implemented in the FLMP framework could serve as accurate multimodal machine recognition models.

These results have important implications for human-machine interactions in that they offer a justification for using embodied conversational agents (ECAs) in various applications. Although many have viewed speech as a natural medium for human-machine interactions, we have learned that a disembodied voice is often much less informative than a full-blown ECA that expresses visible speech, emotion, and gesture.

The results from typically developing children as well as deaf and hard of hearing and autistic children indicate that ECAs embedded in multisensory environments should be ideal for child computer interaction. In the area of speech and language learning. Baldi is now used in a Language Wizard and Player (Bosseler & Massaro, 2003; Massaro 2003) that allow easy creation and deployment of lessons in vocabulary and grammar. In another very different domain, Elizabeth Andre (personal communication) is using Baldi to engage young girls' interest in computer science.

Given the demonstrated value of ECAs, it is still necessary to test their effectiveness in each specific application in which they are used. Research has shown that the agent has to materially contribute to the application to have a positive influence (Andre, 2004) and therefore, it is necessary to distinguish between evaluation of the application versus evaluation of the ECA.

8. ACKNOWLEDGMENTS

The research and writing of the paper were supported by the National Science Foundation (Grant No. CDA-9726363, Grant No. BCS-9905176, Grant No. IIS-0086107), Public Health Service (Grant No. PHS R01 DC00236), a Cure Autism Now Foundation Innovative Technology Award, and the University of California, Santa Cruz. The technology and research resulted from the dedicated effort of many different members of the Perceptual Science Laboratory.

9. REFERENCES

[1] Anastasio, T. J., & Patton, P. E. (2004). Analysis and modeling of multisensory enhancement in the deep superior colliculus. In G. Calvert, C. Spence & B. E. Stein (Eds.), *Handbook of Multisensory Processes* (pp. 265-283). Cambridge, MA: MIT Press.

[2] Andre, E. (2004). Lessons Learned from Evaluating Animated Presentation Agents. *Workshop on Evaluating Embodied Conversational Agents,* Schloß Dagstuhl, Germany.

[3] Bauckhage, C., Fritsch, J., Rohlfing, K., Wachsmuth, S. & Sagerer, G, (2002). Evaluating Integrated Speech- and Image Understanding. *Proceedings of the 4th IEEE international conference on Multimodal interfaces* (pp. 9-14). Pittsburgh, Pennsylvania. October 14-16.

[4] Bosseler, A. & Massaro, D.W. (2003). Development and Evaluation of a Computer-Animated Tutor for Vocabulary and Language Learning for Children with Autism. *Journal of Autism and Developmental Disorders, 33,* 653-672.

[5] Campbell, C. S.; Schwarzer, G.; Massaro, D. W. (2001). Face perception: An information processing perspective. In M.J. Wenger, & J.T. Townsend (Eds.), *Computational, geometric, and process perspectives on facial cognition: Contexts and challenges* (pp. 285-345). Lawrence Erlbaum Associates, Inc., Publishers: Mahwah, NJ.

[6] Chai, J., Pan, S., Zhou, M. & Houck, K. (2002). Context-Based Multimodal Input Understanding in Conversational Systems. *Proceedings of the 4th IEEE international conference on Multimodal interfaces* (pp. 87-92). Pittsburgh, Pennsylvania. October 14-16.

[7] Cohen, M.M., Beskow, J. & Massaro, D.W. (1998). Recent developments in facial animation: An inside view. AVSP '98 (Dec 4-6, 1998, Sydney, Australia). http://mambo.ucsc.edu/psl/avsp98/11.doc

[8] Cohen, M.M., Massaro, D.W. & Clark, R. (2002) Training a talking head. In *Proceedings of ICMI'02, IEEE Fourth International Conference on Multimodal Interfaces.* October 14-16, Pittsburgh, Pennsylvannia.

[9] Corradini, A. Wesson, R, & Cohen, P. (2003). A Map-Based System Using Speech and 3D Gestures for Pervasive Computing. *Proceedings of the 4th IEEE international conference on Multimodal interfaces* (pp. 191-196). Pittsburgh, Pennsylvania. October 14-16.

[10] de Gelder, B. & Vroomen, J. (2000). Perceiving Emotions by Ear and by Eye. *Cognition & Emotion 14,* 289-311.

[11] Erber, N. P. (1972). Auditory, visual, and auditory-visual recognition of consonants by children with normal and impaired hearing. *Journal of Speech and Hearing Research, 15,* 423-422.

[12] Harrison, M. & Thimbleby, H. (1990). Formal methods in human-computer interaction. Cambridge University Press, Cambridge, MA.

[13] Horvitz, E. Kadie, C.M., Paek, T. &. Hovel, D. (2003). Models of Attention in Computing and Communications: From Principles to Applications, *Communications of the ACM 46(3)*, 52-59.

[14] Jesse, A., Vrignaud, N. & Massaro, D.W. (2000/01). The processing of information from multiple sources in simultaneous interpreting. *Interpreting, 5,* 95-115.

[15] Lederman, S. J. & Klatzky, R. L. (2004). Multisensory texture perception. In G. Calvert, C. Spence & B. E. Stein (Eds.), *Handbook of Multisensory Processes.* (pp. 107-122). Cambridge, MA: MIT Press.

[16] Lewkowicz, D. J. & Kraebel, K. S. (2004). The value of multisensory redundancy in the development of intersensory perception. In G. Calvert, C. Spence & B. E. Stein (Eds.), *Handbook of Multisensory Processes* (pp. 655-678). Cambridge, MA: MIT Press.

[17] Massaro, D.W. (1984). Children's perception of visual and auditory speech. *Child Development, 55,* 1777-1788.

[18] Massaro, D.W. (1987). Speech perception by ear and eye: A Paradigm for psychological inquiry. Hillsdale, NJ: Erlbaum.

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

[20] Massaro, D.W. (1989). Testing between the TRACE model and the Fuzzy Logical Model of speech perception. *Cognitive Psychology 21,* 398-421.

[21] Massaro, D.W. (1998). Perceiving talking faces: From speech perception to a behavioral principle. Cambridge, MA: MIT Press.

[22] Massaro, D.W. (1999). From theory to practice: Rewards and challenges. In *Proceedings of the International Conference of Phonetic Sciences* (pp. 1289-1292). San Francisco, CA.

[23] Massaro, D.W. (2000). From "Speech is Special" to Talking Heads in Language Learning. In *Proceedings of Integrating speech technology in the (language) learning and assistive interface, (InSTIL 2000)* (pp.153-161). University of Abertay Dundee, Scotland.

[24] Massaro, D.W. (2002). Multimodal Speech Perception: A Paradigm for Speech Science. In B. Granstrom, D. House & I. Karlsson (Eds.), *Multilmodality in language and speech systems* (pp.45-71). The Netherlands: Kluwer Academic Publishers

[25] Massaro, D.W. (2003). A computer-animated tutor for spoken and written language learning. *Proceedings of the 5th international conference on Multimodal interfaces* (pp. 172-175). Vancouver, British Columbia, Canada.

[26] Massaro, D.W. & Bosseler, A. (2003). Perceiving Speech by Ear and Eye: Multimodal Integration by Children with Autism. *Journal of Developmental and Learning Disorders, 7,* 111-144.

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

[28] Massaro, D.W. & Cohen, M.M. (1999). Speech perception in hearing-impaired perceivers: Synergy of multiple modalities. *Journal of Speech, Language & Hearing Science, 42,* 21-41.

[29] Massaro, D.W. & Cohen, M.M. (2000). Fuzzy logical model of bimodal emotion perception: Comment on "The perception of emotions by ear and by eye" by de Gelder and Vroomen. *Cognition and Emotion, 14(3),* 313-320.

[30] Massaro, D.W., Cohen, M.M., Tabain, M., Beskow, J. & Clark, R. (in press). Animated speech: Research progress and applications. In E. Vatiokis-Bateson, G. Bailly & P. Perrier (Eds.) Audiovisual Speech Processing, Cambridge, MA: MIT Press.

[31] Massaro, D.W. & Friedman, D. (1990). Models of integration given multiple sources of information. *Psychological Review, 97(2)* 225-252.

[32] Massaro, D.W. & Light, J. (2003). Read My Tongue Movements: Bimodal Learning To Perceive And Produce Non-Native Speech /r/ and /l/. In *Proceedings of Eurospeech '03-Switzerland (Interspeech). 8th European Conference on Speech Communication and Technology*. Geneva, Switzerland.

[33] Massaro, D.W. & Light, J. (in press). Using Visible Speech for Training Perception and Production of Speech for Hard of Hearing Individuals. *Volta Review.*

[34] Massaro, D.W. & Stork, D. G. (1998). Sensory integration and speechreading by humans and machines. *American Scientist, 86,* 236-244.

[35] McNeill, D. (1985). So you think gestures are nonverbal? *Psychological Review, 92,* 350-371.

[36] Mesulam, M.M. (1998). From sensation to cognition. *Brain, 121,* 1013-1052.

[37] Moore, M. & Calvert, S. (2000). Brief Report: Vocabulary acquisition for children with autism: Teacher or computer instruction. *Journal of Autism and Developmental Disorders, 30,* 359-362.

[38] Movellan, J. R. & McClelland, J. L. (2001). The Morton-Massaro law of information integration: Implications for models of perception. *Psychological Review, 108,* 113–148.

[39] Munhall, K., & Vatikiotis-Bateson, E. (2004). Spatial and Temporal Constraints on Audiovisual Speech Perception. In G. Calvert, C. Spence & B. E. Stein (Eds.), *Handbook of Multisensory Processes* (pp. 177-188). Cambridge, MA: MIT Press.

[40] Ouni, S., Massaro, D.W., Cohen, M.M. & Young, K. (2003) Internationalization of a talking head. *15th International Congress of Phonetic Sciences*. Barcelona, Spain.

[41] Oviatt, S., Coulston, R., Tomko, S., Xiao, B., Lunsford, R., Wesson, M., Carmichael L. (2003). Toward a theory of organized multimodal integration patterns during human-computer interaction. *Proceedings of the 5th international conference on Multimodal interfaces* (pp. 44-51). Vancouver, British Columbia, Canada.

[42] Pashler, H. E. (1998). The psychology of attention. Cambridge, MA: MIT Press.

[43] Potamianos, G., Neti, C., Gravier, G. & Garg, A. (2003). Automatic Recognition of audio-visual speech: Recent progress and challenges. In *Proceedings of the IEEE, 91(9),* (pp.1306-1326).

[44] Stein, B. E., & Meredith, M. A. (1993). The merging of the senses. Cambridge, MA: MIT Press.

[45] Thompson, L.A. & Massaro, D.W. (1994). Children's Integration of Speech and Pointing Gestures in Comprehension. *Journal of Experimental Child Psychology,* 57, 327-354.

[46] Zadeh, L. A. (1965). Fuzzy sets. *Information and Control, 8,* 338-353.