FROM THEORY TO PRACTICE: REWARDS AND CHALLENGES

Dominic W. Massaro, Michael M. Cohen, and Jonas Beskow

Perceptual Science Laboratory, Dept. of Psychology, Univ. of California, Santa Cruz, CA 95064 USA

ABSTRACT

Language perceivers are viewed as having available multiple sources of information supporting the identification and production of language. This theoretical framework has been successful in accounting for a wide variety of empirical research findings. Our research agenda called for the development of a computer-animated talking head, Baldi, who serendipitously showed promise for language tutoring. This facial animation software was fully integrated into a speech toolkit and is currently an integral part of an NSF Challenge grant to develop interactive learning tools for language training with profoundly deaf children. In addition to Baldi, the tools to date combine the key technologies of speech recognition, speech synthesis, and a rapid application developer platform. We describe our theoretical framework, how it is used to guide language training, and plans for assessment of its efficacy.

1. THEORETICAL FRAMEWORK

We envision speech perception as an instance of a more general process of pattern recognition in which persons use multiple sources of information [11]. Recognition is achieved via a variety of bottom-up and top-down sources of information. Our research addresses both empirical and theoretical issues. At the empirical level, experiments are carried out to determine how visible speech is combined with auditory speech for a broad range of individuals and across a wide variation of situational domains. At the theoretical level, the assumptions and predictions of several models are formalized, analyzed, contrasted, and tested. Various types of model fitting strategies have been employed, with similar outcomes. These model tests have been highly informative with respect to improving our understanding of how spoken language is perceived and understood.

A wide variety of results have been described within the framework of a fuzzy logical model of perception (FLMP). Within this model, perceivers are assumed to have available multiple sources of information supporting the identification and interpretation of the language input. The assumptions central to the model are 1) each source of information is evaluated to give the *continuous* degree to which that source specifies various alternatives, 2) the sources of information are evaluated *independently* of one another, 3) the sources are integrated *multiplicatively* to provide an overall degree of support for each alternative, and 4) perceptual identification and interpretation follows the *relative* degree of support among the alternatives.

Independence of sources is motivated by the principle of category-conditional independence [13]: it isn't possible to predict the evaluation of one source on the basis of the evaluation of another, so the independent evaluation of both sources is necessary to make an optimal category judgment. While sources are thus kept separate at evaluation, they are then integrated to achieve perception, recognition, and interpretation. Multiplicative integration yields a measure of total support for a given category identification. This operation, implemented in the model, allows the combination of two imperfect sources of information to yield better performance than would be possible using either source by itself. However, the output of integration is an absolute measure of support; it must be relativized, which is implemented through a decision stage, which divides the support for one category by the summed support for all categories.

Given this framework, we are able to make a distinction between "information" and "information processing." The sources of information from the auditory and visual channels make contact with the perceiver at the evaluation stage of processing. The reduction in uncertainty effected by each source is defined as information. In the fit of the FLMP, for example, the parameter values indicating the degree of support for each alternative from each modality correspond to information. These parameter values represent how informative each source of information is. Information processing refers to how the sources of information are processed. In the FLMP, this processing is described by the evaluation, integration, and decision stages.

Within this framework, we 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.

This paradigm thus offers a potentially useful framework for the assessment and training of individuals with 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.

2. IMPORTANCE OF TALKING FACES IN DIALOG

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].

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.

3. IMPLICATIONS FOR LANGUAGE TRAINING

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 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. Recently, 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].

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.

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.

5. PSYCHOLOGY OF INSTRUCTION

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.

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.

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] 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.

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

[6] 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.

[7] 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.

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

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

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

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

[12] 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.

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

[14] 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.

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

[16] 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.