Using Visible Speech to Train Perception and Production of Speech for Individuals With Hearing Loss

Dominic W. Massaro
Joanna Light
University of California, Santa Cruz

The main goal of this study was to implement a computer-animated talking head, Baldi, as a language tutor for speech perception and production for individuals with hearing loss. Baldi can speak slowly; illustrate articulation by making the skin transparent to reveal the tongue, teeth, and palate; and show supplementary articulatory features, such as vibration of the neck to show voicing and turbulent airflow to show frication. Seven students with hearing loss between the ages of 8 and 13 were trained for 6 hours across 21 weeks on 8 categories of segments (4 voiced vs. voiceless distinctions, 3 consonant cluster distinctions, and 1 fricative vs. affricate distinction). Training included practice at the segment and the word level. Perception and production improved for each of the 7 children. Speech production also generalized to new words not included in the training lessons. Finally, speech production deteriorated somewhat after 6 weeks without training, indicating that the training method rather than some other experience was responsible for the improvement that was found.

KEY WORDS: visible speech, language learning, hearing loss, speech perception, speech production

In the United States, 1–2 infants per 1,000 have a moderate to severe hearing loss in both ears (U.S. Department of Health and Human Services, 2002). This loss often goes unnoticed for a considerable period of time. If untreated for too long, hearing loss can have severe effects on language learning. According to the Gallaudet Research Institute (1999–2000), 90% of children with hearing loss are born to parents with normal hearing. These parents are forced to decide what communication method they will choose for their child (oral, manual, or a combination of both). Although there is no consensus on the best medium through which children who are deaf should learn language, the communication method that parents choose for their child is one that should optimize language learning and quality of life. Independently of the communication method of choice, the amount and quality of language in and out of the classroom is the number one factor leading to communication and academic success (National Association of State Directors of Special Education, 1992).

Parents often choose to educate their children through the oral communication method. Although it is possible for some children who are profoundly deaf to develop excellent spoken language, many do not (Dodd,

McIntosh, & Woodhouse, 1998). Children with even moderate hearing loss are not exposed to the wealth of auditory input that is available to the hearing child (Sanders, 1988). Because of their degraded auditory language input, children with hearing loss learning oral language must depend on distorted speech and perhaps insufficiently informative mouth movements.

Correct perception and production of all phonemes in a language is essential for spoken language learning (Jusczyk, 1997). Results have indicated that the better a child with hearing loss can perceive spoken language, the better he/she can approximate development of spoken language compared to his/her counterparts with normal hearing (Svirsky, Robbins, Kirk, Pisoni, & Miyamoto, 2000). The better a child is at perceiving and understanding spoken words, the better he/she will be at producing spoken language (Levitt, McGarr, & Geffner, 1987).

Children with early onset deafness generally lag significantly behind their normally hearing peers in all areas involving speech—speech perception and production, oral language development, metaphonological abilities, and reading and spelling (Leybaert, Alegria, Hage, & Charlier, 1998). Listeners often have trouble understanding speakers who are deaf. In one study, inexperienced listeners could understand only about 20% of the speech output of deaf talkers (Gold, 1980). Whether intentional or not, the way one speaks can ultimately affect the way others perceive one (Scherer, 1986). This difficulty in oral communication may result in feelings of social isolation on the part of the deaf individual. Thus, deafness may vastly affect both the child's academic and vocational achievement.

As far back as Hudgins and Numbers (1942, as cited in Ling, 1976), researchers have primarily focused on pinpointing the speech segments that are most difficult for individuals with hearing loss to produce (e.g., Kirk, Pisoni, & Miyamoto, 1997). The most common articulation problems made by individuals with hearing loss are voiced–voiceless errors, omissions/distortions of initial consonants, omission of consonants in clusters, omissions/distortions of final consonants, nasalization, substitution of one consonant for another, and intrusive voicing between neighboring consonants.

Assistive technology is one means by which children experiencing communication difficulties can be helped. Along with the evolving technology already in use (e.g., hearing aids, cochlear implants), technological advancements can potentially provide individuals who are deaf with some of the help they need to perceive and speak more intelligibly. Because speech training is a labor-intensive task, requiring endless hours of one-on-one training between child and clinician, interactive technology may offer a promising and cost-effective means to improve the perception and production skills of speech-impaired individuals. Tailoring training lessons based on the specific needs of the student allows for child-centered instruction, increased time on task, speech training outside of the classroom and treatment setting, and ideally increased competence and confidence in perceiving and producing English speech segments.

Speech and language science evolved under the assumption that speech perception was a solely auditory event (Denes & Pinson, 1963). However, a burgeoning record of research findings reveals that our perception and understanding are influenced by a speaker's face and the accompanying visual information about gestures, as well as the actual sound of the speech (Dodd & Campbell, 1987; Massaro, 1987, 1998; McGurk & MacDonald, 1976). Perceivers expertly use these multiple sources of information to identify and interpret the language input. Information from the face is particularly effective when the auditory speech is degraded because of noise, limited bandwidth, or hearing loss. If only roughly half of a degraded auditory message is understood, for example, adding visible speech can allow comprehension to be almost perfect. The combination of auditory and visual speech has been called superadditive because their combination can lead to accuracy that is much greater than the sum of the accuracies on the two modalities presented alone (Massaro, 1998). 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 nonsense auditory sentence, *My bab pop me poo brive*, is paired with the nonsense visible sentence, *My gag kok me koo grive*, the perceiver is likely to hear, *My dad taught me to drive*. Two sources of nonsense are combined to create a meaningful interpretation (Massaro, 1998; McGurk, 1981).

In addition to the information value of visible speech, 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: robustness of visual speech, complementarity of auditory and visual speech, and optimal integration of these two sources of information.

Speechreading, or the ability to obtain speech information from the face, depends somewhat on the talker, the perceiver, and the viewing conditions (Bernstein, Demorest, & Tucker, 2000; Massaro, 1998; Massaro & Cohen, 1999). Even so, empirical findings show that speechreading is fairly robust (Massaro, 1998). Research has shown that perceivers are fairly good at speechreading even when they are not looking directly at the talker's lips (Smeele, Massaro, Cohen, & Sittig, 1998). Furthermore, accuracy is not dramatically reduced when the facial image is blurred (because of poor vision, for example); when the face is viewed from above, below, or in

profile; or when there is a large distance between the talker and the viewer (Massaro, 1998, chap. 14).

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, most speech distinctions are 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 noncomplementary, or redundant (Massaro, 1998, chap. 14).

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 that both sources are used but 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 Fuzzy Logical Model of Perception (FLMP), which describes an optimally efficient process of combination (Massaro, 1987, chap. 7).

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 (for a review, see Massaro & Cohen, 1999). Although individuals with hearing loss have less auditory information, they integrate information in the same optimal manner as those with typical hearing. There is also some evidence that individuals with hearing loss become experts in speechreading (e.g., Bernstein et al., 2000).

These positive findings encourage the use of multimodal environments for persons with hearing loss. Ling (1976, 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. The working hypothesis of the present study is that visible speech as well as auditory speech can be productively included in the training of speech perception and production.

Although there is a long history of using visible cues in speech training for individuals with hearing loss, these cues have usually been somewhat indirect representations of the actual articulatory and phonetic properties of the speech. In *cued speech* (Cornett, 1988; Morais & Kolinsky, 1994), for example, the talker uses shape and movement of the hands to indicate additional information not transmitted by speechreading alone. For example, the thumb-up cue indicates that the phoneme can be identified /t/, /m/, or /f/. This cue is necessarily symbolic or abstract in that there is nothing in the cue that corresponds to the articulation of these segments. Our goal, on the other hand, is to directly illustrate the vocal tract and articulators during production and to assess whether this information can facilitate the learning of speech perception and production.

A few studies have tested the efficacy of training methods with visible speech for speech perception and speech production (e.g., Dagenais, 1992; Osberger, 1987). Dagenais provided speech training for individuals with hearing loss using glossometry and palatometry techniques along with the traditional aural/oral training method proposed by Ling (1976), which includes amplified residual hearing with oral training of auditory contrasts. Both glossometry and palatometry techniques use a false palate, which is like a dental retainer with an array of sensors that are activated when they are contacted by the tongue. The glossometry system indicates the location of the surface of the tongue in the oral cavity by displaying distances between sensors and the tongue on a monitor. The palatometry system shows the contact made between tongue and palate. The spatial layout of the sensors is shown in a computerized display, which is used to illustrate the sensors that should be contacted for the production of a given segment. The goal of the learner is to match the idealized display using biofeedback techniques. In Dagenais's study, lines were used to signify which sensors should be contacted during the production of a given segment. Dots indicated the sensors that should not be contacted. As feedback, a line changing to a square indicated a correct response and an error was indicated by a dot changing to an asterisk. Each student with hearing loss received three half-hour training sessions per week for the 1st year of training and two half-hour sessions per week for the 2nd year. Training totaled 120 hours over the course of 2 years and involved teaching the vowels /i, æ, a, u/ and the consonants /t, d, k, g, s, z, ʃ/. Contact versus noncontact status of each sensor was recorded and scored in terms of percentage of correct contacts between the tongue and palate during production. Percentage of correct contact scores improved roughly from 71% at pretest to 79% after 6 months (36 hours) of training. These results provide support for the use of physiological feedback and visual presentation for training individuals with hearing loss to produce speech. The effectiveness of this type of training will be used to help evaluate our study, which

uses a new noninvasive visual technique to train speech perception and production.

We have developed, evaluated, and implemented a computer-animated talking head, Baldi, incorporated it into a general speech toolkit, and used this technology to develop interactive learning tools for language training for children with language challenges (Bosseler & Massaro, 2003; Massaro & Light, in press). The facial animation program controls a wireframe model, which is texture 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 can be appropriately aligned with either synthetic or natural speech. Paralinguistic information (e.g., amplitude, pitch, and rate of speech) and emotion are also expressed during speaking (Massaro, Cohen, Tabain, Beskow, & Clark, in press).

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. Baldi has a tongue, hard palate, and three-dimensional teeth, and his internal articulatory movements have been trained with electropalatography and ultrasound data from natural speech (Cohen, Beskow, & Massaro, 1998). These internal structures can be used to pedagogically illustrate correct articulation. The goal is to instruct the child by revealing the appropriate movements of the tongue relative to the hard palate and teeth.

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 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 embellishment 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 Baldi more informative than he normally is.

Students were trained to discriminate minimal pairs of words bimodally (simultaneous auditory and visual input) and were also trained to produce various speech segments by visual information about how the inside oral articulators work during speech production. 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 take 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 with the same contact, hold, and release phases (Ladefoged, 2001). The time course of articulation and how the air escaped the mouth differed (e.g., fricative: slow, consistent turbulent airflow; affricate: quick, abrupt turbulent airflow).

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); neighboring vowel quality (height and front/backness features); and, in the case of consonant cluster training, neighboring consonant quality (place and manner features) 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 pretest and posttest 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 6 weeks after posttest. We expected that performance would be greater than pretest but not as high as posttest levels due to discontinued use of training.

Method

Students

Seven students with hearing loss (2 boys and 5 girls) from the Jackson Hearing Center and JLS Middle School in Los Altos, California, participated in the study. All children were mainstreamed in the school, where oral language was the communication method present in all classrooms. The students ranged in age from 8 to 13 at the start of the study. Their unaided hearing varied to some extent, but all children had a severe hearing loss in at least one ear. As shown in Table 1, the aided hearing threshold levels of our students were still within the mild-to-moderate range.

According to their regular teachers and speech-language pathologists, all of these students had difficulty perceiving and producing certain English phonemes. At our request, these instructors provided us with a set of speech segments that students could benefit from being trained in. These speech segments are shown in Table 2 and are consistent with what has been reported in the past as problematic for individuals with hearing loss to produce (Kirk et al., 1997; Ling, 1976). Every student was trained on this entire set of segments.

Testing and training were carried out individually by the second author (JL) in a quiet room at the child's school. All children wore their personal aids while participating in the study. Students sat at a personal desk that was equipped with a laptop computer, external speakers (Model PCVA-SP1; Sony Corporation, Tokyo, Japan), and an external microphone (Model QS-5841; Quickshot Technology, Inc., El Monte, CA). Stimuli were presented on the laptop monitor and through the speakers set by each child to a comfortable listening level. The sound card was a Maestro Wave/WaveTable Synthesis Device provided by ESS Technology, Inc. (Fremont, CA). Most often, the listening level was set to maximum loudness without distortion (79 dBA; B & K 2203 sound-level meter, Brüel & Kjær, Nærum, Denmark). The sampling rate for digitizing the participants' productions was 8 kHz.

A computer program with Baldi as the instructional agent carried out all of the testing and training. Baldi's articulation was aligned with auditory speech produced by the Festival text-to-speech synthesis system (Black & Taylor, Caley & Clark, 1997), whose rate, pitch, and intensity were specified exactly (speech rate: 140 words per minute; pitch: 120 Hz; pitch range: 20 Hz). The program was designed using the rapid application developer (RAD) in the Center for Spoken Language Understanding (CSLU) speech toolkit (Massaro, Cohen, Beskow, & Cole, 2000). Even though the experimenter was present during each lesson and could be considered a source of distraction, we encouraged the students to attend to the computer screen and look at Baldi's face when he spoke.

Table 1. Individual and average unaided and aided auditory thresholds (dB HL) for four frequencies for the 7 students who participated in the current study.

Student no.	Aids	Unaided/aided auditory thresholds (dB HL) 500 Hz	1000 Hz	2000 Hz	4000 Hz	Age (years)
1	Binaural	60/25	60/15	75/30	85/45	12
2	Binaural	60/28	80/25	85/30	80/30	11
3	Binaural	85/50	90/55	85/55	80/55	8
4	Binaural	45/15	60/35	65/45	60/50	11
5	Cochlear implant, left ear	95/40	110/25	115/25	105/35	13
6	Binaural	65/30	95/30	100/35	110/70	13
7	Right ear	65/30	55/35	65/40	70/60	8
Average		68/31	79/31	84/37	84/49	11

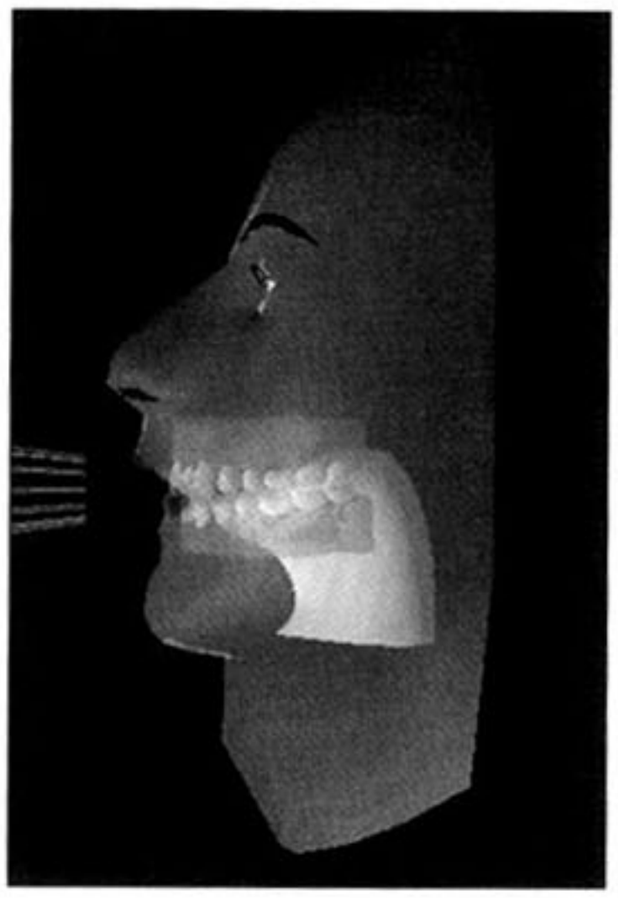
Table 2. The speech segments that were trained in the present study.

Type of training lesson	Segments involved	Example of words involved in test phase of perception and production training	Number of trials in test phase of production training
voiceless vs. voiced	/f/ vs. /v/	Fan vs. van Fine vs. vine Leaf vs. leave Fail vs. veil	12
voiceless vs. voiced	/θ/ vs. /ð/	Bath vs. bathe Breath vs. breathe Teeth vs. teethe	12
voiceless vs. voiced	/s/ vs. /z/	Sip vs. zip Fussy vs. fuzzy Mace vs. maze Sap vs. zap	12
voiceless vs. voiced vs. voiced	/t/ vs. /d/ vs. /b/	Rot vs. rod vs. rob Till vs. dill vs. bill Tie vs. die vs. buy /aɪ/	18
fricative vs. affricate	/ʃ/ vs. /ʧ/	Dish vs. ditch Shop vs. chop Shoe vs. chew	12
Consonant cluster training	/r/ initial clusters	Pray Free Cry	12
Consonant cluster training	/s/ initial clusters	Smile Stare Slit	12
Consonant cluster training	/l/ final clusters	Milk Malt Help	12

Procedures

Based on articulatory difficulties identified by the participants' teachers, eight programs were developed. Four were used to teach the distinction between voiceless and voiced cognates: /f/ versus /v/, /s/ versus /z/, /t/ versus /d/ versus /b/ and /θ/ versus /ð/. Because the instructors indicated that practice with /p/ (the voiceless counterpart of /b/) was not necessary, we combined the three plosives /t/, /d/, and /b/ into one training program. As with the traditional method proposed by Ling (1976), our method included training at the segment and word level. We added supplementary features consisting of visible vibrations (quick back and forth movements of the virtual larynx) in Baldi's neck whenever the segments were voiced. An air stream expelled from Baldi's mouth was also used to differentiate these segments (e.g., a considerable amount of air for voiceless segments and a limited amount for voiced counterparts; see Figure 1). Baldi's speech rate was slowed down to 100 words per minute, 65% of the normal rate (155 words per minute), to illustrate these distinctions. Three programs involved consonant clusters: two word initial clusters involving /r/ (e.g., ***cry***, ***grow***, ***free***) and /s/ (e.g., ***sm****ile*, ***slit***, ***st****are*), and one word-final cluster involving /l/ (e.g., *be****lch***, *mi****lk***, *fie****ld***). For these three programs, Baldi's speech rate was slowed down even further to 47 words per minute, or 30% of the normal rate. As shown in Figure 2, inside oral articulators were also revealed to teach the articulatory processes involved in producing consonant clusters. A final program taught the difference between the fricative /ʃ/ and the affricate /ʧ/. This program used methods that were involved in teaching both voiced versus voiceless distinctions and consonant clusters. Slowing down Baldi's speech to 47 words per minute, 30% of the normal rate, while revealing Baldi's inside oral articulators provided a perceivable difference between these two segments. With Baldi's instruction, the students were able to visibly determine that the starting position

Figure 1. A side view of Baldi giving supplementary features (vocal cord vibration involving vibration of the neck for voiced segments and turbulent airflow involving rays emanating from the mouth for voiceless segments).



of articulation was different for these two segments and that the affricate was actually a combination of two segments produced simultaneously (/t/ + /ʃ/ = /ʧ/). Supplementary voicing features were also used in training this distinction. Figure 3 illustrates the procedure of this study in its entirety.

On the first day, before the pretest was given, each student was required to give specific information about him/herself, including name, age, and date, in order to set up a file for his/her data. On each subsequent day, each student was required to sign in, and their data were recorded and stored in the student's file.

Pretest

The pretest consisted of 104 words, which included all of the training segments in all contexts. Baldi said each word, along with its orthographic captioning presented on the screen beneath his chin, to ensure that the student understood the intended stimulus correctly. The student was required to repeat that word aloud after a tone sounded. The utterances of the students were spoken into the external microphone, saved on the computer for each student, and used for evaluation at a later date. Each subsequent training lesson required the student to log in using only his/her name so that progress could be tracked individually. The date of training was also recorded by the experimenter.

Training

Each student completed two training lessons per week over the course of 21 weeks, including a 2-week break when the schools were closed for holiday vacation. Occasionally, because of the child's absence from school, the scheduled training lesson was simply presented at the next meeting. Each of the eight training lessons lasted for approximately 15 minutes and was completed three times over the course of the study. Thus, each student completed approximately 45 minutes of eight training lessons for a total of 6 hours of training. After the students completed a specific training lesson, the program was modified to take into consideration

Figure 2. The four presentation conditions of Baldi with transparent skin revealing inside articulators (back view, sagittal view, side view, front view).

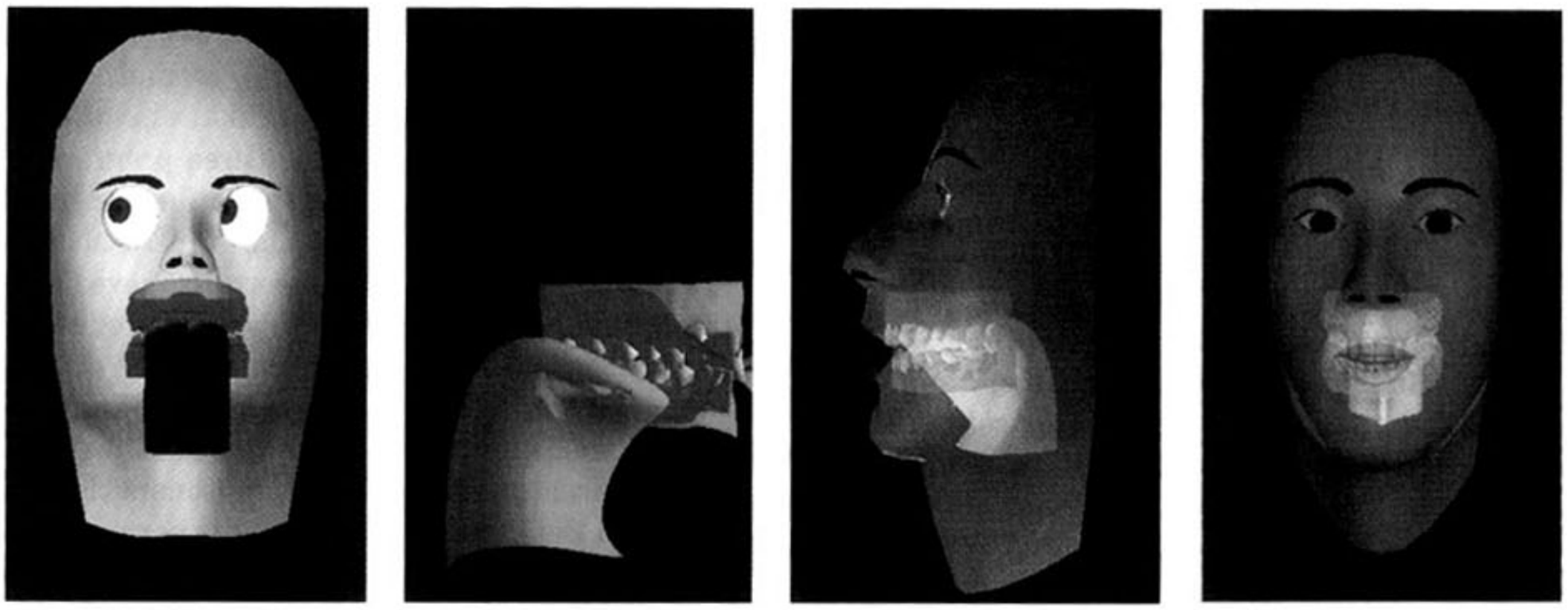
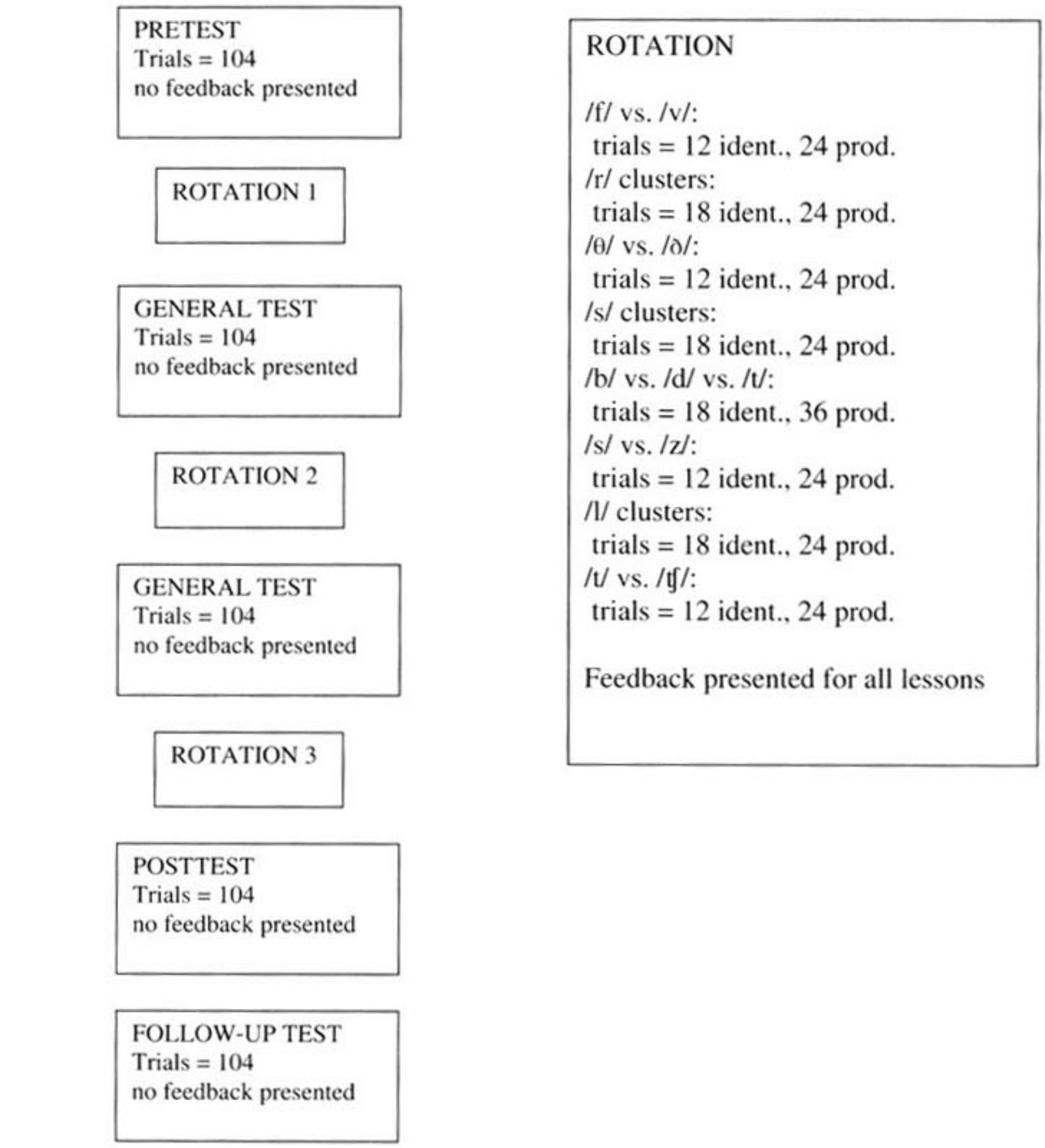


Figure 3. Sequence of procedures involved in testing and training.



their difficulties. For example, the experimenter noted that the /v/ sound was being produced with a nasal quality by a few students. This program was modified so that during the next training session of this cognate, Baldi would instruct the students to pinch their noses and produce the /v/ sound. This modification allowed the student to realize that nasality is not a feature of this sound. Although variations were made from one rotation to the next, the general format of the lessons remained constant from one day to the next. Each student completed a speech perception lesson and a speech production lesson during each day of training. The procedure for each training lesson is described below.

Perception

For speech perception, an identification task was given. For the voiceless versus voiced (/f/ vs. /v/, /s/ vs. /z/, and /θ/ vs. /ð/) and fricative versus affricate (/ʃ/ vs. /ʧ/) training lessons, stimuli consisted of 6 minimal pairs of words, contrasting voiceless and voiced phonemes (e.g., *fat* vs. *vat*, *shoe* vs. *chew*). For the /d/ versus /t/ distinction, /b/ was also included in this program; therefore, six minimal triplets were involved. For the consonant cluster programs, stimuli consisted of six minimal triplets of words, contrasting consonant cluster segments (e.g., for /r/ clusters: ***cr***own, ***fr***own, ***br***own).

Test Phase. During the test phase for all categories (voiced vs. voiceless, fricative vs. affricate, and consonant clusters), Baldi said an isolated word from a pair or triplet while written words were simultaneously presented on the computer monitor. During the first rotation, the experimenter noticed that some of the students were attending to the text rather than to Baldi. In an attempt to redirect the student's attention to Baldi, a delay between the speech and orthographic presentation was added for the second and third rotations. The experimenter's impression was that this modification

seemed to be successful. First, Baldi said an isolated word from the pair or triplet. After a delay of one second, the text of that minimal pair/triplet appeared on the computer monitor. At this time, the student was required to identify the word from the pair/triplet that he/she heard (see Figure 4). The judgment was made by dragging the computer mouse over the appropriate word in the pair/triplet and clicking the mouse. For voiced vs. voiceless pairs as well as the fricative vs. affricate pair, six words were tested in total (one word randomly selected from each pair). For the /b/ versus /d/ versus /t/ distinction, as well as for the distinction between consonant clusters, nine words were presented and tested in total (one word randomly selected from each triplet). Three alternatives were given in the consonant cluster test because it was easier to distinguish between minimal pairs for cluster categories compared to the other categories. Even with three alternatives, the students performed better in the consonant cluster identification tasks compared to the other categories. Feedback was given after each trial. A happy or sad face, representing a correct or incorrect response, appeared on the computer monitor to motivate the student to continue or try harder. This feedback also enabled the students to track their progress. The feedback with two alternatives indicated to the child the correct answer. With the three-alternative task, negative feedback did not give the correct answer; however, this uncertainty did not seem to cause any distress. The next trial was presented 1 s after feedback was given.

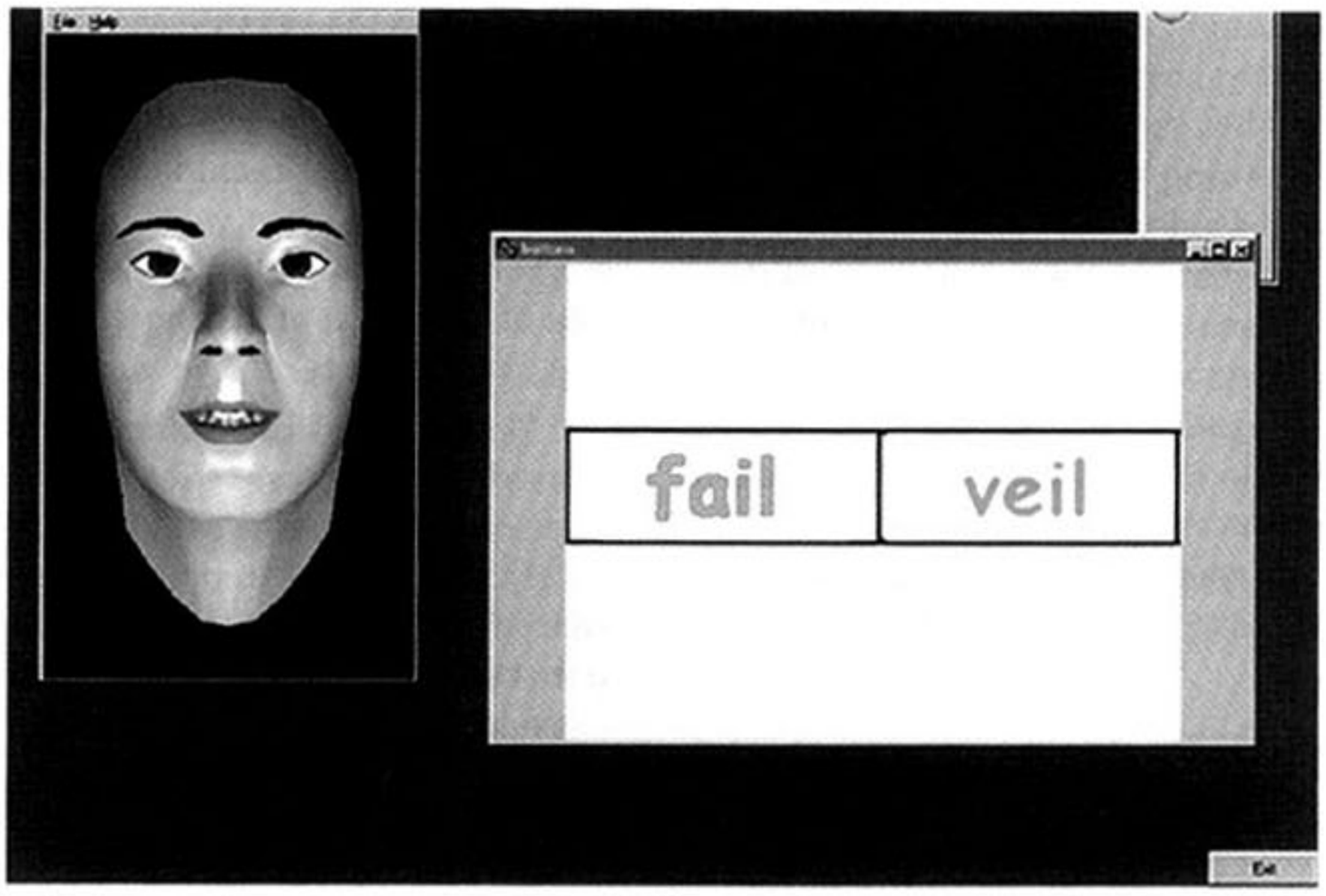
Tutoring Phase. After the six or nine trials were completed (depending on the training lesson), the student moved on to the tutoring phase. This tutoring phase was used to acquaint the students with Baldi's speech so that they could begin to understand how Baldi used his lips, tongue, and teeth to produce each word that was presented. No inside articulators were revealed at this time. First, Baldi told the student to "just listen and watch me. You won't need to click the buttons this time." Baldi said each word involved in the test phase while the accompanying text appeared on the monitor.

Next, Baldi showed the student how he produced each phoneme that was being trained (for example, for the voiceless/voiced distinction: "this is the /s/ sound, watch me and listen: [/s/ said slowly], this is the /z/ sound: [/z/ said slowly]" and for the consonant clusters: "this is the /r/ sound: [/r/ said slowly], this is the /br/ sound, watch closely and listen: [/br/ said slowly], this is the /kr/ sound: [/kr/ said slowly]" and so on). No inside articulatory views were presented during perception training. To mimic speech perception in the real world, we decided that presenting a front view of Baldi with opaque skin would be most appropriate.

Finally, Baldi said each word that was involved in the test phase one last time while the accompanying text of the word appeared on the monitor.

Test Phase. When the tutoring phase was complete, the student went through a second and final test phase. This was exactly the same as the test phase before tutoring. The student was required to identify which instance of the pair/triplet was heard by dragging the computer mouse over the text of the correct word in the minimal pair/triplet and clicking on it. Feedback was again given via a happy or sad face.

Figure 4. A full-screen view of a typical identification task (/f/ vs. /v/ distinction).



After completion of the speech perception training, the student went on to participate in a speech production training lesson.

Production

Test Phase. Baldi said an isolated word that included the target segment. After a tone, the student was instructed to repeat the word. Approximately 2 s after the tone, if a response could not be detected, Baldi asked the student to "please speak after the tone" and the tone replayed. Once a verbal response was detected, the computer captured this utterance in a sound file and it was logged. The production ability of the speaker (i.e., correct vs. incorrect) was determined by the voice recognition system in the CSLU toolkit. Unfortunately, the voice recognizer was not as accurate as we had hoped. Negative feedback was often given for correct responses as judged by the experimenter. This inaccurate feedback could hinder learning and was discouraging to the students so we modified the program by implementing a technique where the experimenter judged the student's response and input this recognition decision via the computer mouse, without the student being aware of the procedure. The experimenter's input to the computer after each response determined the feedback. The next trial was presented 1 s after feedback was given. Six trials of each target segment in the voiceless/voiced minimal pair (e.g., /f/ vs. /v/) or triplet (e.g., /b/ vs. /d/ vs. /t/) distinction and the fricative/affricate distinction (e.g., /ʃ/ vs. /ʧ/) were completed (12 trials for all pairs and 18 trials for the triplet), and placement of the target segment within the word was varied. Twelve trials of each consonant cluster were completed. Order of presentation was randomized, and on completion of the twelve trials, the student moved on to a tutoring phase.

Tutoring Phase. In the tutoring phase, the student was trained in how to produce the target segment. These instructions were composed from various sources (e.g., Ling, 1976; Massaro et al., in press). Different training methods were used to train certain categories. For example, supplementary features such as vocal cord vibration and turbulent airflow were used to visibly indicate the difference between voiceless and voiced contrasts (e.g., /f/ vs. /v/). A side view of Baldi with transparent skin was used during voiced versus voiceless training. This view was most effective for presenting the supplementary voicing features. For consonant cluster training, internal views of the oral cavity were important to show place features of the tongue during production. Slowing down Baldi's speech allowed us to emphasize the articulatory sequence involved in producing a consonant cluster. To teach the fricative versus affricate distinction, supplementary voicing features, internal articulatory views, and slowed down speech were all used in training. Four different internal views of the oral cavity were shown during consonant cluster and fricative versus affricate training: a view from the back of Baldi's head looking in, a sagittal view of Baldi's mouth alone (static and dynamic), a side view of Baldi's whole face where his skin was transparent, and a frontal view of Baldi's face with transparent skin. Each view gave the student a unique perspective on the activity, which took place during production (see Figure 2). We expected these multiple views to facilitate learning and to anticipate individual preferences for different views.

During all training lessons, the student was instructed in how to produce the segments being trained (e.g., /f/ and /v/ for a voiceless vs. voiced contrast; /s/, /sm/, /st/, /sl/, and so on for consonant cluster training; /ʃ/ vs. /ʧ/ for a fricative vs. affricate contrast). The students were also required to produce the segment in isolation as well as in words and were given the ability to hear their productions of certain words by a playback feature during the tutoring of the consonant clusters. No feedback was given during the training stage, but "good job" cartoons were given as reinforcement. The appendix gives a more detailed explanation of the processes involved in each type of training.

The tutoring phase for all lessons ended with Baldi saying, "Okay, now let's see what you've learned."

Test Phase. After the tutoring phase was completed, each student performed the repetition phase once again with feedback. This was identical to the first test phase. Six trials of each segment being tested were presented randomly, and placement of the target segment in the word varied. Baldi said a word and the student was required to say that word back to him.

Posttest

One "rotation" was defined by the completion of all eight training lessons (see Figure 3). After each rotation, the student was given the general test of 104 words. This test was the same as the one given at pretest. The general tests were used as a measure of the degree to which the production abilities of each student changed from pretest to posttest. Three rotations of the eight lessons (6 hours of training), as well as a pretest, two general tests, and a posttest, were completed (see Figure 3).

Follow-Up Test

A follow-up test was given 6 weeks after training ended. This test was exactly the same as the pretest and the posttest (a general test of 104 words). This test was used to see how production ability was retained once the training lessons ended.

Ratings

The productions of the participants were evaluated in two rating experiments, using different groups of judges. After all of the sound files were collected and properly labeled, nine judges recruited from the psychology student pool at the University of California, Santa Cruz, participated in a rating experiment. Words, students' utterances, and pretest/posttest were randomized and presented auditorily one at a time. The judges were asked to rate the intelligibility of a word against the target text, which was simultaneously presented on the computer monitor. Intelligibility was rated on a scale from 1 to 5 (1: unintelligible, 2: ambiguous, 3: distinguishable, 4: unambiguous, 5: good/clear pronunciation). The judges' ratings were later linearly transformed to a scale ranging from 0 to 1. In all cases, the raters had no knowledge of details of the experiment or the testing status of each word production.

Fifty undergraduate linguistics students at the University of California, Santa Cruz, were selected as new judges to rate the intelligibility of the students' speech from pretest to posttest to follow-up. These undergraduate students were enrolled in a phonetics class at the time. We believed that the training these students received in their class would be valuable to our study.

In contrast to the first assessment, where the auditory and written stimuli were presented simultaneously, the new assessment first presented the auditory stimulus with the written stimulus immediately following. Twenty minutes of class time were set aside for the presentation of these stimuli to the class, and the judges' ratings were recorded on Scantron cards (Scantron Corp., Irvine, CA). Time constraints prevented us from including all of the trained categories, so we included the categories that received the lowest pretest ratings by the initial judges (e.g., /s/ vs. /z/, /s/ clusters, and /ʃ/ vs. /ʧ/). Both groups of judges were necessary in our study, for the first group did not provide follow-up ratings and the second group did not provide ratings for all cognate pairs involved in training. Intelligibility ratings provided by the two groups also allowed us to cross-check the judges' results for reliability.

Results

Perception

A two-way repeated analysis of variance (ANOVA) was carried out on the proportion of correct identifications. Student served as the random source of variance ($N = 7$). The eight training categories (/f/ vs. /v/, /s/ vs. /z/, /l/ final clusters, etc.) and test (pretest vs. posttest) were the independent variables.

Figure 5 shows the proportion of correct identifications for each category illustrated by order of presentation to the student during any given training rotation. The proportion of correct identifications increased from .64 at pretest to .86 at posttest, $F(1, 6) = 252.43$, $p < .001$. Performance varied depending on which training category was involved, $F(7, 42) = 6.03$, $p < .001$, and there was an interaction between training category and test, $F(7, 42) = 3.09$, $p < .001$. This interaction reflects the finding that the most difficult items showed the largest improvement. For instance, at pretest, the /l/ final clusters were identified almost perfectly in both the pretest and posttest. The /s/ versus /z/ and the /ʃ/ versus /ʧ/ categories were the most difficult to identify at pretest

Figure 5. Proportion of correct identifications (and standard error bars) during pretest and posttest for each of the eight training categories. The results are graphed from left to right by order of presentation during each training rotation.

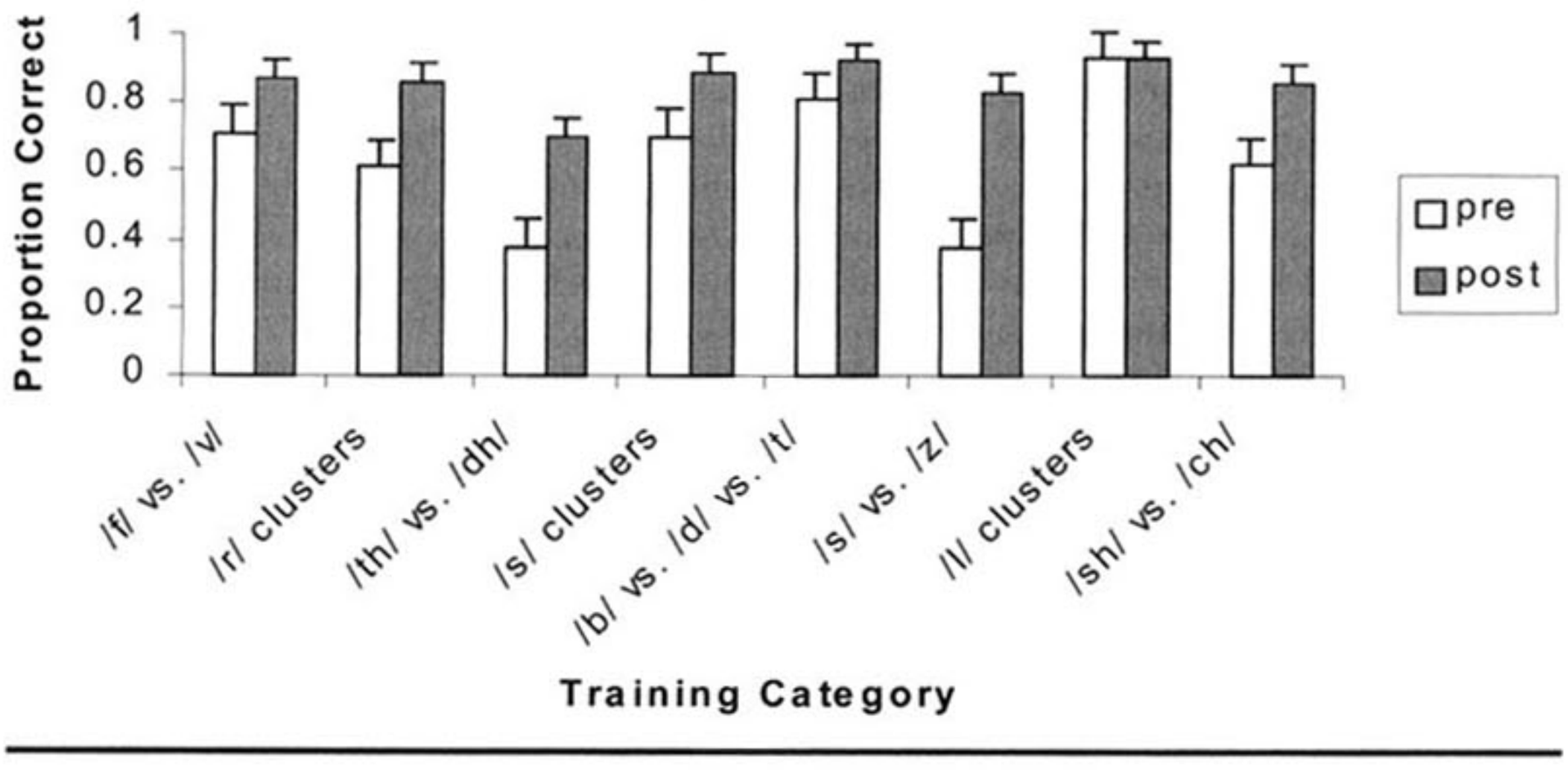
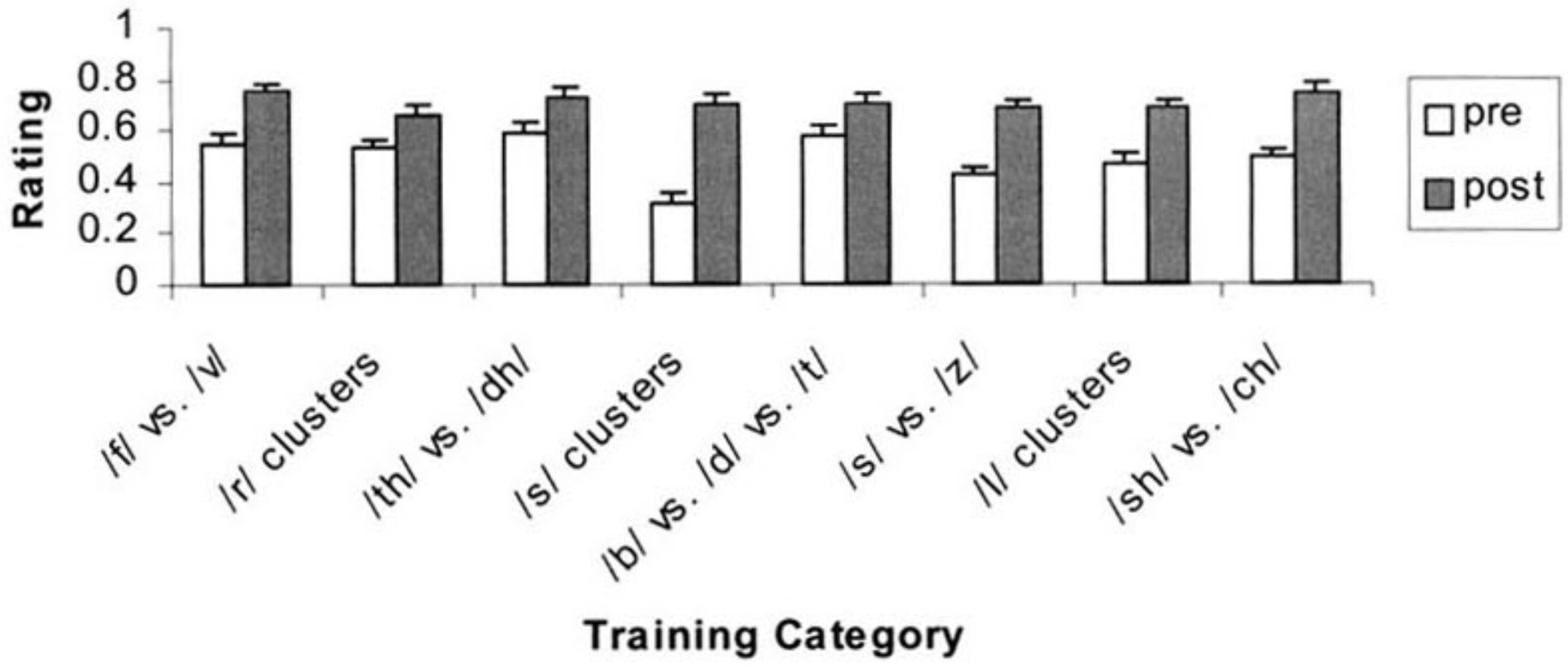


Figure 6. Intelligibility ratings (and standard error bars) of the pretest and posttest word productions for each of the eight training categories. The results are graphed from left to right by order of presentation during each training rotation.



(/s/ vs. /z/ = .38; /ʃ/ vs. /ʧ/ = .38) and showed the most improvement at posttest (/s/ vs. /z/ = .83, .45 improvement; /ʃ/ vs. /ʧ/ = .70, .32 improvement).

Production

Ratings (intelligibility of the auditory stimulus to the target text on a scale from 0 to 1) of the nine psychology student judges were used as a measure of production accuracy. To determine how well the students improved in their speech production from pretest to posttest, a three-way repeated analysis of variance (ANOVA) was performed on the judges' ratings of the students' productions. Student ($N = 7$), test (pretest vs. posttest), and category (/f/ vs. /v/, /s/ vs. /z/, /l/ final clusters, etc.) were the independent variables. Judge ($N = 9$) served as the random source of variance in this design.

Figure 6 gives the pretest and posttest production ratings for each of the eight training categories. Production ratings showed a .21 increase on the 0 to 1 intelligibility scale from pretest to posttest, $F(1, 8) = 67.93$, $p < .001$. To determine the reliability of the judges' ratings, we computed the range of differences in the pretest and posttest scores. These differences varied between .08 and .29 across the nine judges, with seven of the nine judges falling within the .20 to .29 range, showing that there was good reliability across the different judges.

To determine whether each student individually benefited from the program, a separate analysis was performed on each student's results. As can be seen in Table 3, a statistically significant increase in ratings from pretest to posttest was observed for each of the 7 students.

As is shown in Figure 6, performance also varied depending on which training category was involved, $F(7, 56) = 8.31$, $p < .001$, and there was an interaction between test and category, $F(7, 56) = 20.71$, $p < .001$. Although all categories showed an improvement in ratings from pretest to posttest, the categories that were rated lowest at pretest showed the greatest improvement at posttest.

A second analysis was performed to assess the effectiveness of the differentiating information involved in production training: Supplementary voicing features including vocal cord vibration and turbulent airflow, inside articulatory views from multiple angles with slowed-down speech, and a combination of both techniques were the information conditions. A three-way repeated ANOVA was performed on the production ratings (intelligibility on a scale of 0 to 1). Judge ($N = 9$) was the random source of variance. Information condition (voicing features vs. visible articulation vs. both), test (pretest vs. posttest), and student ($N = 7$) were the independent measures.

The production rating increased from pretest to posttest for each information condition. There was a significant rating increase from pretest to posttest, $F(1, 8) = 81.62$, $p < .001$. Production ratings differed depending

Table 3. Change in ratings for each student from pretest to posttest.

Student	Pretest	Posttest	Significance	
			$F(1, 8)$	p
S1	.53	.66	7.42	<.05
S2	.34	.60	46.09	<.001
S3	.43	.68	17.237	<.001
S4	.64	.85	78.38	<.001
S5	.57	.83	22.23	<.001
S6	.46	.72	142.764	<.001
S7	.48	.59	8.095	<.05

on which information condition was used, $F(2, 16) = 9.59$, $p < .002$, and a significant interaction between information condition and test was also shown, $F(2, 16) = 6.19$, $p < .05$. The pretest and posttest ratings were .44 and .67 for the visible articulators, .54 and .72 for the supplementary voicing features, and .49 and .74 for both conditions.

Because information condition was confounded with category involved, a separate analysis was carried out for each information condition individually. In each analysis, judge ($N = 9$) was the random source of variance. Test (pretest vs. posttest) and student ($N = 7$) were the independent variables. All information conditions showed a significant increase in ratings from pretest to posttest, revealing that each of the information conditions was successful; voicing features: $F(1, 8) = 41.22$, $p < .001$; visible articulation: $F(1, 8) = 66.26$, $p < .001$; both: $F(1, 8) = 90.31$, $p < .001$.

To test whether learning generalized to words not involved in training, a three-way repeated ANOVA was carried out on the ratings (intelligibility on a scale of 0 to 1). Judge ($N = 9$) served as the random source of variance. Student ($N = 7$), pretest vs. posttest, and test words (trained vs. untrained) were the independent variables. There was a significant increase in production ratings from pretest to posttest, $F(1, 8) = 10.09$, $p < .01$, with the untrained words showing a greater change in ratings (.45 to .69; .24 change) than the trained words (.53 to .71; .18 change), $F(1, 8) = 14.82$, $p < .005$. An increase in production ratings for the untrained words from pretest to posttest shows that the training generalized to words not in the training lessons.

We assessed retention of training by comparing the production ratings by the 50 linguistics students for the words at pretest, posttest, and follow-up. A three-way repeated ANOVA was carried out. Judge ($N = 50$) was the random source of variance. Student ($N = 7$), test (pretest vs. posttest vs. follow-up), and category (/s/ vs. /z/, /s/ clusters, /ʃ/ vs. /ʧ/) were the independent variables, and rating (intelligibility on a scale of 0 to 1) was the dependent variable. There was a significant main effect for test, $F(2, 98) = 204.27$, $p < .001$. To determine whether the change in ratings was significant from pretest to posttest and also from posttest to follow-up, three separate analyses were performed. In the first analysis, pretest vs. posttest ratings were compared, revealing a significant .21 positive change in ratings from pretest to posttest, $F(1, 49) = 393.43$, $p < .001$. This positive change was identical to that observed with the nine psychology student judges. The ratings from the phonetics students also appeared to be reliable in that the positive change from pretest to posttest varied between .01 and .36 across the 50 judges, with 41 out of the 50 judges falling within the .10 to .30 range.

In the second analysis, posttest versus follow-up ratings were analyzed, revealing a significant –.08 negative change in ratings from posttest to follow-up, $F(1, 49) = 54.75$, $p < .001$. This difference varied between .01 and –.21 across the 50 judges, with 41 out of the 50 judges falling within the 0 to –.20 range. In the third analysis, the ratings for the follow-up productions averaged .13 better than the pretest, which was also significant, $F(1, 49) = 162.83$, $p < .001$, showing that significant learning was retained even 6 weeks after training was completed. This difference varied between –.02 and .28 across the 50 judges, with 41 out of the 50 judges falling within the 0 to .20 range. The statistical significance and the range of the ratings in all cases showed that the ratings were reliable.

Students' Reactions and Evaluation of the Tutoring

Although there were individual differences in aided hearing thresholds, attitude, and cognitive level, the training program helped all of the children (see Tables 1 and 3). Student 1 was cooperative but did not like working with Baldi. She appeared to have had physiological difficulty producing some of the sounds; she was small for her age and her father reported that she had a "short tongue." Student 2 was often not as cooperative as the others during training but he gained certain important skills from Baldi (e.g., voicing instead of nasalizing certain sounds). For the sounds that he knew he had learned, he was confident and impressive. Student 3 loved working with Baldi. He is not a very social child and doesn't usually cooperate in class but Baldi was his favorite part of the day. He was involved and motivated. Student 4 was apprehensive at first but she became more comfortable. Although her speech was already quite intelligible, she could hear and feel an improvement in her own speech. Student 5 was very cooperative and thought Baldi was funny. She recently received a cochlear implant and had gained a lot of confidence in her speech. She thought the program was too easy and didn't think she needed so much practice. By the end, she knew that she improved but perhaps felt it was only marginal. Student 6 was a very attentive and cooperative student. She was always asking questions and wanted to learn as much as she could. She indicated that the program was a great teaching tool and she quickly noticed the benefits of the program. She definitely could hear/feel an improvement in her own speech and would have liked to continue with this program for a longer duration. Student 7 was not as interested in working with Baldi as she was in talking to the second author, who often had to redirect her focus and use rewards to keep her motivated. She was receptive to these instructions and rewards and knew that it was good practice for her.

Discussion

The main goal of this study was to implement Baldi as a language tutor for speech perception and production for individuals with hearing loss. The students' ability to perceive and produce words involving the trained segments did change from pretest to posttest. A second analysis revealed an improvement in production ratings 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). Although training method was confounded with category, an analysis of pretest versus posttest ratings revealed each method to be successful.

Our method of training is similar in some respects to electropalatography (EPG), which has been considered useful in clinical settings because it provides direct visual feedback (in the form of a computer display) on the contact between the tongue and the palate during speech production. The student wears a custom-fitted artificial palate embedded with electrodes, and the clinician may wear one as well. The clinician illustrates a target pattern, and the student attempts to match it. For instance, the student may be presented with a typical contact pattern for /s/, with much contact at the sides of the palate and a narrow constriction toward the front of the palate. Certain speech pathologies result in /s/ being produced as a pharyngeal fricative. The pharyngeal fricative would show up on the screen as a lack of contact on the hard palate. The clinician can then teach the patient how to achieve the target pattern. Dent, Gibbon, and Hardcastle (1995) provide a case study where EPG treatment improved the production of lingual stops and fricatives in a patient who had undergone pharyngoplasty.

EPG treatment has also proved to be useful in teaching children who are deaf to produce normal-sounding lingual consonants (e.g., Crawford, 1995; Dagenais, Critz-Crosby, Fletcher, & McCutcheon, 1994; Fletcher, Dagenais, & Critz-Crosby, 1991). Although the visual feedback from the EPG is deemed to be extremely important to the significant improvement in production, there have been very few systematic evaluations of its effectiveness. In the current study, however, our method appears to have been more successful, with a 21% improvement overall. Dagenais (1992) trained four different segments (e.g., alveolar stops, velar stops, alveolar sibilants, and palatal sibilants) and found an average 8% improvement in linguapalatal contact across 4 participants from pretest to 6 months after commencement of training (71% at pretest vs. 79% at 6 months). Although Dagenais provided many more hours of training (about 36 hours), we found a 21% improvement in production ratings after just approximately 6 hours of training.

Dagenais (1992) also noted that electropalatography methods limited the abilities of the trainees with hearing loss to generalize to novel situations because of the limited tactile feedback that participants received during training. Our untrained words actually showed a somewhat greater improvement from pretest to posttest than did trained words (.24 change and .18 change, respectively). The small difference probably only reflects that the untrained words received lower initial production ratings than did trained words. Learning with our training method therefore appears to generalize to words outside of the training lessons.

The present findings suggest that Baldi is an effective tutor for speech training students with hearing loss. 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 desired, 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 study took place. Baldi appears to offer unique features that can be added to the arsenal of speech-language pathologists.

Ratings of the posttest productions were significantly higher than pretest ratings, indicating significant learning. Given that we did not have a control group, it is always possible that some of this learning occurred independently of our program or was simply based on routine practice. However, the results provided some evidence that at least some of the improvement must be due to our program. Follow-up ratings 6 weeks after our training was complete were significantly lower than posttest 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 directly test the usefulness of Baldi to their treatment methods and focus on which specific training regimens are most effective for particular contrasts.

Acknowledgments

The research and writing of this article were supported by the National Science Foundation (Grant No. CDA-9726363, Grant No. BCS-9905176, Grant No. IIS-0086107), the Public Health Service (Grant No. PHS R01 DC00236), and the University of California, Santa Cruz.

References

Bernstein, L. E., Demorest, M. E., & Tucker, P. E. (2000). Speech perception without hearing. *Perception & Psychophysics, 62*, 233–252.

Black, A. W., & Taylor, P. A. (1997). The Festival Speech

Synthesis System: System Documentation," HCRC/TR-83, v1.1. http://www.cstr.ed.ac.uk/projects/festival/.

Bosseler, A., & Massaro, D. W. (2003). Development and evaluation of a computer-animated tutor for vocabulary and language learning for children with autism. *Journal of Autism and Developmental Disorders, 33,* 653–672.

Cohen, M. M., Beskow, J., & Massaro, D. W. (1998, December). Recent developments in facial animation: An inside view. *Proceedings of Auditory Visual Speech Perception '98* (pp. 201–206). Terrigal-Sydney, Australia: Causal Productions.

Cornett, R. O. (1988). Cued speech, manual complement to lipreading, for visual reception of spoken language: Principles, practice and prospects for automation. *Acta Oto–Rhino–Laryngologica Belgica, 42,* 375–384.

Crawford, R. (1995). Teaching voiced velar stops to profoundly deaf children using EPG: Two case studies. *Clinical Linguistics & Phonetics, 9,* 255–269.

Dagenais, P. (1992). Speech training with glossometry and palatometry for profoundly hearing-impaired children. *The Volta Review, 94,* 261–282.

Dagenais, P., Critz-Crosby, P., Fletcher, S., & McCutcheon, M. (1994). Comparing abilities of children with profound hearing impairments to learn consonants using electropalatography or traditional aural–oral techniques. *Journal of Speech and Hearing Research, 37,* 687–699.

Denes, P. B., & Pinson, E. N. (1963). *The speech chain: The physics and biology of spoken language.* New York: Bell Telephone Laboratories.

Dent, H., Gibbon, F., & Hardcastle, B. (1995). The application of electropalatography (EPG) to the remediation of speech disorders in school-aged children and young adults. *European Journal of Disorders of Communication, 30,* 264–277.

Dodd, B., & Campbell, R. (Eds.). (1987). *Hearing by eye: The psychology of lip-reading.* London, England: Erlbaum.

Dodd, B., McIntosh, B., & Woodhouse, L. (1998). Early lipreading ability and speech and language development of hearing-impaired preschoolers. In R. Campbell, B. Dodd, & D. Burnham (Eds), *Hearing by eye II: Advances in the psychology of speechreading and auditory visual speech* (pp. 229–242). East Sussex, U.K.: Psychology Press.

Fletcher, S. G., Dagenais, P. A., & Critz-Crosby, P. (1991). Teaching consonants to profoundly hearing-impaired speakers using palatometry. *Journal of Speech and Hearing Research, 34,* 929–942.

Gallaudet Research Institute. (1999–2000). *Regional and national summary, annual survey of deaf and hard of hearing children and youth.* Washington, DC: Author.

Gold, T. G. (1980). Speech production in hearing-impaired children. *Journal of Communication Disorders, 13,* 397–418.

Jusczyk, P. W. (1997). *The discovery of spoken language (language, speech and communication).* Cambridge, MA: MIT Press.

Kirk, K. I., Pisoni, D. B., & Miyamoto, R. C. (1997). Effects of stimulus variability on speech perception in listeners with hearing impairment. *Journal of Speech, Language, and Hearing Research, 40,* 1395–1405.

Ladefoged, P. (2001). *A course in phonetics* (4th ed.). Orlando, FL: Harcourt College Publishers.

Levitt, H., McGarr, N., & Geffner, D. (1987). Development of language and communication skills in hearing-impaired children: Introduction. *ASHA Monographs, 26,* 1–8.

Leybaert, J., Alegria, J., Hage, C., & Charlier, B. (1998). The effect of exposure to phonetically augmented lip-speech in the prelingual deaf. In R. Campbell, B. Dodd, & D. Burnham (Eds.), *Hearing by eye II: Advances in the psychology of speechreading and auditory visual speech* (pp. 283–301). East Sussex, U.K.: Psychology Press.

Ling, D. (1976). *Speech and the hearing-impaired child: Theory and practice.* Washington, DC: Alexander Graham Bell Association for the Deaf and Hard of Hearing.

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

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

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.

Massaro, D. W., Cohen, M. M., Beskow, J., & Cole, R. A. (2000). Developing and evaluating conversational agents. In J. Cassell, J. Sullivan, S. Prevost, & E. Churchill (Eds.), *Embodied conversational agents* (pp. 286–318). Cambridge, MA: MIT Press.

Massaro, D. W., Cohen, M. M., Tabain, M., Beskow, J., & Clark, R. (in press). Animated speech: Research progress and applications. In E. Vatikiotis-Bateson, P. Perrier, & G. Bailly (Ed.), *Advances in audio-visual speech processing.* Cambridge, MA: MIT Press.

Massaro, D. W., & Light, J. (in press). Improving the vocabulary of children with hearing loss. *Volta Review.*

McGurk, H. (1981). Listening with eye and ear (paper discussion). In T. Myers, J. Laver, & J. Anderson (Eds.), *The cognitive representation of speech* (pp. 336–337). Amsterdam: North-Holland.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264,* 746–748.

Morais, J., & Kolinsky, R. (1994). Perception and awareness in phonological processing: The case of the phoneme. *Cognition, 50,* 287–297.

National Association of State Directors of Special Education. (1992). *Deaf and hard of hearing students: Educational guidelines.* Alexandria, VA: Author.

Osberger, M. (1987). Training effects on vowel production by two profoundly hearing-impaired speakers. *Journal of Speech and Hearing Research, 30,* 241–251.

Sanders, D. M. (1988). *Teaching deaf children: Techniques and methods.* Boston: Little, Brown.

Scherer, K. (1986). Vocal affect expression: A review and model for future research. *Psychological Bulletin, 99,* 143–165.

Smeele, P. M. T., Massaro, D. W., Cohen, M. M., & Sittig, A. C. (1998). Laterality in visual speech perception. *Journal of Experimental Psychology: Human Perception and Performance, 24,* 1232–1242.

Svirsky, M. A., Robbins, A. M., Kirk, K. I., Pisoni, D. B., & Miyamoto, R. T. (2000). Language development in profoundly deaf children with cochlear implants. *Psychological Science, 11*, 153–158.

Torgesen, J. K., Wagner, R. K., Rashotte, C. A., Lindamood, P., Rose, E., Conway, T., & 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.

U.S. Department of Health and Human Services. (2002). Retrieved March 9, 2004, from http://www.cdc.gov/ncbddd/dd/ddhi.htm

Received February 10, 2003

Accepted July 21, 2003

DOI: 10.1044/1092-4388(2004/025)

Contact author: Dominic W. Massaro, PhD, Department of Psychology, University of California, Santa Cruz, CA 95064. E-mail: massaro@fuzzy.ucsc.edu

Appendix (p. 1 of 2). Tutoring phase.

Voiced Versus Voiceless Distinction

In all of the tutoring, the experimenter was present but did not provide any additional instruction other than repeating what Baldi said if the child did not understand (mostly because of their limited hearing and Baldi's synthetic speech). For the voiceless versus voiced distinctions (e.g., /f/ vs. /v/, /s/ vs. /z/), Baldi first showed the student how to produce the target segments. He asked the student whether he/she could hear the difference between the two target sounds in the minimal pair (e.g., "Can you hear the difference between the /f/ in *fat* and the /v/ in *vat*?"). A 2-s pause allowed the student to respond before Baldi continued. Baldi told the student to watch carefully as he produced these two segments again and then went on to give the student verbal instructions on how he/she should produce the target segments (e.g., where to position the tongue with respect to the teeth, the shape of the tongue and lips, etc.). It should be noted that the instructions are the same for both segments in terms of tongue, teeth, and lip place features. Baldi then produced the voiced segment while the inside articulators were revealed.

Supplementary features such as vocal cord vibration and turbulent airflow were visible when Baldi produced the segment to enhance awareness about articulatory properties. A side view was the only view used during this type of training, for this was the best way to emphasize the supplementary features. The student was then instructed to produce the voiced target segment. No feedback was given at this time. During Baldi's production of the voiced segment, vocal cord vibration was shown. Baldi asked the student if he/she saw his throat vibrate. After a short pause of 2 s, which gave the student a chance to respond, Baldi explained that this segment was voiced, that voicing was made in his throat, and that this is what caused his throat to vibrate. The student was instructed to watch Baldi as he produced the voiced segment and to pay attention to his throat. The student was then asked to "put your hand on your throat. Keep it there and make the X sound." This enabled the student to feel for him/herself whether or not he/she was using his/her throat to make this sound. Baldi told the student that he/she should feel a vibration in his/her own throat. He explained that if he/she didn't feel a vibration, he/she was not making this sound correctly, but not to worry because he/she would have much opportunity to practice and improve.

The same procedure was carried out for the voiceless counterpart. For the voiceless segment, no vocal cord vibration was shown during Baldi's production. Baldi asked the student if he/she saw his throat vibrate. After a short pause of 2 s, which gave the student a chance to respond, Baldi explained that the reason why no vibration occurred was because this sound was voiceless and when making voiceless sounds, you do not use your vocal cords. Baldi let the student know that voicing was one feature that distinguished between the two sounds being trained. The student was then instructed to put his/her hand on his/her throat and produce this sound. This enabled the student to feel no vibration and to understand the difference between voiced and voiceless segments in his/her own speech.

Baldi then went through the same procedure for turbulent airflow, showing the difference between the varying degrees of air that escape the mouth for voiced versus voiceless segments (e.g., a large degree of expulsion for voiceless segments and almost no expulsion for voiced segments). He asked the student to "put your hand in front of your mouth. Keep it there and make the X sound." Having a hand in front of his/her mouth during production allowed the student to feel the air hit his/her hand in varying degrees, which allowed the student to better understand the difference between the production of the two sounds in his/her own speech.

Next, Baldi showed the student how to produce various words involving the voiced segment from the pair. After Baldi produced a word involving the target phoneme, he gave the student helpful tips about tongue positioning and so on, and then he asked the student to repeat this word to him. After the student made an effort to produce this word, he/she was presented with a cartoon on which was written one of various reinforcing statements, including "good job," "way to go," "awesome," and so on, to encourage the student to keep trying, regardless of whether or not he/she had produced the segment correctly. The same procedure was carried out for the voiceless counterpart.

Appendix (p. 2 of 2). Tutoring phase.

Consonant Cluster Training

For consonant cluster training, the student was first instructed to watch Baldi as he produced the target segment in the cluster (e.g., /l/, /r/, /s/). This segment was first produced in isolation, from an inside view, at 30% of the normal speech rate. The student was then asked to try and produce this sound after the beep. Once a response could be detected by the voice recognizer, Baldi gave the student helpful tips about tongue positioning and so on for production of this sound (e.g., for production of /l/: "Remember to point your tongue, put the tip of your tongue behind your upper front teeth, and raise the sides of your tongue.").

Next, Baldi asked the student to say a word after the beep that didn't involve a cluster but did involve the target segment (e.g., *ball*). When a response was detected, the student was instructed to produce the sound of a letter that could easily be added to the previous word to then make a new word involving a consonant cluster (e.g., "now make the /d/ sound"). The student was then encouraged to slowly say the word that was just formed (e.g., *bald*). Baldi explained what a consonant cluster was and that to produce clusters "you have to change the position of your tongue very fast." Baldi allowed the student to view his production from an inside view, slowed his speech down to 30% of the normal rate, and told the student to watch his tongue carefully as it changed from one position to the next while he produced some clusters. While revealing his internal oral cavity, Baldi produced several consonant clusters involving the target segment (e.g., /lm/, /lf/, /lt/). After Baldi produced each cluster, he asked the student to repeat that cluster. This allowed the student to perceive Baldi's speech before producing the word on his/her own.

Baldi told the student that together they were going to have some fun. He asked the student to say a consonant cluster word involving the target segment after the tone. Once the student produced the word, his/her voice was recorded and played back to him/her. This playback feature allowed the student to hear his/her own articulation and identify his/her own mistakes. If the student could not detect all of the segments in the word he/she just said, the student knew which segments he/she needed to work on. Finally, Baldi asked the student to repeat after him as he produced various consonant cluster words involving the target segment. The production of these words was presented from multiple views, allowing the student to understand the articulation by viewing the tongue movements from several angles.

Fricative Versus Affricate Distinction

For the fricative versus affricate training lesson, the difference between the two sounds (/ʃ/ and /ʧ/) was contrasted using a combination of the previous techniques. Baldi first showed the student how to produce the target segments. He asked the student whether he/she could hear the difference between the two target segments within the minimal pair (e.g., "Can you hear the difference between the /ʃ/ in *ship* and the /ʧ/ in *chip*?"). A 2-s pause allowed the student to respond before Baldi continued. Baldi told the student to watch carefully as he produced the two segments again. The student was then asked to produce each segment after the beep. Once a response could be detected, Baldi gave the student verbal instructions on how to produce the /ʃ/ segment (e.g., where to position the tongue with respect to the teeth, the shape of the tongue and lips). The student was instructed to repeat after Baldi as he produced a word with the /ʃ/ segment. The student was reinforced with a cartoon to encourage him/her to continue, but no feedback was given about his/her production. An equation appeared on the screen while Baldi explained that the /ʧ/segment was actually a combination of two segments (/t/ and /ʃ/). He instructed the student in how to produce the /ʧ/ segment by starting off making a /t/ segment but forcing out a /ʃ/ segment really hard. He told the student that if he/she repeated the phrase "meet ship" very quickly, it would segment like "me chip." Baldi told the student to try and say this phrase very fast after the tone. Once a response was detected, the student was instructed to put his/her hand in front of his/her mouth and keep it there. Having a hand in front of his/her mouth during production allowed the student to feel a greater burst of air for the production of /ʧ/ than for the production of /ʃ/. Finally, Baldi showed the student how to produce various words that included the trained segments (e.g., *share* vs. *chair*, *shoe* vs. *chew*). The student was asked to repeat after Baldi once he/she heard the tone. Baldi's speech was slowed down to 30% of the normal rate so that the distinction between the two segments could easily be realized. This repetition phase lasted for six trials (three words involving each segment), and the words were produced from various views while the inside articulators were shown. This allowed the student to understand the articulation by viewing the tongue movement from several angles.