Chapter 9

THE PSYCHOLOGY AND TECHNOLOGY OF TALKING HEADS: APPLICATIONS IN LANGUAGE LEARNING

Dominic W. Massaro
Perceptual Science Laboratory
Department of Psychology, University of California
Santa Cruz, CA 95064, USA
massaro@fuzzy.ucsc.edu

Abstract Given the value of visible speech, our persistent goal has been to develop, evaluate, and apply animated agents to produce accurate visible speech. The goal of our recent research has been to increase the number of agents and to improve the accuracy of visible speech. Perceptual tests indicted positive results of this work. Given this technology and the framework of the fuzzy logical model of perception (FLMP), we have developed computer-assisted speech and language tutors for deaf, hard of hearing, and autistic children. Baldi[1], as the conversational agent, guides students through a variety of exercises designed to teach vocabulary and grammar, to improve speech articulation, and to develop linguistic and phonological awareness. The results indicate that the psychology and technology of Baldi holds great promise in language learning and speech therapy.

Keywords: Facial animation, visible speech, language learning, speech perception, vocabulary tutor, autism, children with language challenges.

1. Introduction

The face presents visual information during speech that is critically important for effective communication. While the auditory signal alone is adequate for communication, visual information from movements of the lips, tongue and jaws enhance intelligibility of the acoustic stimulus (particularly in noisy en-

[1] Baldi® is a registered trademark of Dominic W. Massaro.

J.C.J. van Kuppevelt et al. (eds.), Advances in Natural Multimodal Dialogue Systems, 183–214.

vironments). Moreover, speech is enriched by the facial expressions, emotions and gestures produced by a speaker [Massaro, 1998]. The visual components of speech offer a lifeline to those with severe or profound hearing loss. Even for individuals who hear well, these visible aspects of speech are especially important in noisy environments. For individuals with severe or profound hearing loss, understanding visible speech can make the difference in effectively communicating orally with others or a life of relative isolation from oral society [Trychin, 1997].

Our persistent goal has been to develop, evaluate, and apply animated agents to produce accurate visible speech. These agents have a tremendous potential to benefit virtually all individuals, but especially those with hearing problems (> 28,000,000 in the USA), including the millions of people who acquire age-related hearing loss every year (http://www.nidcd.nih.gov/health/hb.htm), and for whom visible speech takes on increasing importance. One of many applications of animated characters allows the training of individuals with hearing loss to "read" visible speech, and thus facilitate face-to-face oral communication in all situations (educational, social, work-related, etc). These enhanced characters can also function effectively as language tutors, reading tutors, or personal agents in human machine interaction.

For the past ten years, my colleagues and I have been improving the accuracy of visible speech produced by an animated talking face - Baldi (Figure 9.1; [Massaro, 1998, chapters 12-14]). Baldi has been used effectively to teach vocabulary to profoundly deaf children at Tucker-Maxon Oral School in a project funded by an NSF Challenge Grant [Barker, 2003; Massaro, 2000]. The same pedagogy and technology has been employed for language learning with autistic children [Massaro et al., 2003]. While Baldi's visible speech and tongue model probably represent the best of the state of the art in real-time visible speech synthesis by a talking face, experiments have shown that Baldi's visible speech is not as effective as human faces. Preliminary observations strongly suggest that the specific segmental and prosodic characteristics are not defined optimally. One of our goals, therefore, is to significantly improve the communicative effectiveness of synthetic visual speech.

2. Facial Animation and Visible Speech Synthesis

Visible speech synthesis is a sub-field of the general areas of speech synthesis and computer facial animation (Chapter 12 in [Massaro, 1998] organizes the representative work that has been done in this area). The goal of the visible speech synthesis in the Perceptual Science Laboratory (PSL) has been to develop a polygon (wireframe) model with realistic motions (but not to duplicate the musculature of the face to control this mask). We call this technique terminal analogue synthesis because its goal is to simply use the final speech

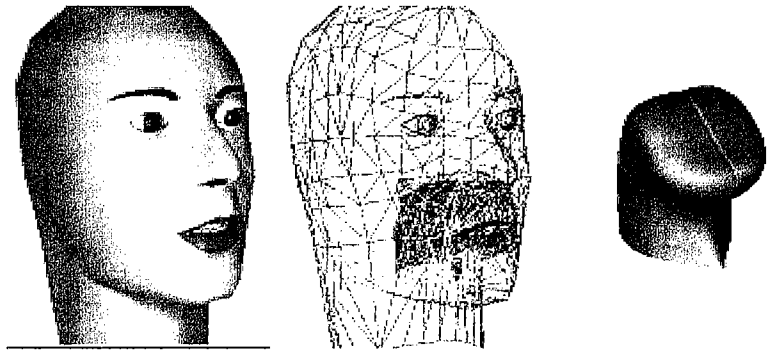

Figure 9.1. Baldi, a computer-animated talking head, in normal and wireframe presentations, and a close-up of the tongue.

product to control the facial articulation of speech (rather than illustrate the physiological mechanisms that produce it). This method of rendering visible speech synthesis has also proven most successful with audible speech synthesis. One advantage of the terminal analogue synthesis is that calculations of the changing surface shapes in the polygon models can be carried out much faster than those for muscle and tissue simulations. For example, our software can generate a talking face in real time on a commodity PC, whereas muscle and tissue simulations are usually too computationally intensive to perform in real time [Massaro, 1998]. More recently, image synthesis, which joins together images of a real speaker, has been gaining in popularity because of the realism that it provides. These systems also are not capable of real-time synthesis because of their computational intensity. Finally, performance-based synthesis, e.g., [Guenter et al., 1998], does not have the flexibility of saying anything at any time in real time, as does our text-to-speech system.

Our own current software [Cohen and Massaro, 1993; Cohen et al., 1996; Cohen et al., 1998; Massaro, 1998] is a descendant of Parke's software and his particular 3-D talking head [Parke, 1975]. Our modifications over the last 8 years have included increased resolution of the model, additional and modified control parameters, three generations of a tongue (which was lacking in Parke's model), a new visual speech synthesis coarticulatory control strategy, controls for paralinguistic information and affect in the face, alignment with natural speech, text-to-speech synthesis, and bimodal (auditory/visual) synthesis. Most of our current parameters move vertices (and the polygons formed from these vertices) on the face by geometric functions such as rotation (e.g. jaw rotation) or translation of the vertices in one or more dimensions (e.g., lower and upper lip height, mouth widening). Other parameters work

by scaling and interpolating different face subareas. Many of the face shape parameters–such as cheek, neck, or forehead shape, and also some affect parameters such as smiling–use interpolation.

We have used phonemes as the basic unit of speech synthesis. In this scheme, any utterance can be represented as a string of successive phonemes, and each phoneme is represented as a set of target values for the control parameters such as jaw rotation, mouth width, etc. Because speech production is a continuous process involving movements of different articulators (e.g., tongue, lips, jaw) having mass and inertia, phoneme utterances are influenced by the context in which they occur by a process called coarticulation. In our visual speech synthesis algorithm [Cohen and Massaro, 1993] [Massaro, 1998, chapter 12], coarticulation is based on a model of speech production using rules that describe the relative dominance of the characteristics of the speech segments. In our model, each segment is specified by a target value for each facial control parameter. For each control parameter of a speech segment, there are also temporal dominance functions dictating the influence of that segment over the control parameter. These dominance functions determine independently for each control parameter how much weight its target value carries against those of neighbouring segments, which will in turn determine the final control values.

Our animated face can be aligned with either the output of a speech synthesizer or natural auditory speech [Massaro et al., 2005]. We have also developed the phoneme set and the corresponding target and coarticulation values to allow synthesis of several other languages. These include Spanish (Baldero), Italian (Baldini, [Cosi et al., 2002]), Mandarin (Bao), Arabic (Badr, [Ouni et al., 2003]), French (Balduin) and German (Balthasar). Baldi, can be seen at: http://mambo.ucsc.edu.

Baldi's synthetic tongue is constructed of a polygon surface defined by sagittal and coronal b-spline curves (see Figure 9.1). The control points of these b-spline curves are controlled singly and in pairs by speech articulation control parameters. There are now 9 sagittal and 3 * 7 coronal parameters that are modified to mimic natural tongue movements. The tongue, teeth, and palate interactions during speaking require an algorithm to prevent the tongue from going into rather than colliding with the teeth and palate. To ensure this, we have developed a fast collision detection method to instantiate the appropriate interactions. Two sets of observations of real talkers have been used to inform the appropriate movements of the tongue. These include 1) three dimensional ultrasound measurements of upper tongue surfaces and 2) EPG data collected from a natural talker using a plastic palate insert that incorporates a grid of about a hundred electrodes that detect contact between the tongue and palate at a fast rate (e.g. a full set of measurements 100 times per second). These measurements were made in collaboration with Maureen Stone at John Hopkins University. Minimization and optimization routines are used to create animated tongue movements that mimic the observed tongue movements [Cohen et al., 1998].

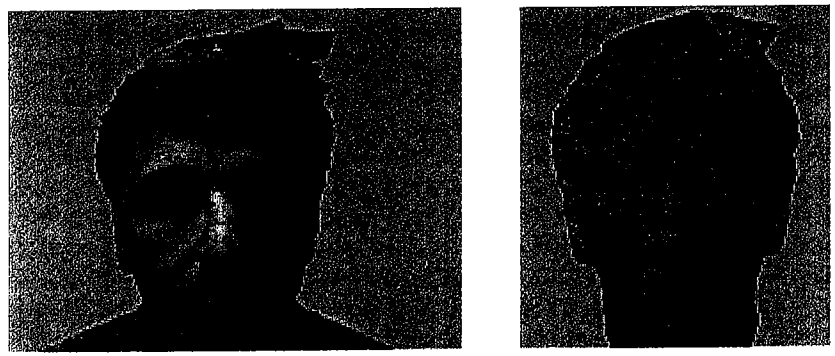

Figure 9.2. High-resolution texture (left) and a high-resolution polygon mesh (right) obtained from a Cyberware laser scan.

2.1 Recent Progress in Visible Speech Synthesis

Important goals for the application of talking heads are to have a large gallery of possible agents and to have highly intelligible and realistic synthetic visible speech. Our development of visible speech synthesis is based on facial animation of a single canonical face, called Baldi (see Figure 9.1; [Massaro, 1998]. Although the synthesis, parameter control, coarticulation scheme, and rendering engine are specific to Baldi, we have developed software to reshape our canonical face to match various target facial models. To achieve realistic and accurate synthesis, we use measurements of facial, lip, and tongue movements during speech production to optimize both the static and dynamic accuracy of the visible speech. This optimization process is called minimization because we seek to minimize the error between the empirical observations of real human speech and the speech produced by our synthetic talker [Cohen et al., 1998; Cohen et al., 2001; Cohen et al., 2002].

2.2 Improving the Static Model

A Cyberware 3D laser scanning system is used to enroll new citizens in our gallery of talking heads. To illustrate this procedure, we describe how a Cyberware laser scan of DWM was made, how Baldi's generic morphology was mapped into the form of DWM, how this head was trained on real data, and how the quality of its speech was evaluated. A laser scan of a new target head produces a very high polygon count representation. Figure 9.2 shows a high-

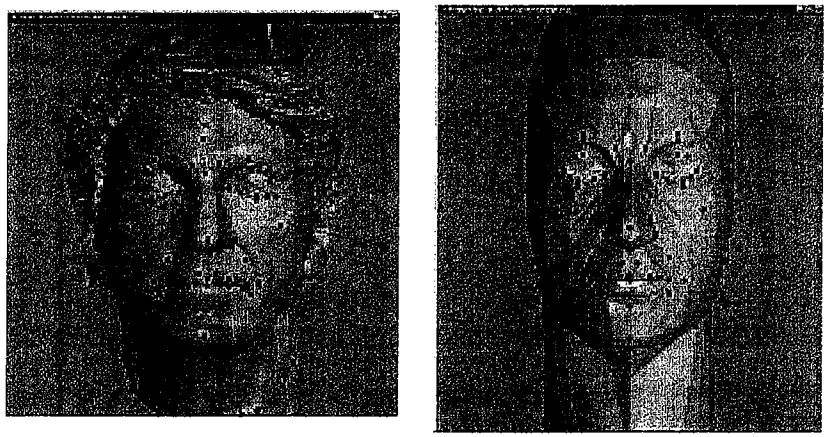

Figure 9.3. Laser scan of high resolution head and canonical Baldi low resolution head with alignment points.

resolution texture mapped Cyberware scan of DWM and the accompanying high-resolution mesh. Rather than trying to animate this high-resolution head (which is impossible to do in real-time with current hardware), our software uses these data to reshape our canonical head to take on the shape of the new target head. In this approach, a human operator marks corresponding facial landmarks on both the laser scan head and the generic Baldi head (Figure 9.3). Our canonical head is then warped until it assumes as closely as possible the shape of the target head, with the additional constraint that the landmarks of the canonical face move to positions corresponding to those on the target face.

This morphing algorithm is based on the work of Kent, Carlson, and Parent [1992]. In this approach, all the triangles making up the source and target models are projected on a unit sphere centred at the origin. The models must be convex or star shaped so that there is at least one point within the model from where all vertices of all triangles are visible. This can be confirmed by a separate vertex visibility test procedure that checks for this requirement. If a model is non-convex or non-star shaped, then it may be necessary to ignore or modify these sections of the model. In order to meet this requirement, portions of the ears, eyes, and lips are handled separately from the rest of Baldi's head.

For the main portion of the head, we first translate all vertices so that the centre point of the model coincides with the coordinate system origin. We then move the vertices so that they are at a unit distance from the origin. At this point, the vertices of the triangles making up the model are on the surface of the unit sphere. This is done to both Baldi's source head and the Cyberware laser scan target head. The landmarks are then connected into a mesh of their

own. As these landmarks are moved into their new positions, the non-landmark points contained in triangles defined by the landmark points are moved to keep their relative positions within the landmark triangles. Then, for each of these source vertices we determine the location on the target model to which a given source vertex projects. This gives us a homeomorphic mapping (1 to 1 and onto) between source and target datasets, and we can thereby determine the morph coordinate of each source vertex as a barycentric coordinate of the target triangle to which it maps. This mapping guides the final morph between the source and target datasets.

A different technique is used to interpolate polygon patches, which were earlier culled out of the target model on account of being non-convex. These patches are instead stretched to fit the new boundaries of the culled regions in the morphed head. Because this technique does not capture as much of the target shape's detail as our main method of interpolation, we try to minimize the size of the patches that are culled in this manner. To output the final topology the program then reconnects all the source polygonal patches and outputs them in a single topology file. The source connectivity is not disturbed and is the same as the original source connectivity.

2.3 Improving the Dynamic Model

To improve the intelligibility of our talking heads, we have developed software for using dynamic 3D optical measurements (Optotrak) of points on a real face while talking. In one study, we recorded a large speech database with 19 markers affixed to the face of DWM at important locations [Cohen et al., 2002].

Fitting of these dynamic data occurred in several stages. To begin, we assigned points on the surface of the synthetic model that best correspond to the Optotrak measurement points. In the training, the Optotrak data were adjusted in rotation, translation, and scale to best match the corresponding points marked on the synthetic face.

The data collected for the training consisted of 100 CID sentences recorded by DWM speaking in a fairly natural manner. In the first stage fit, for each time frame (30 fps) we automatically and iteratively adjusted 11 facial control parameters of the face to get the best fit (the least sum of squared distances) between the Optotrak measurements and the corresponding point locations on the synthetic face. In the second stage fit, the goal was to tune the segment definitions (parameter targets, dominance function strengths, attack and decay rates, and peak strength time offsets) used in our coarticulation algorithm [Cohen and Massaro, 1993] to get the best fit with the parameter tracks obtained in the first stage fit. We first used Viterbi alignment on the acoustic speech data of each sentence to obtain the phoneme durations used to synthesize each

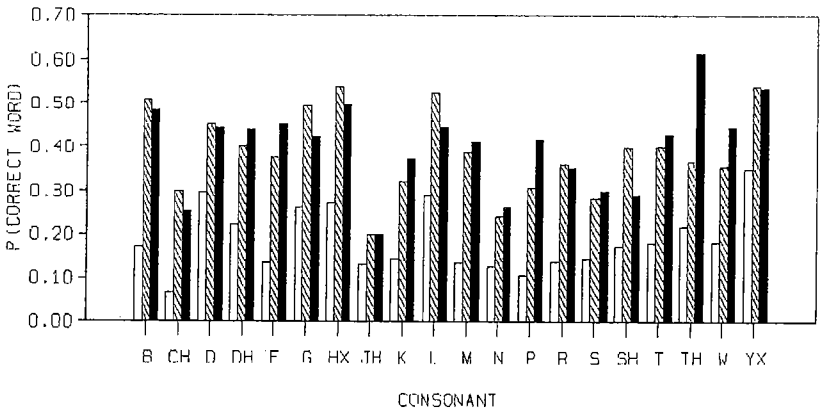

Figure 9.4. Proportion words correct as a function of initial consonant of all words in the test sentences for auditory alone, synthetic and real face conditions.

sentence. Given the phonemes and durations, we used our standard parametric phoneme synthesis and coarticulation algorithm to synthesize the parameter tracks for all 100 CID sentences. These were compared with the parameter tracks obtained from the first stage fit, the error computed, and the parameters adjusted until the best fit was achieved.

2.4 Perceptual Evaluation

We carried out a perceptual recognition experiment with human subjects to evaluate how well this improved synthetic talker conveyed speech information relative to the real talker. To do this we presented the 100 CID sentences in three conditions: auditory alone, auditory + synthetic talker, and auditory + real talker. In all cases there was white (speech band) noise added to the audio channel. Each of the 100 CID sentences was presented in each of the three modalities for a total of 300 trials. Each trial began with the presentation of the sentence, and subjects then typed in as many words as they could recognize. Students in an introductory psychology course served as subjects.

Figure 9.4 shows the proportion of correct words reported as a function of the initial consonant under the three presentation conditions. There was a significant advantage of having the visible speech, and the advantage of the synthetic head was equivalent to the original video of the real face. Overall, the proportion of correctly reported words for the three conditions was 0.22 auditory, 0.43 synthetic face, and 0.42 with the real face.

The results of the current evaluation study, using the stage 1 best fitting parameters is encouraging. In studies to follow, we'll be comparing performance

with visual TTS synthesis based on the segment definitions from the stage 2 fits, both for single segments, context sensitive segments, and also using concatenation of diphone sized chunks from the stage 1 fits. In addition, we will be using a higher resolution canonical head with many additional polygons and an improved texture map.

3. Speech Science

Speech science evolved as the study of a unimodal phenomenon. Speech was viewed as a solely auditory event, as captured by the seminal speech-chain illustration shown in [Denes and Pinson, 1963]. This view is no longer viable as witnessed by a burgeoning record of research findings. Speech as a multimodal phenomenon is supported by experiments indicating that our perception and understanding are influenced by a speaker's face and accompanying gestures, as well as the actual sound of the speech. Many communication environments involve a noisy auditory channel, which degrades speech perception and recognition. Visible speech from the talker's face improves intelligibility in these situations. Visible speech also is an important communication channel for individuals with hearing loss and others with specific deficits in processing auditory information.

The number of words understood from a degraded auditory message can often be doubled by pairing the message with visible speech from the talker's face [Jesse et al., 2001]. The combination of auditory and visual speech has been called super-additive because their combination can lead to accuracy that is much greater than accuracy on either modality alone. Furthermore, 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. 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, 1998].

3.1 Value of Multimodal Speech

There are several reasons why the use of auditory and visual information together in face to face interactions is so successful. These include a) robustness of visual speech, b) complementarity of auditory and visual speech, and c) optimal integration of these two sources of information. Speech reading, or the ability to obtain speech information from the face, is robust in that perceivers are fairly good at speech reading even when 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 [Massaro, 1998, Chapter 14].

Complementarity of auditory and visual information simply means that one of the sources is strong when the other is weak. A distinction between two segments robustly conveyed in one modality is relatively ambiguous in the other modality. For example, the place difference between /ba/ and /da/ is easy to see but relatively difficult to hear. On the other hand, the voicing difference between /ba/ and /pa/ is relatively easy to hear but very difficult to discriminate visually. Two complementary sources of information make their combined use much more informative than would be the case if the two sources were non-complementary, or redundant [Massaro, 1998, pages 424–427].

The final reason is that perceivers combine or integrate the auditory and visual sources of information in an optimally efficient manner. It might seem obvious, given the joint influence of audible and visible speech, that these two sources of information are combined or integrated. Integration is not the only process that can account for an advantage of two sources of information relative to just one, however. There are many possible ways to treat two sources of information: use only the most informative source; use only the auditory source if it is identified and if it isn't use the visible source; average the two sources together; or integrate them in such a fashion in which both sources are used but that the least ambiguous source has the most influence. Research has shown that perceivers do in fact integrate the information available from each modality to perform as efficiently as possible. We now describe a model that predicts this optimally efficient process of combination [Massaro, 1998].

3.2 Fuzzy Logical Model of Perception

The fuzzy logical model of perception (FLMP), shown in Figure 9.5, assumes necessarily successive but overlapping stages of processing. The perceiver of speech is viewed as having multiple sources of information supporting the identification and interpretation of the language input. The model assumes that 1) each source of information is evaluated to give the continuous degree to which that source supports various alternatives, 2) the sources of information are evaluated independently of one another, 3) 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 [Massaro et al., 2001; Massaro, 2002].

The paradigm that we have developed permits us to determine how visible speech is processed and integrated with other sources of information. The results also inform us about which of the many potentially functional cues are actually used by human observers [Massaro, 1987, Chapter 1]. The systematic variation of properties of the speech signal combined with the quantitative test

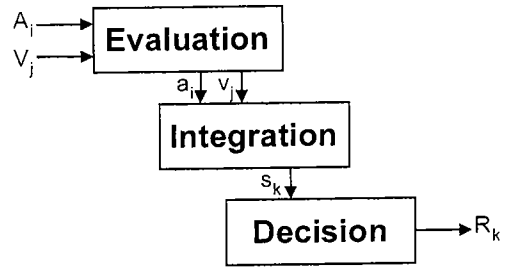

Figure 9.5. Schematic representation of the three processes involved in perceptual recognition. The three processes are shown to proceed left to right in time to illustrate their necessarily successive but overlapping processing. These processes make use of prototypes stored in long-term memory. The sources of information are represented by uppercase letters. Auditory information is represented by A_i and visual information by V_j. The evaluation process transforms these sources of information into psychological 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.

of models of speech perception enables the investigator to test the psychological validity of different cues. This paradigm has already proven to be effective in the study of audible, visible, and bimodal speech perception [Massaro, 1987; Massaro, 1998]. Thus, our research strategy not only addresses how different sources of information are evaluated and integrated, but can uncover what sources of information are actually used. We believe that the research paradigm confronts both the important psychophysical question of the nature of information and the process question of how the information is transformed and mapped into behaviour. Many independent tests point to the viability of the FLMP as a general description of pattern recognition. The FLMP is centred around a universal law of how people integrate multiple sources of information. This law and its relationship to other laws is developed in detail in [Massaro, 1998]. The FLMP is also valuable because it motivates our approach to language learning.

Baldi can display a midsagital view, or the skin on the face can be made transparent to reveal the internal articulators. The orientation of the face can be changed to display different viewpoints while speaking, such as a side view, or a view from the back of the head [Massaro, 1999; Massaro, 2000]. The auditory and visual speech can also be independently controlled and manipulated, permitting customized enhancements of the informative characteristics of speech. These features offer novel approaches in language training, permitting one to pedagogically illustrate appropriate articulations that are usually

hidden by the face. This technology has the potential to help individuals with language delays and deficits, and we have been utilizing Baldi to carry out language tutoring with hard of hearing children and children with autism.

4. Language Learning

As with most issues in social science, there is no consensus on the best way to teach or to learn language. There is agreement, however, about the importance of time on task; learning and retention are positively correlated with the time spent learning. Our technology offers a platform for unlimited instruction, which can be initiated when and wherever the child and/or supervisor chooses. Baldi and the accompanying lessons are perpetual. Take, for example, children with autism, who have irregular sleep patterns. A child could conceivably wake in the middle of the night and participate in language learning with Baldi as his or her friendly guide.

Several advantages of utilizing a computer-animated agent as a language tutor are clear, including the popularity of computers and embodied conversational agents with children with autism. A second advantage is the availability of the program. Instruction is always available to the child, 24 hours a day 365 days a year. Furthermore, instruction occurs in a one-on-one learning environment for the students. We have found that the students enjoy working with Baldi because he offers extreme patience, he doesn't become angry, tired, or bored, and he is in effect a perpetual teaching machine.

4.1 Vocabulary Learning

Vocabulary knowledge is critically important for understanding the world and for language competence in both spoken language and in reading. There is empirical evidence that very young children more easily form conceptual categories when category labels are available than when they are not [Waxman, 2002]. There is also evidence that there is a sudden increase in the rate at which new words are learned once the child knows about 150 words. Grammatical skill also emerges at this time. Even children experiencing language delays because of specific language impairment benefit once this level of word knowledge is obtained. It follows that increasing the pervasiveness and effectiveness of vocabulary learning offers a huge opportunity for improving conceptual knowledge and language competence for all individuals, whether or not they are disadvantaged because of sensory limitations, learning disabilities, or social condition. Finally, it is well-known that vocabulary knowledge is positively correlated with both listening and reading comprehension [Anderson and Freebody, 1981].

Our Language Tutor, Baldi, encompasses and instantiates the developments in the pedagogy of how language is learned, remembered and used. Education research has shown that children can be taught new word meanings by using direct instruction, e.g., [McKeown et al., 1985; Stahl, 1986]. It has also been convincing demonstrated that direct teaching of vocabulary by computer software is possible, and that an interactive multimedia environment is ideally suited for this learning [Berninger and Richards, 2002; Wood, 2001]. As cogently observed by Wood [2001], "Products that emphasize multimodal learning, often by combining many of the features discussed above, perhaps make the greatest contribution to dynamic vocabulary learning. Multimodal features not only help keep children actively engaged in their own learning, but also accommodate a range of learning styles by offering several entry points: When children can see new words in context, hear them pronounced, type them into a journal, and cut and paste an accompanying illustration (or create their own), the potential for learning can be dramatically increased." Following this logic, many aspects of our lessons enhance and reinforce learning. For example, the existing program and planned modifications make it possible for the student to 1) Observe the words being spoken by a realistic talking interlocutor (Baldi), 2) See the word as written as well as spoken, 3) See visual images of referents of the words or view an animation of a meaningful scene, 4) Click on or point to the referent, 5) Hear himself or herself say the word, 6) Spell the word by typing, 7) Observe the word used in context, and 8) Incorporate the word into his or her own speech act.

Other benefits of our program include the ability to seamlessly meld spoken and written language, provide a semblance of a game-playing experience while actually learning, and to lead the child along a growth path that always bridges his or her current "zone of proximal development."

4.2 Description of Language Wizard and Tutor

The Language Tutor and Wizard is a user-friendly application that allows the presentation and composition of lessons with minimal computer experience [Bosseler and Massaro, 2003; Barker, 2003].[2] The lessons encompass and instantiate the developments in the pedagogy of how language is learned, remembered and used.

Figure 9.6 shows a view of the screen from the Presentation exercise in a prototypical lesson. In this lesson, the students learn to identify vegetables. In this exercise, text corresponding to each item is presented when the item is tutored. An outlined region around the zucchini designates the selected object. Emoticon "stickers" (not shown) can also be used as feedback for the responses.

[2]The development of this application was carried out in collaboration with the Center for Spoken Language Understanding at the Oregon Health Sciences University and the Tucker Maxon Oral School, both in Portland, Oregon.

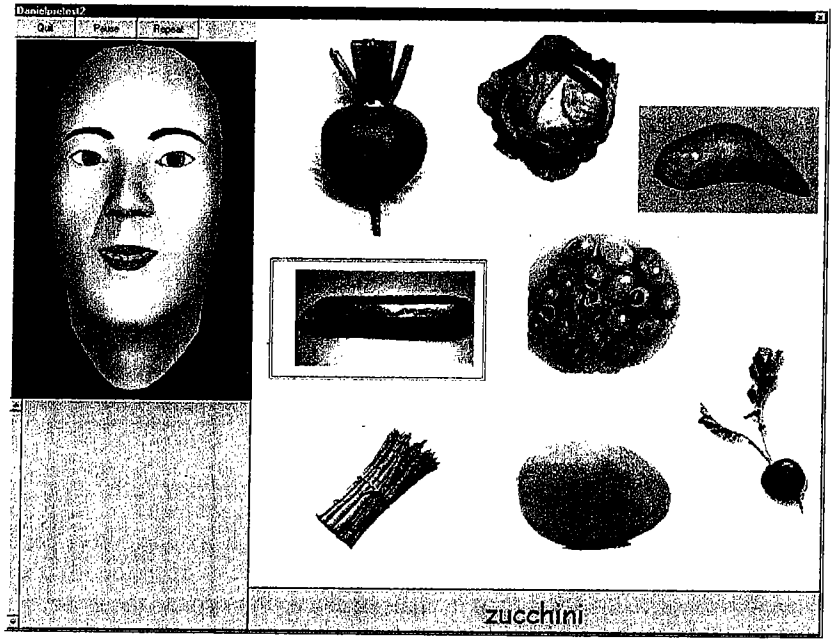

Figure 9.6. A prototypical lesson illustrating the format of the Language Tutor. Each lesson contains Baldi, the vocabulary items and written text and captioning (optional, not shown), and emoticons (not shown). In this application the students learn to identify vegetables. For example, Baldi says "this is a zucchini" in the Presentation exercise.

All of the exercises required the children to respond to spoken directives such as "click on the little chair", or "find the red fox". These images were associated with the corresponding spoken vocabulary words, see [Bosseler and Massaro, 2003] for vocabulary examples). The items became highlighted whenever the mouse passed over that region. The student selected his or her response by clicking the mouse on one of the designated areas.

The Language Wizard consists of 8 different exercises. These exercises are pre-test, presentation, recognition, reading, spelling, imitation, elicitation, and post-test. The Wizard is equipped with easily changeable default settings that determine what Baldi says and how he says it, the oral feedback and emoticons given for responses, the number of attempts permitted for the student in each exercise, and the number of times each item is presented. The program automatically creates and writes all student performance information to a log file stored in the student's directory.

5. Research on the Educational Impact of Animated Tutors

Research has shown that this pedagogical and technological program is highly effective for both children with hearing loss and children with autism. Processing information presented via the visual modality reinforces learning [Courchesne et al., 1994] and is consistent with the TEEACH [Schopler et al., 1995] suggestion for the use of visually presented material. These children tend to have major difficulties in acquiring language, and they serve as particularly challenging tests for the effectiveness of our pedagogy. We now describe some recent research carried out to evaluate our animated tutor to teach both children with hearing loss [Barker, 2003; Massaro et al., 2003] and children with autism [Bosseler and Massaro, 2003].

5.1 Improving the Vocabulary of Hard of Hearing Children

It is well-known that hard of hearing children have significant deficits in vocabulary knowledge. In many cases, the children do not have names for specific things and concepts. These children often communicate with phrases such as "the window in the front of the car," "the big shelf where the sink is," or "the step by the street" rather than "windshield," "counter," or "curb" [Barker, 2003, citing Pat Stone]. The Language Tutor has been in use at the Tucker Maxon Oral School in Portland, Oregon, and Barker [2003] evaluated its effectiveness. Students were given cameras to photograph objects at home and surroundings. The pictures of these objects were then incorporated as items in the lessons. A given lesson had between 10 and 15 items. Students worked on the items about 10 minutes a day until they reached 100% on the post-test. They then moved on to another lesson. About one month after each successful (100%) post-test, they were retested on the same items. Ten girls and nine boys the "upper school" and the "lower school" participated in the applications. There were six hard of hearing children and one hearing child between 8 and 10 years of age in the lower school. Ten hard of hearing and two hearing children, between 11 and 14 years of age, participated from the upper school.

Similar results were found for both age groups. Students knew about one-half of the items without any learning, they successfully learned the other half of the items, and retained about one-half of the newly learned items when retested 30 days later. These results demonstrate the effectiveness of the language Tutor for learning and retaining new vocabulary.

The results of the Barker evaluation [Barker, 2003] indicated that hard of hearing children learned a significant number of new words, and retained about half of them a month after training ended. No control groups were used in that evaluation, however, and it is possible the children were learning the words

outside of the Language Tutor environment. Furthermore, the time course of learning with the Language Tutor was not evaluated. It is of interest how quickly words can be learned with the Language Tutor to give some idea of how this learning environment would compare to a real teacher. Finally, both identification and production of the words should be assessed given that only identification was measured previously.

5.2 Testing the Validity of the Vocabulary Tutor

To address these issues, Massaro and Light [2004a] carried out an experiment based on a within student multiple baseline design [Baer et al., 1968] where certain words were continuously being tested while other words were being tested and trained. Although the student's instructors and speech therapists agreed not to teach or use these words during our investigation, it is still possible that the words could be learned outside of the Language Tutor environment. The single student multiple baseline design monitors this possibility by providing a continuous measure of the knowledge of words that are not being trained. Thus, any significant differences in performance on the trained words and untrained words can be attributed to the Language Tutor training program itself rather than some other factor.

Eight hard of hearing children, 2 male ages 6 and 7, 6 female ages 9 and 10, were recruited from The Jackson Hearing Center in Los Altos, California and were given parental consent to participate. The male students were in grade 1 and the female students in grade 4 respectively and all students needed help with their vocabulary building skills as suggested by their regular day teachers. One child had a cochlear implant and the seven other children had hearing aids in both ears except for one child with an aid in just a single ear. Using the Language Wizard, the experimenter developed a set of lessons with a collection of vocabulary items that was individually tailored for each student. Each collection of items was comprised of 24 items, broken down into 3 categories of 8 items each. Three lessons with 8 items each were made for each child.

Images of the vocabulary items were presented on the screen next to Baldi as he spoke, as illustrated in Figure 9.6. Some of the exercises required the child to respond to Baldi's instructions such as "click on the cabbage", or "show me the yam", by clicking on the highlighted area or by moving the computer mouse over the appropriate image until an item was highlighted and then clicking on it. Two other exercises asked the child to recognize the written word and to type the word, respectively. The production exercises asked the child to repeat after Baldi once he named the highlighted image or to name the highlighted image on their own, followed by Baldi's naming of the image.

Figure 9.7 gives the results of identification and production for one of the eight students. The results were highly consistent across the eight students.

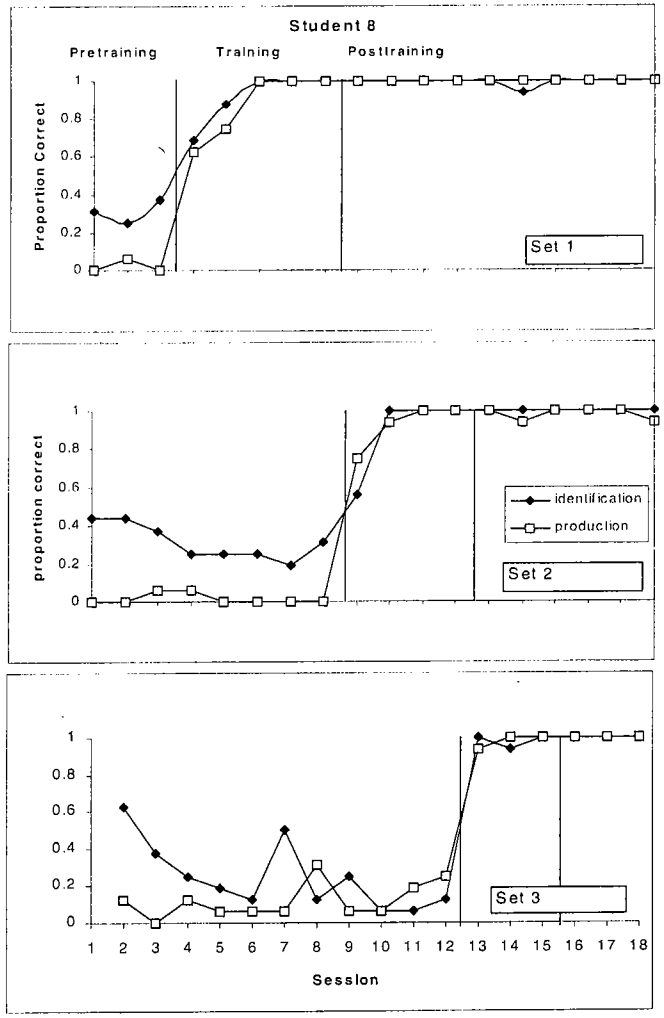

Figure 9.7. Proportion of correctly identified (solid black diamonds) and correctly produced (empty white squares) items across the testing sessions for student 1. The training on a set of words occurred between the two vertical bars. The figure illustrates that once training was implemented identification performance increased dramatically, and remained accurate without further training (from [Massaro and Light, 2004a]).

As expected, identification accuracy (mean = .72) was always higher than production accuracy (mean = .64). This result is not unexpected because a stu-

dent could know the name of an item without being able to pronounce it correctly. There was little knowledge of the test items shown without training, even though these items were repeatedly tested for many days. Once training began on a set of items, however, performance improved fairly quickly until asymptotic knowledge was obtained. This knowledge did not degrade after training on these words ended and training on other words took place. In addition, a reassessment test given about 4 weeks after completion of the experiment revealed that the students retained the items that were learned.

The average number of trials required to reach criterion was 5, 4.3, and 3.4 for mastering the first, second, and third sets of categories. Given that the word lists were randomized across participants, differences in the difficulty of the word sets was probably not responsible for this difference. Learning vocabulary actually involves three different things to be learned. Stimulus learning involves recognizing the stimulus, response leaning requires acquiring the appropriate response, and stimulus-response learning requires an association of the stimulus and the response. The testing of the items actually gave the students experience that could contribute to both stimulus learning and response learning. Thus when the items were finally trained, the students only had to master the stimulus-response association. This would give an advantage of learning the second and third sets of words relative to the first and second, once training was initiated.

5.3 Improving the Vocabulary of Autistic Children

Autism is a spectrum disorder characterized by a variety of characteristics, which usually include perceptual, cognitive, and social differences. Among the defining characteristics of autism, the limited ability to produce and comprehend spoken language is the most common factor leading to diagnosis [American Psychiatric Association, 1994]. The language and communicative deficits extend across a broad range of expression [Tager-Flusberg, 1999]. Individual variations occur in the degree to which these children develop the fundamental lexical, semantic, syntactic, phonological, and pragmatic components of language including those who fail to develop one or more of these elements of language comprehension and production.

Approximately one-half of the autistic population fails to develop any form of functional language [Tager-Flusberg, 2000]. Within the population that does develop language, the onset and rate at which the children pass through linguistic milestones are often delayed compared to non-autistic children (e.g. no single words by age 2 years, no communicative phrases by age 3) [American Psychiatric Association, 1994]. The ability to label objects is often severely delayed in this population as well as the deviant use and knowledge of verbs and adjectives. Van Lancker et al. [1991] investigated the abilities of autistic

and schizophrenic children to identify concrete nouns, non-emotional adjectives, and emotional adjectives. The results showed that the performance of children with autism was below controls in all three areas.

Despite the prevalence of language delays in autistic individuals, formalized research has been limited, partly due to the social challenges inherent in this population [Tager-Flusberg, 2000]. Intervention programs for children with autism typically emphasize developing speech and communication skills (e.g. TEEACH, Applied Behavioural Analysis). These programs most often focus on the fundamental lexical, semantic, syntactic, phonological, and pragmatic components of language. The behavioural difficulties speech therapists and instructors encounter, such as lack of cooperation, aggression, and lack of motivation to communicate, create difficult situations that are not optimal for learning. Thus, creating motivational environments necessary to develop these language skills introduces many inherent obstacles [Tager-Flusberg, 2000].

In this study [Bosseler and Massaro, 2003], the Tutors were constructed and run on a 600 MHz PC with 128 MB RAM hard drive running Microsoft Windows NT 4 with a Gforce 256 AGP-V6800 DDR graphics board. The tutorials were presented on a Graphic Series view Sonic 20" monitor. All students wore a Plantronics PC Headset model SR1. Students completed 2 sessions a week, a minimum of 2 lessons per session, and an average of 3, and sometimes as many as 8. The sessions lasted between 10 and 40 minutes. A total of 559 different vocabulary items were selected from the curriculum of both schools for a total of over 84 unique vocabulary lessons.

A series of observations by the experimenter during the course of each lesson led to many changes in the program, including the use of headsets, isolating the student from the rest of the class and removal of negative verbal feedback from Baldi (such as, "No, (student) that's not right". The students appeared to enjoy working with Baldi. We documented the children saying such things as "Hi Baldi" and "I love you Baldi". The stickers generated for correct (happy face) and incorrect (sad face) responses proved to be an effective way to provide feedback for the children, although some students displayed frustration when he or she received more than one sad face. The children responded to the happy faces by saying such things like "Look, I got them all right", or laughing when a happy face appeared. We also observed the students providing verbal praise to themselves such as "Good job", or prompting the experimenter to say "Good job" after every response. For the autistic children, several hundred vocabulary tutors were constructed, consisting of various vocabulary items selected from the curriculum of two schools. The children were administered the tutorial lessons until 100% accuracy was attained on the post-test exercise. Once 100% accuracy was attained on the final post-test module, the child did not see these lessons again until reassessment approximately 30 days later.

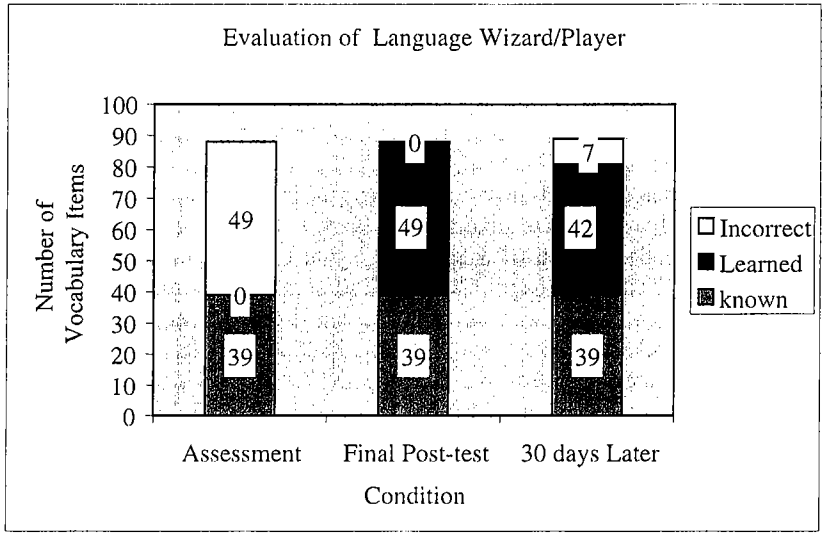

Figure 9.8. The mean observed proportion of correct identifications for the initial assessment, final posttest and reassessment for each of the seven students. reassessment. The results reveal these seven students were able to accurately identify significantly more words during the reassessment than the initial assessment (from [Bosseler and Massaro, 2003]).

Figure 9.8 shows that the children learned many new words, grammatical constructions, and concepts, proving that the language tutors are a valuable learning environment for these children. In order to assess how well the children would retain the vocabulary items that were learned during the tutorial lesson, we administered the assessment test to the student at least 30 days following the final post-test. As can be seen in Figure 9.8, the students were able to recall 85% of the newly-learned vocabulary items at least 30 days following training.

5.4 Validity and Generalization

Although all of the children demonstrated learning from initial assessment to final reassessment, the children might have been learning the words outside of our program, for example, from speech therapists, at home, or in their school curriculum. Furthermore, we questioned whether the vocabulary knowledge would generalize to new pictorial instances of the words. To address these issues we conducted a second experiment. Collaborating with the children's instructors and speech therapists, we gathered an assortment of vocabulary words that the children supposedly did not know. We used these words in the Horner

and Baer [1978] single subject multiple probe design. We randomly separated the words to be trained into three sets, established individual pre-training performance for each set of vocabulary items, and trained on the first set of words while probing performance for both the trained and untrained sets of words.

Once the student was able to attain 100% identification accuracy during a training session, a generalization probe to new instances of the vocabulary images was initiated. If the child did not meet the criterion, he or she was trained on these new images. Generalization training continued until the criterion was met, at which time training began on the next set of words. Probe tests continued on the original learned set of words and images until the end of the study. We continued this procedure until the student completed training on all three sets of words. Our goal was to observe a significant increase in identification accuracy during the post-training sessions relative to the pre-training sessions.

Figure 9.9 displays the proportion of correct responses for a typical student during the probe sessions conducted at pre-training and post-training for each of the three word sets. The vertical lines in each of the three panels indicates the last pre-training session before the onset of training. Some of the words were clearly known prior to training, and were even learned to some degree without training. As can be seen in the figure, however, training was necessary for substantial learning to occur. In addition, the children were able to generalize accurate identification to four instances of untrained images.

In summary, the goal of these investigations was to evaluate the potential of using a computer-animated talking tutor for children with language delays. The results showed a significant gain in vocabulary. We also found that the students were able to recall much of the new vocabulary when reassessed 30 days after learning. Follow-up research showed that the learning is indeed occurring from the computer program and vocabulary knowledge can transfer to novel images.

It should be emphasized that the present studies used a within-subject design (with fewer participants) than the field's general acceptance of between-subject designs. There are several limitations with between-subject designs, however. First, different groups might differ on a pretest so that differences in the pretest-posttest differences might not be a valid dependent measure [Loftus, 1978]. Second, even if subjects do not differ in the pretest, they might differ in their learning ability. Thus group differences might reflect other differences rather than the learning conditions. These potential limitations are less problematic in the within-subject design used in the current research. Furthermore, the multiple baseline design enforces a within-subject comparison in which the effectiveness of the independent variable can be evaluated directly. In our case, a child was continuously tested on items that had not yet been learned, were currently being learned, or had already been learned. The direct comparisons

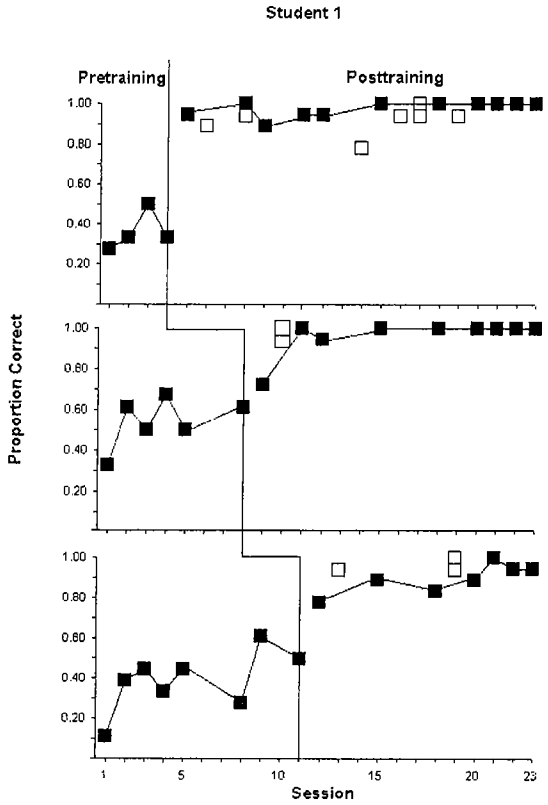

Figure 9.9. Proportion correct during the Pretraining, Posttraining, and Generalization for one of the six students. The vertical lines separate the Pretraining and Posttraining conditions. Generalization results are given by the open squares (from [Bosseler and Massaro, 2003]).

among these conditions showed conclusively that the training program was responsible for the learning.

An obvious question is how effective our training program is relative to a live teacher. Given the fairly fast rate of learning for both the hard of hearing and autistic children, the potential advantage of a live teacher cannot be very large. Even if learning with the live teacher is significantly faster, our program is still valuable because it is low cost and always available,

5.5 Value of the Face for Autistic Children

We believe that the children in our investigation profited from having the face and that seeing and hearing spoken language can better guide language learning than either modality alone. A direct test of this hypothesis would involve comparing learning with and without the face. The purpose of this investigation was to evaluate whether pairing an animated tutor, Baldi, with audible speech in vocabulary training facilitates learning and retention relative to presenting the audible speech alone [Massaro and Bosseler, 2003]. If children with autism do not extract meaningful information from the face, then we would expect to see no difference in learning between the two conditions. This evaluation was carried out using two within-subject experimental conditions: training with the face and the voice and training with the voice only. We evaluated whether the face would increase the rate of learning for both receptive and verbal production measures. To accomplish our goal we compared the two conditions according to an alternating treatment design [Baer et al., 1968; Horner and Baer, 1978] in which each student received each of the two learning conditions concurrently, the order of presentation of the two conditions counterbalanced across days. This alternating treatment design permitted us to assess the individual performance of word identification and production, eliminating inter-subject variability and permitted a direct observation of the two treatment conditions [Baer et al., 1968; Horner and Baer, 1978]. To determine the effect of training on retention, an additional testing block was carried once training was terminated.

Figure 9.10 gives the identification results for one of the five students for the pre-training, training Post-test, and post-training blocks. The left and right vertical lines in figure separate these three conditions. As can be seen in the figure, this student showed a very large advantage of having the face during learning. Across all 5 students, there was faster learning and more retention with the face than without the face.

In summary, the studies using the Language Wizard/Tutor show that children with language challenges were able to learn a significant amount of new vocabulary. By implementing experimental controls in our evaluation, we were able to conclude that the Language Wizard/Tutor was responsible for this learning. Finally, by systematically varying whether Baldi's face was present during the language tutoring, we learned that the presence of the face contributed significantly to the language learning.

5.6 Training Speech Production

Baldi can actually provide more information than a natural face. Baldi has a tongue, hard palate and three-dimensional teeth and his internal articulatory movements have been trained with electropalatography and ultrasound data

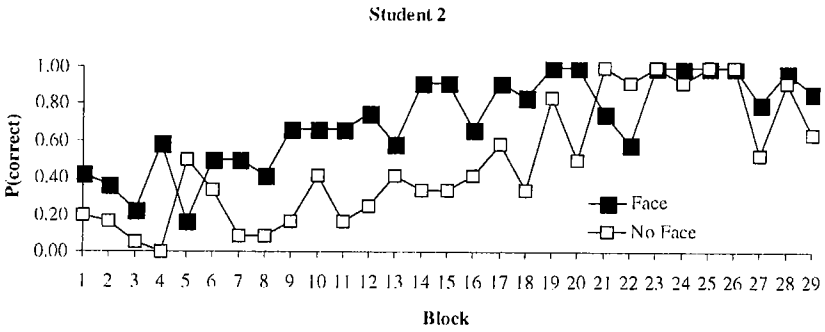

Figure 9.10. Mean proportion correct receptive responses for one of the 5 students during pre-training (first 3 blocks), training, and post-training (last 3 blocks) as a function of training block for the face and no face conditions.

from natural speech [Cohen et al., 1998]. Baldi can be programmed to display a midsagital view, or the skin on the face can be made transparent to reveal the internal articulators. The orientation of the face can be changed to display different viewpoints while speaking, such as a side view, or a view from the back of the head [Massaro, 1999]. The auditory and visual speech can also be independently controlled and manipulated, permitting customized enhancements of the informative characteristics of speech. These features offer novel approaches in language training, permitting one to pedagogically illustrate appropriate articulations that are usually hidden by the face. More generally, additional research should investigate whether the influence of several modalities on language processing provide a productive approach to language learning.

Children with hearing loss require guided instruction in speech perception and production. 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 use visible speech when providing our stimuli. Based on reading research [Torgesen et al., 1999], we expected that visible cues would allow for heightened awareness of the articulation of these segments and assist in the training process.

Although 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 tract is visible, or we can present a cutaway view of the head along the sagittal plane. As an example, a unique view of Baldi's internal articulators can be presented by rotating the exposed head and vocal tract to be oriented away from the student. It is possible that this back-of-head view would be much more conducive to learning language production. The tongue in this view moves away from and towards the student in the same way as the student's own tongue would move. This correspondence between views

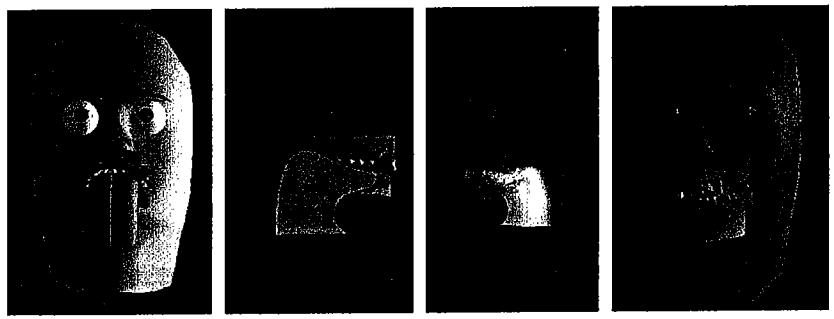

Figure 9.11. The four presentation conditions of Baldi with transparent skin revealing inside articulators (back view, sagittal view, side view, front view).

of the target and the student's articulators might facilitate speech production learning. One analogy is the way one might use a map. We often orient the map in the direction we are headed to make it easier to follow (e.g. turning right on the map is equivalent to turning right in reality).

Another characteristic of the training is to provide additional cues for visible speech perception. Baldi can illustrate the articulatory movements, and he can be made even more informative by embellishing of the visible speech with added features. Several alternatives are obvious for distinguishing phonemes that have similar visible articulations, such as the difference between voiced and voiceless segments. For instance, showing visual indications of vocal cord vibration and turbulent airflow can be used to increase awareness about voiced versus voiceless distinctions. These embellished speech cues could make the face more informative than it normally is.

5.6.1 Empirical test of speech production training. In the Massaro and Light [2004b] study, hard of hearing students were trained to discriminate minimal pairs of words bimodally (auditorily and visually), and were also trained to produce various speech segments by visual information about how the inside oral articulators work during speech production. As shown in Figure 9.11, the articulators were displayed from different vantage points so that the subtleties of articulation could be optimally visualized. The speech was also slowed down significantly to emphasize and elongate the target phonemes, allowing for clearer understanding of how the target segment is produced in isolation or with other segments.

During production training, different illustrations were used to train different distinctions. Although any given speech sound can be produced in a variety of ways, a prototypical production was always used. Supplementary visual indications of vocal cord vibration and turbulent airflow were used to distinguish

the voiced from the voiceless cognates. The major differences in production of these sounds are the amount of turbulent airflow and vocal cord vibration that takes place (e.g. voiced segments: vocal cord vibration with minimal turbulent airflow; voiceless segments: no vocal cord vibration with significant turbulent airflow). Although the internal views of the oral cavity were similar for these cognate pairs, they differed on the supplementary voicing features. For consonant clusters, we presented a view of the internal articulators during the production to illustrate the transition from one articulatory position to the next. Finally, both the visible internal articulation and supplementary voicing features were informative for fricative versus affricate training. An affricate is a stop followed by a homorganic fricative. The time course of articulation and the how the air escaped the mouth (e.g. fricative: slow, consistent turbulent airflow, affricate: quick, abrupt turbulent airflow) differed.

The production of speech segments was trained in both isolated segments and word contexts. Successful perceptual learning has been reported to depend on the presence of stimulus variability in the training materials [Kirk et al., 1997]. In the present study, we varied the trained speech segments on various dimensions such as segment environment (beginning/end of word) and neighbouring vowel quality (height and front/backness features), and neighbouring consonant quality (place and manner features) in the case of consonant cluster training, to optimize learning. Ideally, training of a target segment would generalize to any word, trained or untrained. In an attempt to assess whether or not the learning of specific segments was restricted to the words involved in our training, we included both trained and untrained words in our pre-test and post-test measures. This contrast allowed us to test whether the training generalized to new words. A follow up measure allowed us to evaluate retention of training six weeks after post-test. We expected that performance would be greater than pre-test but not as high as post-test levels due to discontinued use of training.

The main goal of this study was to implement Baldi as a language tutor for speech perception and production for hard of hearing individuals. The student's ability to perceive and produce words involving the trained segments improved from pre-test to post-test. A second analysis revealed an improvement in production no matter which training method was used (e.g. vocal cord vibration and turbulent airflow vs. slowed down speech with multiple internal articulatory views vs. a combination of both methods).

The present findings suggest that Baldi is an effective tutor for speech training hard of hearing students. There are other advantages of Baldi that were not exploited in the present study. Baldi can be accessed at any time, used as frequently as wished and modified to suit individual needs. Baldi also proved beneficial even though students in this study were continually receiving speech training with their regular and speech teachers before, during and after this

Figure 9.12. Supplementary features indicating from left to right, vocal cord vibration, frication as in /s/, and nasal as in /n/ (the red nasal opening cannot be seen in the black and white illustration).

study took place. Baldi appears to offer unique features that can be added to the arsenal of Speech-language pathologists.

The post-test productions were significantly better than the pre-test productions, indicating significant learning. Given that it is always possible that some of this learning occurred independently of our program or was simply based on routine practice, there is some evidence that at least some of the improvement must be due to our program. Follow-up ratings six weeks after our training was complete were significantly lower than post-test ratings, indicating some decrement due to lack of continued use. From these results we can conclude that our training program was a significant contributing factor to the change in ratings seen for production ability. Future studies can now focus on which specific training regimens for which contrasts are most effective.

5.6.2 Learning speech in a new language. A recent study by Massaro and Light [2003] demonstrated the effectiveness of Baldi for teaching non-native phonetic contrasts, by including instruction illustrating the internal articulatory processes of the oral cavity, as in the previous study with children with hearing loss. Eleven Japanese adult speakers of English as a second language were bimodally to identify and produce American English /r/ and /l/. The adults learned to produce these segments more accurately, indicating that Baldi holds great promise for second language learning.

5.7 Learning to Read

We now turn from speech reading to reading: the out-of-the-ordinary problems that a number of children encounter in learning to read and spell. Dyslexia

is a category used to pigeonhole children who have much more difficulty in reading and spelling than would be expected from their other perceptual and cognitive abilities [Fleming, 1984; Willows et al., 1993]. Psychological science has established a tight relationship between the mastery of written language and the child's ability to process spoken language [Morais and Kolinsky, 1994]. That is, it appears that many dyslexic children also have deficits in spoken language perception. The difficulty with spoken language can be alleviated through improving children's perception of phonological distinctions and transitions, which in turn improves their ability to read and spell. Visible speech could only enhance the instruction of phonological awareness [Torgesen et al., 1999] and therefore offers provide another dimension of information for the children to use in identifying segments and mastering phonological awareness.

Baldi can be embellished to signal characteristics of the speech signal that could aid in the teaching of phonological awareness. Figure 9.12 illustrates some potential features that could be displayed along with the typical information given by visible speech. Today, almost all personal computers have the capability to support a bimodal text-to-speech system, which would make it possible to incorporate bimodal speech in reading exercises. This treatment holds great promise, and we believe that adding visual speech will significantly enhance the positive results that have already been demonstrated with audible speech alone.

6. Summary

Speech and language science and technology evolved under the assumption that speech was a solely auditory event. However, a burgeoning record of research findings reveals that our perception and understanding are influenced by a speaker's face and accompanying gestures, as well as the actual sound of the speech. Perceivers expertly use these multiple sources of information to identify and interpret the language input. Given the value of face-to-face interaction, our persistent goal has been to develop, evaluate, and apply animated agents to produce realistic and accurate speech. Baldi is an accurate three-dimensional animated talking head appropriately aligned with either synthesized or natural speech. Baldi has a realistic tongue and palate, which can be displayed by making his skin transparent.

Based on this research and technology, we have implemented computer-assisted speech and language tutors for children with language challenges and persons learning a second language. Our language-training program utilizes Baldi as the conversational agent, who guides students through a variety of exercises designed to teach vocabulary and grammar, to improve speech articulation, and to develop linguistic and phonological awareness. Some of the advantages of the Baldi pedagogy and technology include the popularity and

effectiveness of computers and embodied conversational agents, the perpetual availability of the program, and individualized instruction. The science and technology of Baldi hold great promise in language learning, dialog, human-machine interaction, education, and edutainment.

Acknowledgements

The research and writing of the chapter 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 author is highly grateful for the dedication of the PSL team, particularly Michael Cohen, Alexis Bosseler, Joanna Light, Rashid Clark, and Slim Ouni.

References

American Psychiatric Association (1994). *Diagnostic and Statistical Manual of Mental Disorders Manual of Mental Disorders, DSM-IV*. Washington, DC, 4th edition.

Anderson, R. C. and Freebody, P. (1981). Vocabulary Knowledge. In Guthrie, J. T., editor, *Comprehension and Teaching: Research Perspectives*, pages 71–117. Newark, DE: International Reading Association.

Baer, D. M., Wolf, M. M., and Risley, T. R. (1968). Some Current Dimensions of Applied Behavior Analysis. *Journal of Applied Behavior Analysis*, 1:91–97.

Barker, L. J. (2003). Computer-Assisted Vocabulary Acquisition: The CSLU Vocabulary Tutor in Oral-Deaf Education. *Journal of Deaf Studies and Deaf Education*, 8:187–198.

Berninger, V. W. and Richards, T. L. (2002). *Brain Literacy for Educators and Psychologists Educators and Psychologists*. San Diego: Academic Press.

Bosseler, A. and 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.

Cohen, M. M., Beskow, J., and Massaro, D. W. (1998). Recent Developments in Facial Animation: An Inside View. In Burnham, D., Robert-Ribes, J., and Vatikiotis-Bateson, E., editors, *Proceedings of Auditory-Visual Speech Processing (AVSP)*, pages 201–206, Sydney, Australia.

Cohen, M. M., Clark, R., and Massaro, D. W. (2001). Animated Speech: Research Progress and Applications. In Massaro, D. W., Light, J., and Geraci, K., editors, *Proceedings of Auditory-Visual Speech Processing (AVSP)*, page 201, Aalborg, Denmark. Santa Cruz, CA: Perceptual Science Laboratory.

Cohen, M. M. and Massaro, D. W. (1993). Modeling Coarticulation in Synthetic Visual Speech. In Thalmann, M. and Thalmann, D., editors, *Computer Animation '93*, pages 139–156. Tokyo: Springer Verlag.

Cohen, M. M., Massaro, D. W., and Clark, R. (2002). Training a Talking Head. In *Proceedings of Fourth International Conference on Multimodal Interfaces (ICMI)*, pages 499–510, Pittsburgh, Pennsylvannia, USA.

Cohen, M. M., Walker, R. L., and Massaro, D. W. (1996). Perception of Synthetic Visual Speech. In Stork, D. G. and Hennecke, M. E., editors, *Speechreading by Humans and Machines*, pages 153–168. New York: Springer.

Cosi, P., Cohen, M. M., and Massaro, D. W. (2002). Baldini: Baldi Speaks Italian. In *Proceedings of the Seventh International Conference on Spoken Language Processing (ICSLP)*, pages 2349–2352, Denver Colorado.

Courchesne, E., Townsend, J., Ashoomoff, N. A., Yeung-Courchesne, R., Press, G., Murakami, J., Lincoln, A., James, H., Saitoh, O., Haas, R., and Schreibman, L. (1994). A New Finding in Autism: Impairment in Shifting Attention. In Broman, S. H. and Grafman, J., editors, *Atypical Cognitive Deficits in Developmental Disorders: Implications for Brain Function*, pages 101–137. Hillsdale, NJ: Lawrence Erlbaum.

Denes, P .B. and Pinson, E. N. (1963). The Speech Chain. In *The Physics and Biology of Spoken Language*. New York: Bell Telephone Laboratories.

Fleming, E. (1984). *Believe the Heart*. San Francisco: Strawberry Hill Press.

Guenter, B., Grimm, C., Wood, D., Malvar, H., and Pighin, F. (1998). Making Faces. In *Proceedings of the 25th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*, pages 55–66. NewYork, NY: ACM Press.

Horner, R. D. and Baer, D. M. (1978). Multiple-Probe Technique: A Variation of the Multiple Baseline. *Journal of Applied Behavior Analysis*, 11:189–196.

Jesse, A., Vrignaud, N., and Massaro, D. W. (2000/2001). The Processing of Information from Multiple Sources in Simultaneous Interpreting. *Interpreting*, 5:95–115.

Kent, J. R., Carlson, W. E., and Parent, R. E. (1992). Shape Transformation for Polyhedral Objects. In *Proceedings of ACM SIGGRAPH Computer Graphics*, volume 26:2, pages 47–54. New York, NY: ACM Press.

Kirk, K. I., Pisoni, D. B., and Miyamoto, R. C. (1997). Effects of Stimulus Variability on Speech Perception in Listeners with Hearing Impairment. *Journal of Speech & Hearing Research*, 40:1395–1405.

Loftus, G. R. (1978). On Interpretation of Interactions. *Memory & Cognition*, 6:312–319.

Massaro, D. W. (1987). *Speech Perception by Ear and Eye: A Paradigm for Psychological Inquiry*. Hillsdale, NJ: Lawrence Erlbaum.

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

Massaro, D. W. (1999). From Theory to Practice: Rewards and Challenges. In *Proceedings of the International Conference of Phonetic Sciences*, pages 1289–1292, San Francisco, CA.

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)*, pages 153–161, Dundee, Scotland.

Massaro, D. W. (2002). Multimodal Speech Perception: A Paradigm for Speech Science. In Granström, B., House, D., and Karlsson, I., editors, *Multimodality in Language and Speech Systems*, pages 45–71. Kluwer Academic Publishers, Dordrecht, The Netherlands.

Massaro, D. W. and Bosseler, A. (2003). Perceiving Speech by Ear and Eye: Multimodal Integration by Children with Autism. *The Journal of Developmental and Learning Disorders*, 7:111–146.

Massaro, D. W., Bosseler, A., and Light, J. (2003). Development and Evaluation of a Computer-Animated Tutor for Language and Vocabulary Learning. In *Proceedings of the 15th International Congress of Phonetic Sciences*, Barcelona, Spain.

Massaro, D. W., Cohen, M. M., Campbell, C. S., and Rodriguez, T. (2001). Bayes Factor of Model Selection Validates FLMP. *Psychonomic Bulletin & Review*, 8:1–17.

Massaro, D. W., Cohen, M. M., Tabain, M., Beskow, J., and Clark, R. (2005). Animated Speech: Research Progress and Applications. In Vatiokis-Bateson, E., Bailly, G., and Perrier, P., editors, *Audiovisual Speech Processing*. Cambridge: MIT Press. In press.

Massaro, D. W. and Light, J. (2004a). Improving the Vocabulary of Children with Hearing Loss. *Volta Review*, 104(3):141–174.

Massaro, D. W. and Light, J. (2004b). Using Visible Speech for Training Perception and Production of Speech for Hard of Hearing Individuals. *Journal of Speech, Language, and Hearing Research*, 47(2):304–320.

McKeown, M., Beck, I., Omanson, R., and Pople, M. (1985). Some Effects of the Nature and Frequency of Vocabulary Instruction on the Knowledge and Use of Words. *Reading Research Quarterly*, 20:522–535.

Morais, J. and Kolinsky, R. (1994). Perception and Awareness in Phonological Processing: The Case of the Phoneme. *Cognition*, 50:287–297.

Ouni, S., Massaro, D. W., Cohen, M. M, Young, K., and Jesse, A. (2003). Internationalization of a Talking Head. In *Proceedings of the 15th International Congress of Phonetic Sciences (ICPhS)*, Barcelona, Spain.

Parke, F. I. (1975). A Model for Human Faces that allows Speech Synchronized Animation. *Computers and Graphics Journal*, 1:1–4.

Schopler, E., Mezibov, G. B., and Hearsey, K. (1995). Structured Teaching in the TEACCH System. In Schopler, E., Mesibov, G. B., and Hearsey, K., editors, *Learning and Cognition in Autism. Current Issues in Autism*, pages 243–268. New York: Plenum Press.

Stahl, S. A. (1986). Three Principals of Effective Vocabulary Instruction. *Journal of Reading*, 29:662–668.

Tager-Flusberg, H. (1999). A Psychological Approach to Understanding the Social and Language Impairments in Autism. *International Review of Psychiatry*, 11:355–334.

Tager-Flusberg, H. (2000). Language Development in Children with Autism. In Menn, L. and Ratner, N. B., editors, *Methods For Studying Language Production*, pages 313–332. New Jersey: Mahwah.

Torgesen, J. K., Wagner, R. K., Rashotte, C. A., Lindamood, P., Conway, E. Rose T., and Garvan, C. (1999). Preventing Reading Failure in Young Children with Phonological Processing Disabilities: Group and Individual Responses to Instruction. *Journal of Educational Psychology*, 91:579–593.

Trychin, S. (1997). Guidelines for Providing Mental Health Services to People who are Hard of Hearing. Technical report, Gallaudet University, Washington D.C.

van Lancker, D., Cornelius, C., and Needleman, R. (1991). Comprehension of Verbal Terms for Emotions in Normal, Autistic, and Schizophrenic Children. *Developmental Neuropsychology*, 7:1–18.

Waxman, S. R. (2002). Early Word-Learning and Conceptual Development: Everything had a Name, and each Name Gave Birth to a New Thought. In Goswami, U., editor, *Handbook of Childhood Cognitive Development*, pages 102–126. Malden, MA: Blackwell Publishing.

Willows, D. M., Kruk, R. S., and Corcos, E., editors (1993). *Visual Processes in Reading and Reading Disabilities*. Hillsdale, NJ: Lawrence Erlbaum.

Wood, J. (2001). Can Software Support Children's Vocabulary Development? *Language Learning & Technology*, 5:166–201.

Chapter 10

EFFECTIVE INTERACTION WITH TALKING ANIMATED AGENTS IN DIALOGUE SYSTEMS

Björn Granström and David House
Centre for Speech Technology (CTT)
Department of Speech, Music and Hearing, KTH, Stockholm, Sweden
{bjorn, davidh}@speech.kth.se

Abstract At the Centre for Speech Technology at KTH, we have for the past several years been developing spoken dialogue applications that include animated talking agents. Our motivation for moving into audiovisual output is to investigate the advantages of multimodality in human-system communication. While the mainstream character animation area has focussed on the naturalness and realism of the animated agents, our primary concern has been the possible increase of intelligibility and efficiency of interaction resulting from the addition of a talking face. In our first dialogue system, Waxholm, the agent used the deictic function of indicating specific information on the screen by eye gaze. In another project, Synface, we were specifically concerned with the advantages in intelligibility that a talking face could provide. In recent studies we have investigated the use of facial gesture cues to convey such dialogue-related functions as feedback and turn-taking as well as prosodic functions such as prominence. Results show that cues such as eyebrow and head movement can independently signal prominence. Current results also indicate that there can be considerable differences in cue strengths among visual cues such as smiling and nodding and that such cues can contribute in an additive manner together with auditory prosody as cues to different dialogue functions. Results from some of these studies are presented in the chapter along with examples of spoken dialogue applications using talking heads.

Keywords: Audio-visual speech synthesis, talking heads, animated agents, spoken dialogue systems, visual prosody.

1. Introduction

As we contribute to advances in spoken dialogue systems and see them being integrated into commercial products, we are witnessing a transformation

J.C.J. van Kuppevelt et al. (eds.), Advances in Natural Multimodal Dialogue Systems, 215–243.