From Speech is Special to Computer Aided Language Learning

Dom Massaro and Ron Cole

Department of Psychology, University of California, Santa Cruz, CA 95064 U.S.A.
CSLR, University of Colorado, Boulder, CO
massaro@fuzzy.ucsc.edu, cole@cslr.colorado.edu

Abstract

After describing the belief that speech is special, empirical and theoretical research is reviewed undermining the tenets of this belief. A new framework is presented as a theoretical framework for language learning. Central to this framework is the natural ease of multimodal perception, particularly the value of visible speech. The value of synthetic talking heads is described along with their potential in language learning. Embedded in a speech toolkit platform, containing speech synthesis, recognition and interactive tools, an immediate goal is to teach vocabulary, speech, and reading.

1. Introduction to Speech is Special

A central issue in speech perception and psycholinguistics is the so-called modularity of speech and language. Noam Chomsky (1980) envisioned language ability as dependent on an independent language organ (or module), analogous to other organs such as our digestive system. This organ follows an independent course of development in the first years of life and allows the child to achieve a language competence that cannot be elucidated in terms of traditional learning theory. This mental organ, responsible for the human language faculty and our language competence, matures and develops with experience, but the mature system does not simply mirror this experience. The language user inherits rule systems of highly specific structure. This innate knowledge allows us to acquire the rules of the language, which cannot be induced from normal language experience because (advocates argue) of the paucity of the language input. The data of language experience are so limited that no process of induction, abstraction, generalization, analogy, or association could account for our observed language competence. Somehow, the universal grammar given by our biological endowment allows the child to learn to use language appropriately without learning many of the formal intricacies of the language. At the same time, however, psychologists are finding that infants are highly influenced by experience (e.g., Saffran, 1999) and linguists are documenting that the child's language input is not as sparse as the nativists had argued (Sampson, 1989).

Although speech does not have an advocate as charismatic and influential as Chomsky, a similar description is given for speech perception. In addition, advocates of the special nature of speech are encouraged by Fodor's influential proposal of the modularity of mind. Our magnificent capabilities result from a set of innate and independent systems, such as vision, hearing, and language (Fodor, 1983). Speech-is-special theorists now assume that a speech module is responsible for speech perception (Liberman & Mattingly, 1989; Mattingly & Studdert-Kennedy, 1991). Given the environmental information, the speech module analyzes this information in terms of possible articulatory sequences of speech segments. The perceiver of speech uses his or her own speech-motor system to achieve speech recognition.

The justification for a speech module is analogous to the one for language more generally. Performance is not easily accounted for in terms of the language input. In speech, it is asserted that the acoustic signal is deficient and that typical pattern recognition schemes could not work. Put another way, it is reasoned that speech exceeds our auditory information-processing capabilities. In terms of the modularity view, our speech perception system is linked with our speech production system--and our speech perception is somehow mediated by our speech production. For theorists in the speech-is-special camp, the objects of speech perception are articulatory events or gestures. These gestures are the primitives that the mechanisms of speech production translate into actual articulatory movements and are also the primitives that the specialized mechanisms of speech perception recover from the signal.

2. Some History of Research

Speech perception wasn't always considered specialized. The turn of the nineteenth century was a heady time for psychologists. Fechner, Donders, Wundt, and their converts had paved the way for an experimental study of mental life. With tools such as a tachistoscope to present visual displays for short measurable intervals, experimenters could gain control over stimuli and derive stimulus-response relationships. Some of the best known work involved reading written words. One of the main findings to surface from this research was the important influence of context on reading. As documented in Edmund B. Huey's (1908) seminal text, our knowledge about spelling, syntax, and meaning facilitates the recognition of the letters on a page of text.

In contrast to the plethora of studies carried out on the written word, apparently only one was done on the spoken word. William Chandler Bagley's dissertation

Massaro, D.W., & Cole, R. (2000). From "Speech is special" to talking heads in language learning. In Integrating Speech Technology in the (Language Learning and Assistive Interface, University of Abertay Dundee, Dundee, Scotland, August 29-30, 153-161.

under Edward Titchener showed influences in speech perception that were analogous to those found in written language (Cole & Rudnicky, 1983). Members of Cornell's psychology department were asked to recognize mutilated words with missing segments. This manipulation is reminiscent of Pillsbury's (1897) studies of the recognition of written words with missing letters. In Bagley's (1900) experiment, the naturally spoken words were recorded and played back on Edison phonograph cylinders. The results demonstrated that the context of the sentence improved recognition (and even perception) of the mutilated words. Word recognition was improved if the word was placed in the middle of a sentence, for example. This intuitive result was published in the leading psychological journal of the time, but was quickly forgotten, and speech more or less fell outside the domain of experimental psychology. Bagley's seminal study was not cited in Woodworth's Experimental Psychology (1938) and a twentieth century survey of psychology in America omitted any reference to speech perception (Hilgard, 1987). It also remained somewhat foreign during the "cognitive revolution," at the end of twentieth century, and only the technical goal of speech recognition by machine delegated speech perception its almost fair share of attention from experimental psychologists and other explorers of the mind.

At the beginning of the twentieth century, the psychological study of speech perception came, not from within psychology, but from an applied problem: a reading machine for blinded veterans returning from World War II. The goal was to design a machine that would read typewritten English and convert the letters into distinct sounds. The nonsighted listener would learn to recognize these sounds and read by ear. The scientists quickly found that the words spoken by machine were very difficult to understand and were not easily learned. This led Alvin Liberman and his colleagues to question why humans recognize natural speech so easily. Their inspiration was that we perceive speech via the same mechanisms used to produce speech: Speech was special. The nonsense sounds emanating from the speaking machine had little to do with how speech was spoken and, therefore, were gibberish to the listener. The next three decades of research from Haskins Laboratory was centered on the theme of the specialized nature of speech perception.

2.1 Categorical Perception and its Demise

The strongest evidence harnessed by Haskins Laboratories was categorical perception (CP), or the perceived equality of instances within a category. The CP of phonemes has been a central concept in the experimental and theoretical investigation of speech perception and has also spilled over into other domains such as face processing [1]. CP was operationalized in terms of discrimination performance being limited by identification performance. Over 40 years ago, researchers at Haskins Laboratories [2] used synthetic speech to generate a series of 14 consonant-vowel syllables going from /be/ to /de/ to /ge/ (/e/ as in gate). The onset frequency of the second formant transition of the initial consonant was changed in equal steps to produce the continuum. In the identification task, observers identified random presentations of the sounds as /b/, /d/, or /g/. The discrimination task used the ABX paradigm. Three stimuli were presented in the order ABX; A and B always differed and X was identical to either A or B. Observers were instructed to indicate whether X was equal to A or B. This judgment was supposedly based on auditory discrimination in that observers were instructed to use whatever auditory differences they could perceive.

The experiment was designed to test the hypothesis that listeners can discriminate the syllables only to the extent that they can recognize them as different phoneme categories. The CP hypothesis was quantified in order to predict discrimination performance from the identification judgments. The authors concluded that discrimination performance was fairly well predicted by identification. This rough correspondence between identification and discrimination has provided the major source of support for CP.

Research in the study of CP has remained oblivious to the valuable scientific strategies of Karl Popper [3] and John Platt [4]. To provide a proper assessment of any theory, it is necessary to determine how closely the predicted performance matches what is observed and to compare the accuracy of this prediction with other the predictions of other theories. When this strategy is followed, one immediately notices just how poorly the categorical describes the results. The problem is that observed discrimination is almost always significantly better than that predicted by identification. Furthermore, it has been shown that is no more accurate in its predictions than is CP [MA87].

We have accumulated, as have other investigators, a variety of sources of evidence against the concept of categorical speech perception (Massaro, 1998). One approach to the question of categorical speech perception is the use of continuous rather than discrete perceptual judgments. Relative to discrete judgments, continuous judgments provide a more direct measure of the listener’s perceptual experience. For example, scientists have found that a binary response proved insensitive to the manipulation of an independent variable whereas confidence ratings revealed significant effects of this variable. In these tasks, subjects were asked to rate the degree to which they felt that the speech stimulus represented one alternative or the other, rather than simply indicating which alternative was presented. Categorical and continuous models of speech perception can be formalized and evaluated against the distribution of repeated rating responses to each test stimulus along a synthetic speech continuum [9]. Categorical and continuous models of speech perception make different predictions about the distribution of repeated rating judgments to a given stimulus along some speech continuum. The results of both synthetic auditory and synthetic visual speech studies provide conclusive evidence that there is continuous information available in speech perception. In agreement with these observations, bimodal speech is also perceived continuously rather than categorically [8].

One might question why we have been so concerned about current theories of speech and language when the emphasis here is speech technology in language learning. The reason is that an understanding of language is fundamental to how we might use technology in language learning. If indeed speech is special and categorically perceived, then the outlook for language learning would in my mind be very grim and I would be at a loss at determining what strategies would be called for. If speech and language can be understood in terms of general principles of perception and learning, on the other hand, we can base our learning paradigm on these principles.

2.2 Fuzzy Logical Model of Perception (FLMP)

These empirical results offer the promise that general principles of perception, memory, and learning are relevant to language learning. Our theoretical framework of the fuzzy logical model of perception (FLMP) also provides an optimistic approach to language learning. We have learned that there are no auditory discontinuities in speech; each distinction has multiple stimulus attributes; and experience is critical. These conclusions are the bedrock of the framework of the FLMP. 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 the FLMP. The three processes involved in perceptual recognition and shown in Figure 1 are evaluation, integration, and decision. These processes make use of prototypes stored in long-term memory. The evaluation process transforms these sources of information into psychological values, which are then integrated to give an overall degree of support for each speech alternative. The decision operation maps the outputs of integration into some response alternative. 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 continuous degree to which that source specifies various alternatives, 2) the sources of information are evaluated independently of one another, 3) the sources are integrated to provide an overall continuous 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.

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 [6, Chapter 1]. These results show how visible speech is processed and integrated with other sources of information. The systematic variation of the properties

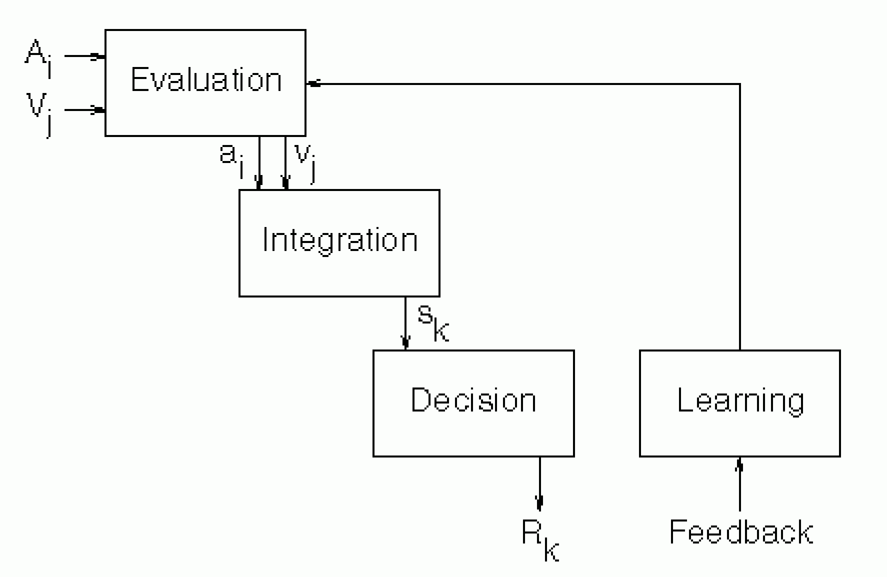

Figure 1. Schematic representation of the FLMP to include learning with feedback. The three perceptual processes are shown to proceed left to right in time to illustrate their necessarily successive but overlapping processing. These processes make use of prototypes stored in long-term memory. The sources of information are represented by uppercase letters. Auditory information is represented by Ai and visual information by Vj. The evaluation process transforms these sources of information into psychological values (indicated by lowercase letters ai and vj) These sources are then integrated to give an overall degree of support, sk, for each speech alternative k. The decision operation maps the outputs of integration into some response alternative, Rk. The response can take the form of a

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 [6,8]. Thus, this research strategy addresses how different sources of information are evaluated and integrated, and can identify the sources of information that are actually used.

Within this framework, we analyze information and information-processing differences among different individuals. Perceivers with hearing loss obviously have less auditory information, but we can also ask whether they differ in terms of information processing. We can ask whether the integration process works the same way regardless of the degree of hearing loss. By comparing individuals using hearing aids to those with cochlear implants [15], we can also address information and information-processing questions in terms of the nature of the assistive device. For example, it is conceivable that integration of the two modalities is more difficult with cochlear implants than with hearing aids.

2.2.1 Learning in the FLMP

Figure 1 also illustrates how learning is conceptualized within the model by specifying exactly how the feature values used at evaluation change with experience. Following the development in Friedman et al. (1995) and Kitzis et al. (1999), learning in the FLMP can be described by the following algorithm. The initial feature value representing the support for an alternative is initially set to .5 (since .5 is neutral in fuzzy logic). A learning trial consists of a feature (such as closed lips at onset) occurring in a test item followed by informative feedback (such as the syllable /ba/). After each trial, the feature values would be updated according to the feedback, as illustrated in Figure 1. Thus, the perceiver uses the feedback to modify the prototype representations and these in turn will become better tuned to the informative characteristics of the patterns being identified.

2.3 Learning Speechreading

Given the importance of the visual modality for spoken language understanding, a significant question is to what extent skill in speechreading can be learned. In addition, it is important to determine whether the FLMP can describe speech perception at several levels of skill. Following the strategy of earlier training studies (e.g., Walden et al., 1977), long-term training paradigm in speechreading was used to test the FLMP across changes in experience and learning (Massaro, Cohen, & Gesi, 1993). The experiment provided tests of the FLMP at several different levels of speechreading skill.

Subjects were taught to speechread 22 initial consonants in three different vowel contexts. Training involved a variety of discrimination and identification lessons with the consonant-vowel syllables. Throughout their training, subjects were repeatedly tested on their recognition of syllables, words, and sentences. The test items were presented visually, auditorily, and bimodally, and presented at normal rate or three times normal rate. Subjects improved in their speechreading ability across all three types of test items. Figure 2 gives their individual performance on the syllables across 7 sessions. The results are plotted in terms of correct viseme classifications, which groups similar visible consonants together. As can be seen in the figure, all six participants improved over training. Replicating previous results (reference), the present study illustrates that substantial gains in speechreading performance are possible.

The FLMP was tested against the results at both the beginning and end of practice. According to the model, a subject would have better information after training than before. To implement this gain in information, we simply assume more informative feature values before and after training. However, the audible and visible sources should be combined in the same manner regardless of training level. Consistent with these assumptions, the FLMP gave a good description of performance at both levels of speechreading skill. Thus, the FLMP was able to account for the gains in bimodal speech perception as the subjects improved their speechreading and listening abilities. This success suggests that the FLMP and its distinction between information and information processing would provide a valuable framework for the study of language learning.

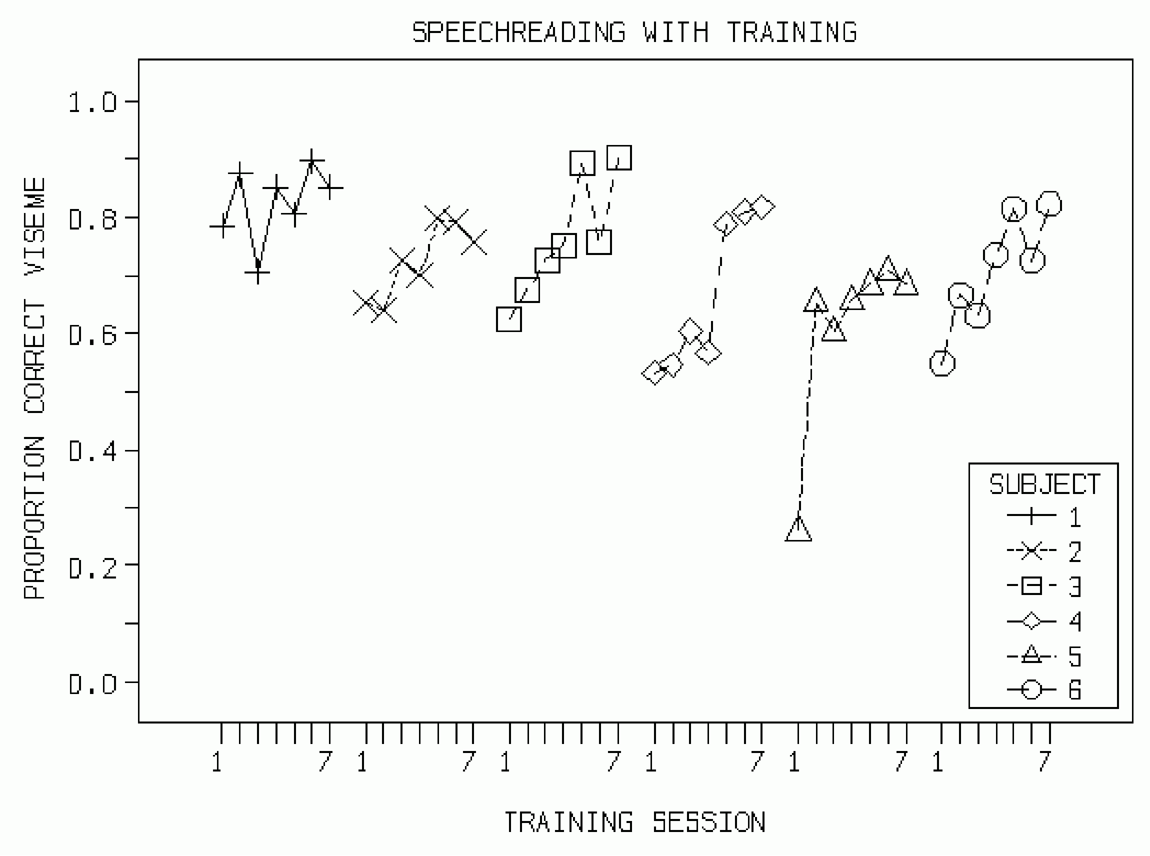

Figure 2. Proportion of correct viseme recognition of the initial consonant in the visible presentation of consonant-vowel syllables, as a function of the seven sessions of training in speechreading for each of the six subjects.

3. Language Learning

This paradigm thus offers a potentially useful framework for the assessment and training of individuals with language delay due to various factors such as hearing impairment [see also 6,7]. Recent research has shown that the FLMP accounts for speech perception in individuals with normal hearing and with hearing loss. An important empirical claim about this algorithm is that while information may vary from one perceptual situation to the next, the manner of combining this information—called information processing--is invariant. With our algorithm, we thus propose an invariant law of pattern recognition describing how continuously perceived (fuzzy) information is processed to achieve perception of a category.

Many communication environments involve a noisy auditory channel, which degrades speech perception and recognition. Visible speech from the talker’s face (or from a reasonably accurate synthetic talking head) improves intelligibility in these situations. Another applied value of visible speech is its potential to supplement other (degraded) sources of information for individuals with hearing loss because it allows effective communication within spoken language for disabled individuals [12,14].

These observations are 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 [11]. Information in the face is particularly effective when the auditory speech is degraded, because of noise, limited bandwidth, or hearing loss. If, for example, only roughly half of a degraded auditory message is understood, its pairing with visible speech can allow comprehension to be almost perfect. 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 [11,13].

3.1 Multimodal Language Learning

There are several reasons why the use of auditory and visual information together is so successful, and why they hold so much promise for language tutoring. These include a) robustness of visual speech, b) complementarity of auditory and visual speech, and c) optimal integration of these two sources of information.

Empirical findings show that speech reading, or the ability to obtain speech information from the face, is robust. Research has shown 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.

Complementarity of auditory and visual information simply means that one of the sources is most informative in those cases in which the other is weakest. Because of this, a speech distinction is differentially supported by the two sources of information. That is, two segments that are robustly conveyed in one modality are relatively ambiguous in the other modality. For example, the difference between /ba/ and /da/ is easy to see but relatively difficult to hear. On the other hand, the difference between /ba/ and /pa/ is relatively easy to hear but very difficult to discriminate visually. The fact that two sources of information are complementary makes their combined use much more informative than would be the case if the two sources were non-complementary, or redundant [11].

The final characteristic is that perceivers combine or integrate the auditory and visual sources of information in an optimally efficient manner. There are many possible ways to treat two sources of information: use only the most informative source, average the two sources together, or integrate them in such a fashion in which both sources are used but that the least ambiguous source has the most influence. Perceivers in fact integrate the information available from each modality to perform as efficiently as possible. A wide variety of empirical results have been described by the FLMP, which describes an optimally efficient process of combination.

Our recent analysis of research from several different laboratories has shown that both children and adults with hearing loss benefit greatly from having visible speech presented jointly with the necessarily degraded audible speech. Normal-hearing participants also show a much larger influence of visible speech when the auditory speech is degraded [10, pp.42-43]. According to our perspective, this result is entirely understandable. Observers with hearing loss integrate information in the same manner as those with normal hearing, but they have less auditory information. One type of observer can be made to resemble the other by assigning the appropriate quality of information.

Recent research with individuals with hearing loss has confirmed many of the principles derived from recent experimental and theoretical studies of individuals with normal hearing [12]. Experiments with individuals with hearing loss tend to be more ecologically valid in that many more stimuli and response alternatives are used. The extension of the FLMP to these data sets was successful along several dimensions. First, the assumptions of the model appear to be equally powerful in describing the confusion matrices as they are in describing simpler experiments using expanded factorial designs. Second, the FLMP was extended to incorporate features as sources of information in speech perception.

These positive findings encourage the use of multimodal environments for persons with hearing loss. Ling [8, p. 51], however, reports that clinical experience seems to show that "children taught exclusively through a multisensory approach generally make less use of residual audition." For these reasons, speech-language pathologists might use bimodal training less often than would be beneficial. To evaluate multisensory control of speech production, the same type of research design used for the study of speech perception is in place to study speech production. It is well known that individuals with severe or profound hearing loss tend to have poorer speech production skills. An experiment is underway in which the children with hearing loss are asked to produce speech given auditory, visual, or bimodal speech input. The working hypothesis is that speech production will be better (and learned more easily) given bimodal input relative to either source of information presented alone.

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 and articulators. Our goal is to create an articulatory simulation as accurate as possible, and to assess whether this information can guide speech production. We know from children born without sight that the ear alone can guide language learning. Our question is whether the eye can do the same, or at least the eye supplemented with degraded auditory information from the ear.

4. Advantages of Synthetic Talking Heads

We have developed, evaluated and implemented a computer-animated talking head, Baldi [11], incorporated it into a general speech toolkit, and are using it as part of an NSF Challenge Grant to develop interactive learning tools for language training with children with severe hearing loss [2,3]. The synthesis program controls a wireframe model, which is textured mapped with a skin surface. Realistic speech is obtained by animating the appropriate facial targets for each segment of speech along with the appropriate coarticulation Baldi is controlled by text-to-speech synthesis and can be appropriately aligned with either synthetic or with natural speech. Paralinguistic information and emotion are also expressed during speaking.

The fact that this technology is always available, whenever the user chooses, meshes well with what is known about maximizing learning and memory. Learning increases with the time spent on the task. This law, called the total time function, can be summarized by the aphorism, "you get what you pay for." Or, to put it another way, "no pain, no gain." A second important variable is how a given amount of time on a task is distributed. Research by psychologists has repeatedly demonstrated that spacing practice over a longer time leads to better learning than massing practice within a shorter time. This outcome is highly general and holds across an amazing variety of skills. Baldi and accompanying instruction is available 24 hours a day, 365 days a year. Baldi doesn't become tired or bored and isn't waylaid by everyday distractions; he is in effect a perpetual motion machine. For this reason, students can spend an inordinate amount of time on task and can also space this practice rather than massing it into a short time frame.

Children with hearing-impairment 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 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. We have augmented the internal structures of our talking head both for improved accuracy and to pedagogically illustrate correct articulation [1]. A new tongue, hard palate, and three-dimensional teeth are present, along with target values that have been computed from electropalatography and ultrasound data. The goal is to instruct the child by revealing the appropriate articulation via the hard palate, teeth and tongue.

Visible and bimodal speech instruction poses many issues that must be resolved before training can be optimized. We are confident that an 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, as also suggested by other researchers [16]. 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. Finally, we will evaluate the influence of providing visual feedback about the student’s own articulation. There is some evidence that video feedback from their own speech production improved the speech production of adults with profound hearing loss [4].

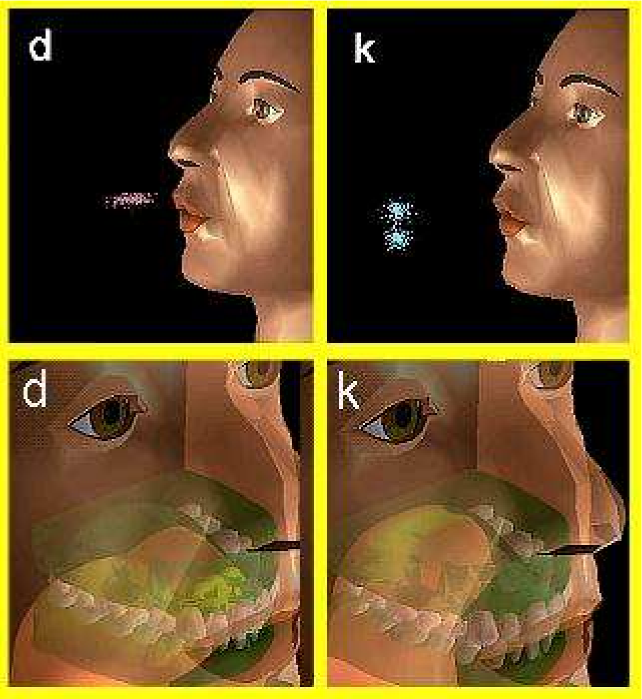

Figure 3 displays two types of visual enhancements designed to teach phonological awareness. In the top panel, Baldi is shown with supplementary visual features displaying the shape and location of the stop bursts following the release of consonants in the syllables /du/ and /ku/. Note that these snapshots were taken after the consonant release while the face is transitioning to the following vowel /u/. These patterns were designed to show that the burst of the voiceless stop /k/ is longer than the burst of the voiced stop /d/ in these syllables (it travels further from the lips). The bursts were also designed to reflect their different acoustic and temporal characteristics. In the bottom panel, Baldi is shown with a half-face display in semi-transparent mode, showing the position of the tongue at the onset of the syllables /da/ and /ka/. In color images of this figure, the area of the tongue touching the palate is highlighted (see http://mambo.ucsc.edu/psl/xxx).

We also expect progress will result from both hard work and serendipitous discoveries. To mention just one instance of serendipity, language tutoring has always necessarily proceeded by the student watching a frontal (or perhaps a profile) view of the instructor. As already mentioned, one downside to this interaction is that the skin hides much of the vocal tract. These vital parts can be revealed within Baldi’s mouth by making his skin transparent or by presenting a mid-sagittal view. One interesting observation was that a unique view could 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 of the target and the student's own production apparatus might facilitate speech production learning. An analogy is using a map. We tend to orient the map to 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 goal is to enhance the cues for visible speech perception. Baldi can be made to be not only realistic, he can be made superrealistic by overarticulating and adding other somewhat natural embellishments of the visible speech. Several alternatives are obvious for distinguishing phonemes within a viseme class. A major confusion is between voiced and voiceless segments. Baldi's neck could be made to vibrate during voicing. In this way, a vibrating neck would occur during voiced but not voiceless segments. The segments /s,z/ tend to be longer in duration than the similarly looking segments /t,d/. This cue is somewhat subtle, but apparently can be learned. To emphasize it, the articulation of /s,z/ could be made more distinctive by spreading the lips more, clenching the teeth more, and even grinning during the articulation [5]. The overlap of the upper teeth on the lower lip could be made more extreme for the segments /f,v/. To distinguish /k,g/ from /t,d/, the jaw could be moved downward to a greater extent. Also, some throat movement might be made to signify an articulation further back in the throat. The segment /h/ could be uttered with some breathy aspiration. The vowels could be made more distinctive by accentuating the height, width, and depth of the lip movements. Also duration could be made more distinctive for the normally long and short vowels. This hyperarticulated speech along with additional cues could make the face more informative than it normally is. Finally, supplementary visual displays for English consonants can be presented along with the face to help teach the articulatory and acoustic properties of the segments thereby enhancing phonological awareness [16].

5. Psychology of Instruction

Our experiences have convinced us that several new trends and challenges come to the forefront with technology-driven education. We envision several new roles for teachers. Rather than actively teaching, the technology promotes the teacher to a more interactive role in the classroom. They become much more active, collaborative and effective, since they can watch each student interact with the program they designed, understand individual problems, and assist when necessary. The classroom becomes an interactive learning environment with as many tutors as students, and with the teacher monitoring learning, Within this new learning environment, teachers become less didactic and more collaborative and thus are implicitly fulfilling a goal of reflective rather than standard education [9].

A second new role for teachers involves acquiring and providing a degree of technology literacy, which was not anticipated in their formal training or experience. To exploit the assistive technology tools, the teachers have to become facile in the use of the speech toolkit and to assume the role of technologist when there are failures in the classroom. Of course, teachers are expected to be much more than computer jocks but some expertise appears to be a necessary dimension of this enterprise.

Imagine a teacher and a doctor, both from the last century, returning to life today. The doctor would be absolutely useless in today's medical environment. The teacher, on the other hand, would be fairly comfortable in the current educational establishment. Education has progressed much slower than medicine. We believe that psychological theory combined with technology will dramatically change this situation.

5.1 Components of Learning Episodes

Any learning episode seems to have four essential components. The first is a goal in terms of the target behavior to be achieved. The specific goal we chose was to instruct children with hearing loss on speech production in order to determine whether speech production could improve. What we immediately discovered, however, was that the tools we provided were recruited for instructional domains well beyond what we had originally envisioned. As described in the accompanying papers of this symposium, Baldi and the toolkit have been integrated into every aspect of the child's learning environment. Baldi's presence, guidance, and support are part and parcel of the child's school day. These one-on-one exercises provide the child with a focused time on task that is not feasible without computer-assisted instruction. Given this expanded domain of our pedagogy and technology, our specific goal of assessment of language tutoring could easily have been compromised. Although the children are receiving concentrated language experiences in a variety of domains, we are in the midst of testing our specific research hypothesis.

The second component is an understanding of the processes involved in achieving the target behavior. At present, we know very little about language tutoring of speech production and even less about the first-language acquisition of children with hearing loss. Our research goals should help fill this gap in knowledge.

The third component is a curriculum for assessment of the initial state of the student and intermediate states during the learning experience. Assessment is very difficult but not impossible within our application setting. We do not have complete control over the school or classrooms, and it is very difficult to isolate some contribution of the technology relative to just a general learning experience. Even so, we expect to be able to test specific hypotheses about learning on an individual student basis.

The fourth is some final assessment of the achievements of the students. A final assessment in our situation is not appropriate because learning and its application should not end. Proponents of situational learning point out that traditional classroom instruction appears to generalize very little to everyday life. They advocate an integration of the curriculum with the needs and goals of the students. It is critical that our learning applications are designed to transfer as much as possible to everyday life.

Acknowledgements

This work was supported in part by NSF grant ECS-9726645, NSF CHALLENGE grant CDA-9726363, Public Health Service (PHS R01 DC00236), National Science Foundation (23818), Intel Corporation, and the University of California Digital Media Program..

References

[1] Cohen, M. M., Beskow, J., & Massaro, D.W. (1998). Recent developments in facial animation: An inside view. Proceedings of the International Conference on Auditory-Visual Speech Processing—AVSP'98 (pp. 201-206). Terrigal, Australia.

[2] Cole, R., Carmell, T., Connors, P., Macon, M., Wouters, J., deVilliers, J., Tarachow, A., Massaro, D.W., Cohen, M.M., Beskow, J., Yang, J., Meier, U., Waibel, A., Stone, P., Fortier, G., Davis, A., Soland, C. (1998). Intelligent Animated Agents for Interactive Language Training. Proceedings of Speech Technology in Language Learning. Stockholm, Sweden.

[3] Cole, R., Massaro, D. W., Villiers, J., Rundle, B., Shobaki, K., Wouters, J., Cohen, M., Beskow, J., Stone, J., Connors, P., Tarachow, A., Solcher, D. (1999). New tools for interactive speech and language training: Using animated conversational agents in the classrooms of profoundly deaf children. Proceedings of ESCA/Socrates sponsored Method and Tool Innovations for Speech Science Education (MATISSE) workshop. London: University College London.

[4] Cole, R. A.; Rudnicky, A. I. (1983). What's new in speech perception? The research and ideas of William Chandler Bagley, 1874-1946. Psychological Review, 90, 94-101.

[5] De Filippo, C. L,; & Sims, D. G. (1995). Linking visual and kinesthetic imagery in lipreading instruction. Journal of Speech and Hearing Research, 38, 244-256.

[6] Erber, N. P. (1996). Communication therapy for adults with sensory loss. Melbourne, Australia: Clavis.

[7] Friedman, D., Massaro, D.W., Kitzis, S.N., & Cohen, M.M., (1995) "A Comparison of Learning Models," Journal of Mathematical Psychology, 39, 164-178.

[8] Grant, K. W., & Walden, B. E. (1995). Predicting auditory- visual speech recognition in hearing-impaired listeners. Proceedings of the XIIIth International Congress of Phonetic Sciences, 3, 122-129.

[9] Grant, K. W.; Walden, B. E.; Seitz, P. F. (1998). Auditory-visual speech recognition by hearing-impaired subjects: Consonant recognition, sentence recognition, and auditory-visual integration. Journal of the Acoustical Society of America, 103, 2677-2690.

[10] Kitzis, S.N., Kelley, H., Berg, E., Massaro, D.W., & Friedman, D., (1999), "Broadening the tests of learning models," Journal of Mathematical Psychology, 42, 327-355.

[11] Ling, D. (1976). Speech and the hearing-impaired child: Theory and practice. Washington, DC: Alexander Graham Bell.

[12] Lipman, M. (1991). Thinking in Education. New York: Cambridge University Press.

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

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

[15] Massaro, D. W. (1999). Speechreading: Illusion or window into pattern recognition. Trends in Cognitive Sciences, 3, 310-317.

[16] Massaro, D.W., & Cohen, M.M. (1999). Speech Perception in Perceivers with Hearing Loss: Synergy of Multiple Modalities. Journal of Speech, Language, and Hearing Research, 42,21-41.

[17] Massaro, D.W., Cohen, M.M., & Gesi, A.T., (1993) "Long-term Training, Transfer, and Retention in Learning to Lipread," Perception and Psychophysics, 53(5), 549-562.

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

[19] Oerlemans, M., & Blamey, P. (1998). Touch and auditory-visual speech perception. In Campbell, R., Dodd, B., & Burnham, D. (Eds.), Hearing by Eye II (pp. 267-281). East Sussex, UK: Psychology Press.

[20] Saffran, J. R.; Johnson, E. K.; Aslin, R. N.; Newport, E.L. (1999). Statistical learning of tone sequences by human infants and adults. Cognition, 70, 27-52.

[21] Schindler, R.A. & Merzenich, M.M. (1985) Cochlear Implants. New York: Raven.

[22] Summerfield, Q. (1987). Some preliminaries to a comprehensive account of audio-visual speech perception. In B. Dodd and R. Campbell (Eds.).Hearing by eye: the psychology of lip-reading (pp. 3-51) Hillsdale, NJ: Lawrence Erlbaum Associates.

[23] Walden, B., Prosek, R., Montgomery, A., Scherr, C. K., & Jones, C. J. (1977) Effects of training on the visual recognition of consonants. Journal of Speech and Hearing Research, 20, 130-145.