Developing and Evaluating Conversational Agents

Dominic W. Massaro, Michael M. Cohen, Jonas Beskow, Sharon Daniel, and Ronald A. Cole

Perceptual Science Laboratory
University of California
Santa Cruz, CA 95060
massaro@fuzzy.ucsc.edu

Abstract

Conversation agents present a challenging agenda for research and application. We describe the development, evaluation, and application of Baldi, a computer animated talking head. Baldi's existence is justified by the important contribution of the face in spoken dialog. His actions are evaluated and modified to mimic natural actions as much as possible. Baldi has the potential to enrich human-machine interactions and serve as a tutor in a wide variety of educational domains. We describe one current application of language tutoring with children with hearing loss.

Embodied Characters

The title of this conference is "Embodied Conversational Characters." Why not just "Conversational Characters?" What does embodiment add to our quest for a simulacrum of some human agent? Traditionally, the success of artificial intelligence was deemed to be contingent on creating a thinking machine, encompassing all of the rationality, logic, and abstract knowledge possessed by humans. The achievement of this goal doesn't appear to be much closer now than it was at the onset of the cybernetic age four decades ago. Today, an expanding cadre of cognitive scientists is offering a new image of thinking and intelligence. The functions of mind are embedded in a body, which is situated in a physical world of action. Very few of the functions of mind take place outside of this embodiment. As summarized by Andy Clark, (1997, p. 1), "Minds are not disembodied logical reasoning devices." Natika Newton (1997), for example, proposes that schematized sensorimotor images are the cognitive basis of understanding.

Because our simulacrum does not experience the physical world of action, does this mean that our agent will never understand, or at least give the impression of understanding? (Of course, this question touches on the infamous Chinese Room scenario defended by Searle, 1992). Glenberg (1997), a memory and language processing researcher, has proposed an indexical hypothesis for understanding. We supposedly index words to objects or perceptual symbols. Making use of a variant of Gibson's (1979) notion of affordances, he proposes that important guides for our behavior are the relations between bodies and objects in the world. These necessarily change with our goals and our experiences. These various influences provide a set of constraints that must mesh appropriately for a behavioral action. Glenberg's empirical behavioral experiments indicate that this framework provides a better understanding of peoples' meaningful judgments of sentences than does more abstract approaches such as latent semantic analysis. Perhaps it is premature to judge how the ideas of embodiment and their implications for conversational agents will play out. One potentially valuable implication, however, is that convincing agents might be instantiated, even if or because they think less and do more.

In this paper, we describe the progress we have made in the development of a conversational agent. Our current use of this agent is a classroom tutor for language training for children with hearing loss. There are several technological and psychological components to this research endeavor and, as expected, some are more realized than others. We begin with a description of Baldi, a computer-animated talking head. We then describe its application within a speech toolkit that allows the teacher and student to develop lessons that incorporate spoken language input and output.

Enhancing Our Experience

Few of us would actually admit how much time we spend on computers and the internet. Sara Kiesler (1997) asked Pittsburghians about both their depression and the amount of time spent on the internet. She found a positive relationship in that more time on the internet was correlated with more depression. For every additional hour spent online, you are about 1% more depressed (as measured on a depression rating scale). The media and press in all their wisdom magnified the findings by interpreting correlation as causation–e.g., "Can spending time on the Internet really make you lonely and depressed?" (Time, September 14, 1998). In reality, we shouldn't be surprised if, in fact, time spent with machines contributes to depression. Perhaps the internet addicts are simply better informed, so that time spend reading newspapers might have the same outcome. Accepting that

time spent with people would not make the same contribution, or at least less of one, our goal should be to make our experience with machines more like our experience with people (but not depressing people). This type of reasoning supports our goal for creating lifelike conversational agents on our desktops, appliances, and perhaps even in our social arenas.

It should be mentioned that “intelligent agents” are not equally valued by scientists, educators, and the general public. Jaron Lanier (1996) has defended the opinion that they are not only infeasible, but that they are also evil. He argues that agents will become controlling and restraining. Rather than agents becoming smarter, people will become dumber. We believe the questions of intelligent agents are only beginning to be formulated, yet alone being answered. The heart of Lanier’s criticisms rests on several debatable assumptions having to do with the nature of the self and the nature of experience. His view of a person conforms closely to western thought; other views such as those of eastern philosophy might not find artificial agents so aversive. Today experience is limited to living things, but who are we to say that this will always be true. We do not intend our agents to be evil but rather to effect a more natural relationship between us and the information and cognitive artifacts that are becoming increasingly common in everyday life. The students’ and teachers’ experience with Baldi as a language tutor has been extremely positive, and might serve as an initial test bed for this controversy. This technology has clearly helped free the children the constraints of hearing loss, although the exact amount and long-term benefit is still under investigation.

Our goal for conversational agents is to create a lifelike presence in our work and play environment. In the atmosphere of the Renaissance, the goal is to fool the eye (trompe l'oeil) and in our scenario, the mind. Reeves and Nass (1996) work on the media equation reveals that users are active participants (conspirators?) in this deception, all too willing to project personalities and social behaviors onto machines. Creating realism in the agent's appearance and actions is an understandable and reasonably modest goal.

We desire our conversational agent to be sensitive to our personality, moods, and wishes, but how sensitive do we want the agent to be? We seem to be somewhat successful in keeping some secrets from our loved ones, friends, and colleagues, and we would hope the same would be true with respect to our simulated conversational agent. Recently, researchers have claimed that a person emits various subtle cues that can reveal his or her true feelings or state of mind. For example, a fake smile can be distinguished from a real smile depending on whether or not there is also crinkling about the eyes. A lie might be distinguished from truth by the detection of various subtle body movements such as the shrugging of the shoulders or a contradictory movement of the lips while speaking (Ekman, 19xx). Perhaps our conversational agent might learn to read our moods and situate our interaction appropriately? Rosalind Picard (1997) plans to have her Digital DJ read the user's mood and then select the appropriate music. In this case, the agent would be reacting to many different sources of information including touch, voice, and face. We could imagine that our computer agent might learn to be more discerning than most real agents, if the many inputs to the machine exceed those that are easily processed by real people. For example, we touch the remote control of our machine in cases where touching other people would be inappropriate.

We believe wholeheartedly that computer users will benefit from interaction with conversational agents and access to the many sources of information they would provide. For example, there is valuable and effective information for the perception and recognition of speech when viewing a speaker's face. To provide this information, we have developed a completely animated synthetic talking head. With this synthetic speaker we can control and study the informative aspects and psychological processes in face-to-face dialogs. Our talking head communicates paralinguistic as well as linguistic information, and is controlled by either a text-to-speech system or a recording of natural speech that has been phonetically transcribed and then is aligned automatically.

We hope to advance the development of our talking head, its design and its accompanying technology, and to create a human-computer interface centered on a virtual, conversational agent. Such agents will interact with human users in the most natural manner possible including the ability to listen and understand, as well as speak fluently. Agents should facilitate and enrich interaction between humans and machines. Communication among humans might also be enhanced when mediated by virtual agents (either personal avatars or autonomous agents). Our work involves the development of the conversational agent, the design of the agent interface and its environment, and the psychological evaluation of its contribution to human language acquisition, communication, and productivity.

Talking Heads in Action

Technology advocates have always hoped that natural conversational behavior would be the primary medium of communication between people and machines. Our talking head, as a conversational agent, takes us one step closer to that realization. Each of us could have our own agent (in our own image, if we wish) to handle our communications. A conversational agent does not get tired or bored, isn't waylaid by a sore throat and, (as of yet) belongs to no union – in short, it's a perpetual 'talking and understanding' machine. As we develop the technology, talking heads will be able to speak in any language, at any rate of speed or level of complexity, with the appropriate emotional

affect. In addition to our current use of visible speech to facilitate learning and language acquisition for the hearing impaired, we envision applications of this technology in a variety of domains, including, but not limited to, education, entertainment, and human-machine interaction. For example, our talking head could serve as a useful aid in learning second languages and in improving the phonological and reading skills of dyslexic children. As we continue our research we expect tht the talking head will play an important role in the enhancement of auditory synthetic speech as well as an educational tool in linguistics and speech science.

Access to the Web and other sources of digital information not only requires a computer, but also the ability to interact with computers in a competent manner. Language and literacy barriers, as well as physical disabilities provide significant obstacles to many of our citizens. ``Today, every aspect of computers, from the out-of-the-box experience to the surfing the Internet, is a joy to ``technoguys" and an unpleasant challenge to ordinary citizens." (Tognazinni, 1997, p. 277). We propose to develop and test a computer-animated conversational agent that will increase the access to digital information and qualitatively enhance human-machine interaction. Our research is guided by the hypothesis that a conversational agent will increase access to communication and information technology, and encourage and support learning through digital media. The primary function of the conversational agent is to make the machine more human-like and thus, to empower the user, who is able to interact more naturally and productively with other humans than with machines.

In order to support the development of conversational agents for universal access and learning, research is necessary on a number of core technologies and their integration. These are dialogue modeling, natural language processing, automatic auditory, visual, and gestural recognition of a speaker’s linguistic, emotional, and conversational cues, speech synthesis, gestural, and facial animation. Many of the key research challenges and potential advancements to the state of the art lie at the boundaries where these disciplines meet. This work is based on the experimental investigation of the functional value of conversational systems.

The Importance of Talking Faces in Dialog

Why haven't engineers and computer scientists been able to program a computer to recognize and understand speech as well as a 3 year-old child? One reason is that speech recognition systems use just one or only a few sources of information. People, on the other hand, seem to use many sources of information and are able to combine several them in an optimal fashion. A second reason is that computers are programmed to process clear-cut categories. Interpreting ambiguous or fuzzy data, however, is natural for humans. This is best seen in face-to-face communication. Experiments have revealed conclusively 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, 1998). For example, if the ambiguous auditory sentence, *My bab pop me poo brive*, is paired with the visible sentence, *My gag kok me koo grive*, the perceiver is likely to hear, *My dad taught me to drive.* Two ambiguous sources of information are combined to create a meaningful interpretation (Massaro & Stork, 1998).

Information in the face is particularly effective when the auditory speech is degraded, because of noise, limited bandwidth, or hearing-impairment. However, the strong influence of visible speech is not limited to situations with degraded auditory input. A perceiver's recognition of an auditory-visual syllable reflects the contribution of both sound and sight. Our most recent findings show that speechreading, or the ability to obtain speech information from the face, is not compromised by oblique views, partial obstruction or visual distance. Humans are fairly good at speechreading even if they are not looking directly at the talker's lips. 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.

With our completely animated, synthetic, talking head we can control the parameters of visible speech and study its informative aspects. Figure 1 shows the talking head, called Baldi. As can be seen in the Figure, there is not much behind Baldi's attractive exterior. His existence and functionality depend on computer animation and text-to-speech synthesis. His speech is controlled by 33

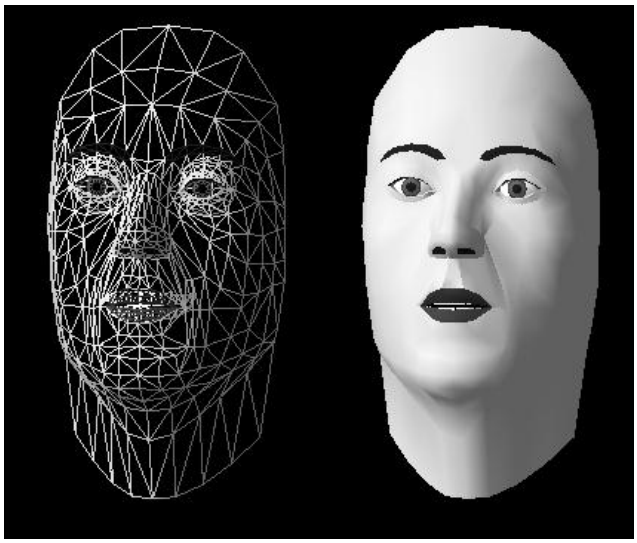

Figure 1 [http://mambo.ucsc.edu/psl/pela/wg.jpg] shows the talking head, called Baldi. As can be seen in the Figure, there is not much behind his attractive exterior.

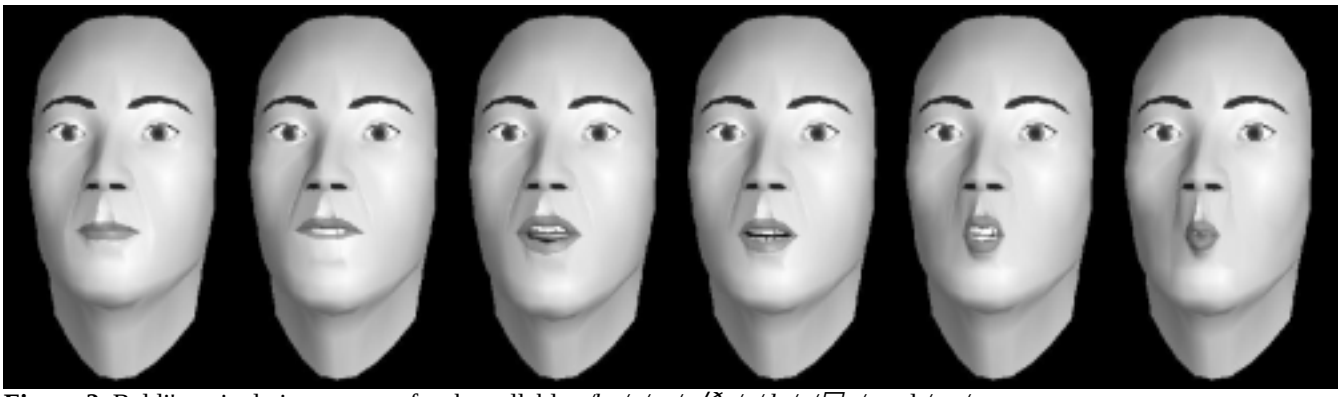

Figure 2. Baldi's articulation at onset for the syllables */ba/, /va/, /ða/, /da/, /□a/,* and */wa/.*

parameters including; jaw rotation and thrust, horizontal mouth width, lip corner and protrusion controls, lower lip f tuck, vertical lip position, horizontal and vertical teeth offset, tongue angle, width, and length. Figure 2 illustrates Baldi's articulation at onset for the syllables */ba/, /va/, /ða/, /da/, □a/,* and */wa/.* Experiments by Cohen, Walker, and Massaro (1996) and Massaro (1998) have shown that visible speech produced by the synthetic head, even in its adumbrated form, is almost comparable to that of a real human.

A primary objective of our research is to identify the informative properties of the human face by evaluating the effectiveness of various properties in our synthetic face. The value of facial animation in the development of synthetic speech is analogous to the important contribution of auditory speech synthesis in speech perception research. The development of a realistic, high-quality, facial display has provided a powerful tool to continue the investigation of a number of questions in auditory-visual speech perception. This visible speech synthesis permits the type of experimentation necessary to determine 1) what properties of visible speech are used, 2) how they are processed, and 3) how this information is integrated with auditory and other contextual sources of information in speech perception. In some of our research we systematically manipulate audible and visible speech independent of one another (Massaro, 1998).

Analysis of real speech articulation has guided our research in visible speech synthesis. Perception experiments have indicated how well the synthesis simulates real speakers. An understanding of visible speech perception derived from these experiments has assisted in our development of visible speech. Our goal at the Perceptual Science Laboratory (PSL) has been to create a talking head whose facial motions look realistic, not to duplicate the musculature of the face. We have chosen to develop visible speech synthesis in the same manner that has proven successful with audible speech synthesis. We call this technique terminal analogue synthesis. Its goal is, simply, to mimic the final speech product rather than the physiological mechanisms that produce it. One advantage of terminal analogue synthesis is that calculations for changing the surface shapes of the polygon models can be carried out much more rapidly than calculations for muscle and tissue simulations. It also may be easier to achieve the desired facial shapes directly rather than in terms of the constituent muscle actions. The real-time animation of the synthetic head (at up to 60 frames per second) was developed on a Silicon Graphics Crimson Reality Engine with 96 megabytes of RAM and a 100MHz R4000 microprocessor using the IrisGL graphics library. Because our implementation is efficient, we have been able to port the animation algorithm to a PC platform, (an Intel Pentium-based Windows NT or Windows 95 platform, using the newer OpenGL graphics library,) for use in language training with profoundly deaf children at the Tucker Maxon Oral School in Portland, Oregon. (Cole et al., 1998).

The talking head is made of polygons that have been joined together and smooth shaded. The polygon topology and animation is controlled through a set of parameters (Cohen & Massaro, 1993), and is made up of approximately 900 surfaces connected at the edges to create the 3-D head with eyes, pupil, iris, sclera, eyebrows, nose, skin, lips, tongue, teeth, and neck (see Figure 1). The tongue is implemented as a parametrically-controlled shaded surface made of a polygon mesh.

Our implementation of visible speech synthesis has progressed over the last 3 years to include additional and modified control parameters, two generations of tongues, a visual speech synthesis control strategy, text-to-speech synthesis, bimodal (auditory/visual) synthesis, and controls for paralinguistic information and affect in the face. Figure 3 illustrates Baldi's facial expressions for *happy, angry, surprise, fear, sadness, and disgust.* Most of our current parameters move points on the face geometrically by rotation (e.g. jaw rotation) or translation, in one or more dimensions (e.g., lower and upper lip height, mouth widening). Other parameters are changed by interpolation between patches on alternate faces. Examples of this type of parameter include cheek shape, neck shape, and smiling. The synthesis program, which consists of about 20,000 lines of C code, runs in real-time on both SGI and PC platforms.

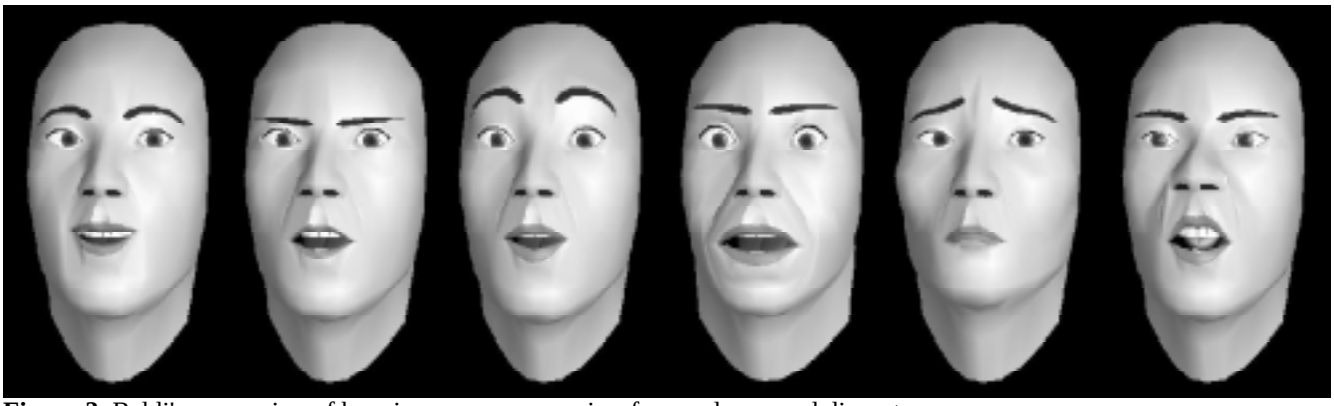

Figure 3. Baldi's expression of happiness, anger, surprise, fear, sadness, and disgust.

Evaluating Talking Heads

While developing the animation and speech synthesis, we have continuously evaluated and improved its speech. It is a sobering fact that auditory speech synthesis is still falls far short of natural speech after 30 years of intensive research and development (Cohen et al., 1996; Massaro, 1998, Chapter 13). We have not allowed ourselves to be swayed by the subjective comments of many viewers who claim, "How natural Baldi seems." Instead, we have systematically performed experiments to compare the quality of the synthetic speech to natural speech. The relative realism of Baldi's visible speech is measured in terms of its intelligibility to speechreaders. The experiments measure comparative intelligibility to determine where and how Baldi falls short of natural speakers. The synthesis is then modified accordingly - bringing it more in line with natural visible speech. In a series of over a dozen such evaluation experiments, with appropriate adjustments to the synthesis, the quality of the visible speech has become almost as good as a very good natural speaker. (Massaro, 1998).

In a typical study, we presented silently for identification monosyllabic English words (e.g. *sing, bin, dung, dip, seethe*) produced either by a natural speaker (Bernstein & Eberhardt videodisk, 1986) or our synthetic talker randomly intermixed. The synthetic stimuli used a specific set of parameter values and dominance functions for each phoneme and our blending function for coarticulation. The AT&T text-to-speech (TtS) module was utilized to provide the phonemic representation for each word and the durations of the speech segments, in addition to synthesizing the auditory speech (presented as feedback) (Olive, 1990). Other characteristics such as speaking rate and average acoustic amplitude were equated for the natural and synthetic talker. The speech on the videodisk was articulated in citation form and thus had a relatively slow speaking rate

College students who were native speakers of American English served as subjects, in two 40-minute sessions each day for two days. Up to four at a time were tested in separate sound attenuated rooms under control of the SGI-Crimson computer, with video from the laserdisk (the human talker) or the computer being presented over 13" color monitors. On each trial they were first presented with a silent word from one of the two faces and then typed in their answer on a terminal keyboard. Only actual monosyllabic English words were accepted as valid answers from a list of about 12,000 derived mainly from the Oxford English dictionary. After all subjects had responded, they received feedback by a second presentation of the word, this time with auditory speech (natural or synthetic) and with the word in written form on the left side of the video monitor. There were 264 test words, and each word was tested with both synthetic and natural speech, for a total of 2 x 264 = 528 test trials.

By comparing the overall proportion correct and analyzing the perceptual confusions, we can determine how closely the synthetic visual speech matches the natural visual speech. The questions to be answered are what is the extent of confusions and how similar are the patterns of confusions for the two talkers. This analysis is be simplified by ignoring confusions that take place between visually similar phonemes. Because of the data-limited property of visible speech in comparison to audible speech, many phonemes are virtually indistinguishable by sight, even from a natural face, and so are expected to be easily confused. To eliminate these non-serious confusions from consideration, we group visually indistinguishable phonemes into categories called viseme. Some confusions do take place between viseme categories, however. This is partly because of the difficulty of speechreading. But also, as with most categories, visemes are not sharply defined (i.e. they are "fuzzy"), and any sharp definitions imposed are therefore somewhat arbitrary and inaccurate. Even so, it is worthwhile to use some standard viseme groupings in order to assess how well the more meaningful visible speech differences are perceived.

We also found that the confusion matrices for natural and synthetic speech are very similar to one another. Figure 4 presents the word-initial consonant viseme confusions for a

typical recent unpublished experiment for natural (left panel) and synthetic (right panel) speech. The area of each circle indicates the proportion of each response to a given stimulus. Of course we don't expect perfect performance for either talker, but it is the similarity of the pattern of responses that is of interest. We computed the correlation of the synthetic and natural talker data which yielded a correlation of r=.927. The overall proportion correct responses for the natural speech (.689) was slightly higher than that for the synthetic talker (.652).

One additional type of analysis that we carried out comparing the synthetic and natural talkers was the information transmission metric which we have previously used in assessing the degrading effect of spatial quantization on perception of visible speech (Campbell & Massaro, 1997). The results of this analysis is that for initial consonant phoneme perception, the percentage of information transmitted was 44.1% for synthetic speech versus 41.7% for natural, and for initial consonant viseme, 56.1% for synthetic versus 62.1% for natural. While these findings are not exactly congruent with the identification performance statistic, they do indicate that the two talkers are roughly equivalent in terms of their informativeness. We have significantly improved the quality of our synthetic visible speech over the course of our successive modifications and tests. The overall viseme accuracy across four successive studies improved significantly. The average deficit relative to natural speech was .222, .179, .104, and .086 across the four successive experiments. Given the proposed measurement, synthesis, and evaluation studies, we are optimistic that our synthetic speech will be very close to natural speech

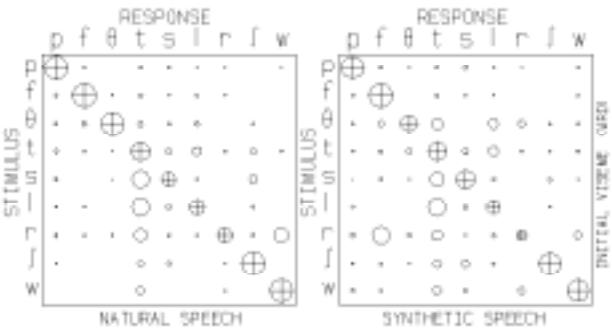

Figure 4. Proportion of viseme responses as a function of initial stimulus viseme of the test word. The area of the circle is proportional the value.

Sometimes uncontrolled experiences with new technology are as informative as systematic studies. In addition to the experimental evidence, we have inadvertently obtained expert testimonials. A deaf speech scientist claims that Baldi's is the only existing synthetic visible speech that he can accurately speechread. Also, a deaf man, Daniel Souther, from San Antonio, Texas, has independently downloaded Baldi for use in speech production training. He has formally presented this technology to his college, Sunshine College, with the hope of making it accessible to all of the students there.

Robustness of Bimodal Speech Perception

There are several impressive properties of visible speech and bimodal speech perception. First, people naturally integrate visible and auditory speech in a variety of situations. Speechreading, or the ability to obtain speech information from the face, is not compromised by a number of variables. Humans are fairly good at speechreading even if they are not looking directly at the talker's lips. Furthermore, accuracy is not dramatically reduced when the facial image is blurred (because of poor vision, for example), when the image is rotated towards a profile view, when viewed from above or below, and when there is a large distance between the talker and the viewer. Another example of the robustness of the influence of visible speech is 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. These findings indicate that speechreading is highly functional in a variety of nonoptimal situations. It follows that the pursuit of visible speech technology could be of great practical value in many spheres of communication.

The Fuzzy Logic Model of Perception

Our work has combined sophisticated experimental designs and quantitative model testing to understand speech perception and pattern recognition more generally. A wide variety of results have been described within a framework of a fuzzy logical model of perception (FLMP). Figure 5 gives a schematic representation of the three processes involved in perceptual recognition. The three processes are shown in sequence, left to right, to illustrate their necessarily successive, but overlapping, procedures.

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 assumptions central to the model are: 1) each source of information is evaluated to determine the degree to which that source specifies various alternatives, 2) the sources of information are evaluated independently of one another, 3)

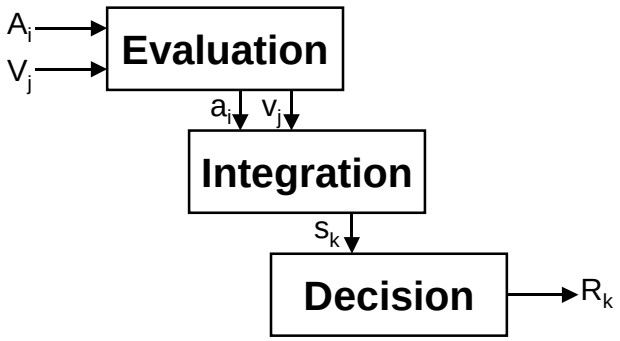

Figure 5. Schematic representation of the three processes involved in perceptual recognition.

the sources are integrated 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. In the course of our research, we have found the FLMP to be a universal principle of perceptual cognitive performance that accurately models human pattern recognition. People are influenced by multiple sources of information in a diverse set of situations. In many cases, these sources of information are ambiguous and any particular source alone does not usually specify completely the appropriate interpretation.

Outcomes from a variety of experiments are well fit by the implementation of our universal principle (Massaro, 1998). For example, Baldi has been used to study how people perceive emotion based on information in the face and in the voice. Two features of facial affect, brow deflection and mouth deflection, were manipulated systematically. An expanded-factorial design was used, with four levels of brow deflection crossed with four levels of mouth deflection, as well as their corresponding half-face conditions. Participants identified these faces as either happy or angry. Both the brow and the mouth influenced their judgments, and these cues were combined in an optimal way as described by the FLMP. Another study examined how emotion is perceived from both facial and vocal cues. Participants were presented all possible permutations, i.e. visual cues alone, vocal cues alone and visual and vocal cues together. When asked to judge the emotion as either happy or angry, participants appear to evaluate and integrate information from both modalities to perceive emotion. The influence of one modality is greater to the extent that the other is ambiguous, a result well described by the FLMP. We have learned that people use many sources of information in perceiving and understanding speech, emotion, and other aspects of the environment. The results from a many studies are consistent with the FLMP, which describes a universal law of behavior - that people naturally use both the information in the speaker's face and the sound to perceive and understand the message.

In speech perception multiple sources of information are available to support the identification and interpretation of language. The experimental paradigm that we have developed allows us to determine which of the many potentially functional cues are actually used by human observers (Massaro, 1998, Chapter 1). These results show how visible speech is processed and integrated with other sources of information. The systematic variation of the properties of the speech signal and quantitative tests of models of speech perception allow the investigator to interpret 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 addresses how different sources of information are evaluated and integrated, and can identify the sources of information that are actually used.

This knowledge provides the basis for the theory of interface design, which is central to the proposed research. Our experience in the design and implementation of interactive interfaces based on verbal, vocal, gestural, and tactile input (including a hardware-free virtual reality interface populated by video avatars and active agents) supports this theory. A thorough aesthetic, conceptual and psychological evaluation of a wide range of existing interfaces has led us to conclude that users will be empowered if they are afforded greater control over their systems and communication will be enhanced by more natural, conversational interfaces.

Communicating Paralinguistic Information

One of the most frequent complaints about synthetic speech is that it is very monotonous. The stress and intonation pattern is far removed from natural speech. The presence of a talking face has the potential to improve not only the perceived quality of the segmental articulations but also the intonation pattern. We use facial expressions and corresponding gestures that communicate intonation. For example, in our case, we use the height of the eyebrows since there is evidence of a positive correlation between eyebrow height and voice pitch (Ohala, 1994). Cave et al. (1996), for example, found that 71% of F0 rise were accompanied by eyebrow rises. Some linguists have remarked that doing their transcriptions as students was made easier by surreptitiously watching their teachers' eyebrows. Along these lines, Pelachaud, Badler, and Steedman (1996) have extended the Facial Action Coding System (FACS, Ekman & Friesen, 1977) approach to simulate facial expressions conveying information that is normally correlated with intonation of the voice.

An important dimension of communication is emotional expression, which will also be developed as an integral characteristic of our conversational agent. There is no doubt that the production of facial expressions is an effective means of communicating emotion. Although

many different facial movements are possible, human facial expressions tend to be classifiable within relatively few emotional categories. Some evidence exists, based on studies of animals, as well as preliterate, and isolated societies, that the production of facial expressions such as rage, surprise, startle, fear, and pleasure, are reliably developing (innate) and universal.

We have built emotional cues into our synthetic face based on empirical research. Baldi now conveys six basic emotions – *happiness, anger, fear, disgust, surprise and sadness* – in a very realistic manner (see Figure 3). For example, brow displacement and mouth corner displacement are varied to create happy and angry. A slightly elevated and arched brow and a mouth fully curled up at the corner create a prototypically happy expression. A fully depressed and flattened brow and a mouth fully curled down at the corner create a prototypically angry expression. These emotional expressions are implemented to convey the appropriate emotional state of our conversational agent, tutor, or whatever role the synthetic actor happens to be playing. We plan to examine the hypothesis that communication dialog is more efficient and enjoyable when emotional expression is inherently part of the message.

Baldi's Interface Environment

The design of the agents' interface environment is crucial to the successful and natural transfer of information. Currently our synthetic talking head floats in a void. It can be rotated, translated and scaled. Baldi is, in essence, a mask. When the head is rotated the inside of the mask is exposed. This works well for applications in speech training and language acquisition since users often need to examine the relation of tongue and teeth, etc. As we increase the realism of the model with the addition of hands and shoulders, back, top, side of head and neck, we intend to maintain this functionality. We are providing tools that allow the user to deconstruct and control the transparency of the model for access to detailed physiological views in speech training applications.

Adding additional features, however, particularly shoulders and hands, will necessitate consideration of a more natural relation of the figure in its screen space. In other words, the view of the model will have to be cropped in some way. The role of the agent or avatar in the context of the user's screen space will be addressed. We intend to design active objects and icons that will be integrated into the agent/avatars environment. The user will define their functionality. The use of background textures as aids to communication and expression will be examined. While the agent will be able to interact with the active objects and icons in its own environment the role of the agent in the user's operating and communications systems should also be considered. For example, the agent's spatial environment might be visually and functionally integrated into the user's desktop or browser. A "window or desktop" would be transformed into a "conversation space," which would support information architectures, persistent communications and social relationships within functioning virtual communities.

Language Tutoring

Our facial animation software is currently an integral part of a three-year project, funded by an NSF Challenge grant, now about a year old, to develop interactive learning tools for language training with profoundly deaf children. In addition to Baldi, the tools to date have combined key technologies: speech recognition, developed at the Oregon Graduate Institute; and speech synthesis, developed at the University of Edinburgh and modified at OGI. The goal of the project is to provide teachers, students, parents and other interested individuals with state of the art tools and technologies for interactive learning and language training. These tools and technologies are integrated into the CSLU toolkit, a software environment for building and researching interactive systems (http://www.cse.ogi.edu/CSLU/toolkit/; Cole et a., 1998).

To date, language-training applications have been developed using Baldi as a language tutor. The acoustic speech is provided by either the Festival TTS system or a natural speaker synchronized with the animated face. The speaker records an utterance and types in the words. The system then produces a time-aligned phonetic transcription that is used to drive the synthetic face. When facial animation and speech generation are combined with computer speech recognition, students are able to have limited conversations with Baldi. These educational modules are created by CSLU's toolkit developers and by educators at the Tucker-Maxon Oral School, using the toolkit's graphical authoring environment, the Rapid Application Developer (RAD).

Language training software is being developed and tested in collaboration with the educators and deaf students at the Tucker-Maxon Oral School in Portland, Oregon. The students in this school are profoundly deaf. Hearing aids, cochlear implants, or a combination of both are used to enhance their hearing. During the first year of the project, our main challenge has been to adapt and extend the toolkit to the needs of the teachers and students. To this end, we have followed principles of participatory design, in which the users of the software participate as much as possible in all phases of its design and development.

In the class of the oldest students between 9 and 12 years old, instructor George Fortier has taken full advantage of the RAD to create interactive multi-media lessons. These dialogues introduce and/or review concepts and vocabulary from the social studies and science curricula. For example, he created a geography lesson by scanning in a drawing that depicted a mountain range, lake, river, stream, plateau,

waterfall, coast, and ocean. When a student starts this program, one of the geographic features is selected randomly and highlighted. Baldi asks: "What landform is this?" A correct answer by the student causes the system to display another highlighted landform and ask another question. If the child does not know the answer she says, "Help." Baldi then says the word three times at a slightly slower rate and asks the question again. If the system does not recognize the student's speech Baldi asks the student to try again. After two recognition failures, Baldi again says the word three times. This "correct speech routine" continues until the student's speech is recognized. Because of the children's varying levels of speech ability the teacher sets the recognition threshold differently for each child. This threshold is increased daily so that the student's speech quality must improve for recognition to occur. This program has been an effective learning tool. None of the six students knew all of these landforms before using the application. After four days of working individually, all 6 students knew the names of each landform and could say them with such clarity that the system accepted their answer on the first or second trial.

Mr. Fortier has created 15 different applications for his class. Working on a home PC in the evening, he uses the toolkit's authoring tools to create sophisticated interactive dialogues that incorporate different graphics at specific points in the dialogue. All of these applications exercise and test language comprehension, speech production and content learning.

Figure 6. Baldi engaging students in the classroom.

Alice Davis, the instructor of the younger students between 7 and 9 years old, has created a variety of applications for different content areas to supplement her curriculum (see Figure 6). For example, she has built modules for the study of math, spelling, reading, listening comprehension and writing. She has even used the toolkit as a way for the children to listen to and practice reciting their own haiku. To test math concepts, Ms. Davis created interactive math quizzes. For example, the child was shown a picture of seven bears in a forest. Baldi would then say: "There are seven bears in the forest. Three bears ran away. How many bears are left?" When the child produced the correct answer, she received the next question. Applications of this type sharpen speechreading and listening skills and speech production skills, while testing knowledge of content.

The speech therapist, Chris Soland, in collaboration with Tim Carmell, developed a series of listening drills based on minimal pair distinctions. For example, the screen shows a picture of "mail" and "veil". The system then says one of the words, and the student uses a mouse to select the picture that corresponds to the word that she perceived. In another task, the system produces a pair of words, e.g., "veil, veil" or "mail, veil," and the student says "same" or "different" after each pair. In each application, the system produces immediate visual feedback and informs the child of her score at the end of the lesson.

Future Plans

Our technology clearly provides an automated method of providing dedicated one-on-one language training and instruction. We learned that Baldi is highly motivating. Children have persisted in some of the lessons until they were mastered–even though it wasn't required. We expect more than this from animated agents, however. One general expectation is to have agents do what no real person can. To this end, we have augmented the internal structures of our talking head both for improved accuracy and to pedagogically illustrate correct articulation.

Figure 7. New palate and tongue embedded in the talking head.

Figure 7 shows the new palate, teeth, and tongue

embedded in Baldi's mouth. High-resolution models of palate and teeth were reduced to a relatively small number of polygons for real-time animation. We are using 3D ultrasound data and electropalatography (EPG) with error minimization algorithms to educate our parametric B-spline based tongue model to simulate realistic speech. In addition, a high-speed algorithm has been developed for detection and correction of collisions, to prevent the tongue from protruding through the palate and teeth, and to enable the real-time display of synthetic EPG patterns (see Figure 8).

One immediate motivation for developing a hard palate, teeth and tongue is their potential utility in language training. Children with hearing-impairment require guided instruction in speech perception and production. Although there is a long history of using visible cues in speech training for individuals with hearing loss, these cues have usually been abstract or symbolic rather than direct representations of the vocal tract. Some of the distinctions in spoken language cannot be heard with degraded hearing–even when the hearing loss has been compensated by hearing aids or cochlear implants. To overcome this limitation, we plan to use visible speech to provide speech targets for the child with hearing loss. In addition, many of the subtle distinctions among segments are not visible on the outside of the face. The skin of our talking head can be made transparent so that the inside of the vocal track is visible, or we can present a cutaway view of the head along the sagittal plane. The goal is to instruct the child by revealing the appropriate articulation via the hard palate, teeth and tongue.

Visible speech instruction poses many issues that must be resolved before training can be optimized. We are confident that illustration of articulation will be useful in improving the learner's speech, but it will be important to assess how well the learning transfers outside the instructional situation. Another issue is whether instruction should be focused on the visible speech or whether it should include auditory input. If speech production mirrors speech perception, then we expect that multimodal training should be beneficial. We expect that the child could learn multimodal targets, which would provide more resolution than either modality alone. Another issue concerns whether the visible speech targets should be illustrated in static or dynamic presentations. We plan to evaluate both types of presentation and expect that some combination of modes would be optimal. Finally, the size of the instructional target is an issue. Should instruction focus on small phoneme and open-syllable targets, or should it be based on larger units of words and phrases? Again, we expect training with several sizes of targets would be ideal.

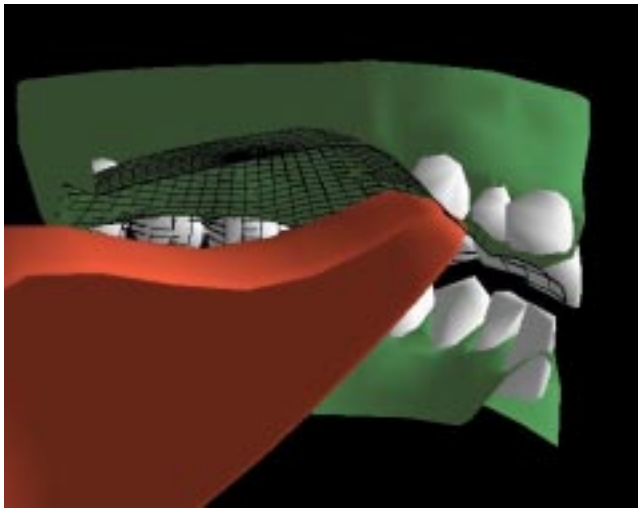

Figure 8. Liner structure shown for palate and upper teeth with longitude and latitude lines. We see the left half of the structures (tongue, palate, gums and teeth) cut at the sagittal plane. The front teeth are to the right in this figure.

Finally, although we have not discussed evaluation as much as the development and application of the technology, it is an integral part of our project. Earlier we described how the quality of Baldi's speech is assessed and compared to that of a real speaker. Although more difficult, we also carry out systematic evaluations of the quality of Baldi's instruction. For language training, we provide periodic tests of both speech perception and speech production, and chart the improvement across training. This improvement is compared to available norms and /or other types of training regiments

Acknowledgments. The research was supported by grants from the National Institute of Deafness and Other Communicative Disorders, the National Science Foundation, and Intel Corporation.

References

Bernstein, L. E., & Eberhardt, S. P. (1986). *Johns Hopkins lipreading corpus videodisk set*. Baltimore, MD: The Johns Hopkins University.

Campbell, C. S., & Massaro, D. W. (1997).Visible Speech Perception: Influence of Spatial Quantization. *Perception, 26*, 627-644.

Cave, C., Guaitella, I., Bertrand, R., Santi, S., Harlay, F. & Espesser, R. (1996). About the relationship between eye-brow movements and F0 variations. Proceedings of ICSLP 96. Wilmington, DE: Univ. of Delaware. 2175-2178.

Clark, A. (1997). Being There: Putting Brain, Body and World Together Again. Cambridge, MA: MIT Press.

Cohen, M. M., Beskow, J., & Massaro, D. W. (1998). Recent developments in facial animation: An inside view. *Proceedings of Auditory-Visual Speech Processing 98*. Sydney, Australia.

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., Walker, R. L.& Massaro, D. W. (1996). Perception of synthetic visual speech. In D. G. Stork & M. E. Hennecke (Eds.) Speechreading by humans and machines (pp. 153-168). New York: Springer-Verlag.

Cole, R. et al. (1998). Intelligent animated agents for interactive language training. Proceedings of Speech Technology in Language learning. Stockholm, Sweden.

Ekman, P. & Friesen, W. V. (1977). Manual for the Facial Action Coding System. Palo Alto, CA: Consulting Psychologists Press.

Gibson, J. J. (1979). *The ecological approach to visual perception*. Boston: Houghton Mifflin.

Glenberg, A. M. (1997). Mental models, space, and embodied cognition. In Creative thought: An investigation of conceptual structures and processes (pp. 495-522); T. B. Ward, S. M. Smith, J. Vaid, (Eds). Washington, D.C. American Psychological Association.

Kiesler, S. (1997). Culture of the Internet. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Lanier, J. (1996). Agents of alienation. (www.well.com/user/jaron/agentalien.html)

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

Massaro, D.W. (1998). Perceiving Talking Faces: From Speech Perception to a Behavioral Principle. Cambridge, MA: MIT Press.

Massaro, D.W., & Stork, D.G. (1998). Speech recognition and sensory integration. American Scientist, 86, 236-244.

Newton, Natika (1996). Foundations of understanding. Philadelphia, PA: John Benjamins Publishing

Ohala, J. J. (1994). The frequency code underlies the sound symbolic use of voice pitch. In L. Hinton, J. Nichols, and J.

Ohala (Eds.), Sound symbolism. New York: Cambridge University Press. 325-347.

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

Pelachaud, C., Badler, N. I., & Steedman, M. (1996). Generating facial expressions for speech. Cognitive Science, 20, 1-46.

Picard, R. W. (1997). Affective computing. Cambridge, MA: The MIT Press;

Reeves, B. & Nass, C. (1996). The media equation: How people treat computers, television, and new media like real people and places. Stanford: CSLI Publications.

Tognazzini, B. (1997). More Than Screen Deep: Toward Every-Citizen Interfaces to the Nation's Information Infrastructure. In A. W. Biermann, T. Bikson, T. Defanti, G. Fischer, B.J. Grosz, T. Landauer, J. Makhoul, B. Tognazzini, G. Vanderheiden, and S. Weinstein (Eds.). National Academy Press, 271-278.