Massaro, D. W., Liu, Y., Chen, T. H., & Perfetti, C. A. (2006). A Multilingual Embodied Conversational Agent for Tutoring Speech and Language Learning. Proceedings of the Ninth International Conference on Spoken Language Processing (Interspeech 2006 - ICSLP, September, Pittsburgh, PA), 825-828.Universität Bonn, Bonn, Germany.

A Multilingual Embodied Conversational Agent for Tutoring Speech and Language Learning

Dominic W. Massaro[1], *Ying Liu*[2], *Trevor H. Chen*[1], *and Charles Perfetti*[2]

[1]Department of Psychology, Perceptual Science Laboratory
University of California, Santa Cruz
Santa Cruz, CA 95060 U.S.A.
http://mambo.ucsc.edu/dwm
[2]Learning Research and Development Center
University of Pittsburgh, Pittsburgh, PA15260
Massaro@fuzzy.ucsc.edu, liuying+@pitt.edu, t8chen@ucsc.edu, perfetti@pitt.edu

Abstract

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

Based on this research and technology, we have implemented computer-assisted speech and language tutors for children with language challenges and for all persons learning a second language. Our language-training program utilizes Baldi (or his likeness) as the conversational agent, who guides students through a variety of exercises designed to teach vocabulary and grammar, to improve speech articulation, and to develop linguistic and phonological awareness. We have also implemented multilingual agents, using a client/server architecture system. This system has been used to develop Bao, a Mandarin talker, which has been used in an initial training study for college students learning Mandarin as a new language. The results address the potential for using visible speech technology and pedagogy in language learning of both similar segments in the two languages and new speech segments in the new language. Although visible speech did not facilitate pronunciation learning relative to just auditory speech, we expect that a more prolonged training period would show an advantage of visible speech.

Some of the advantages of the Baldi pedagogy and technology include the popularity and proven effectiveness of computers and embodied conversational agents, the perpetual availability of the program, and individualized instruction. The science and technology of Baldi holds great promise in language learning, dialog, human-machine interaction, education, and edutainment.

Index Terms: Speech Animation, Language Learning

1. Introduction

In conjunction with its speech science research, the Perceptual Science Laboratory (PSL-UCSC) has aimed to create embodied computer-animated agents that produce accurate auditory and visible speech, as well as realistic facial expressions, emotions, and gestures. The invention of such agents has a tremendous potential to benefit virtually all individuals in learning speech and language. Our talking head, Baldi®[1], has been used as a vocabulary tutor for children with language challenges, including hard of hearing and autistic children. Baldi has also been used for speech training of both hard of hearing children and adults learning a second language. The animated characters that we are developing have also been used to train autistic children to "read" visible speech and to recognize emotions in the face and voice [1].

1.1 Facial Animation and Visible Speech Synthesis

There have been several approaches to facial animation, including muscle models to simulate the muscle and tissues during talking [2], performance-based synthesis that tracks a live talker [3], and image synthesis, which joins together images of a real speaker [4][5]. The facial animation used in the current applications, however, is a descendant of Parke's software and his particular 3-D talking head [6]. Modifications have included increased resolution of the underlying wireframe model; additional and modified control parameters that have been tuned to agree with measurements of natural talkers; a realistic tongue trained on electropalatography and ultra-sound data; a tested coarticulation model; paralinguistic information and affect in the face; alignment with either natural speech or text-to-speech synthesis; and real-time bimodal (auditory/visual) synthesis on a commodity personal computer. Most of the parameters move vertices (and the polygons formed from these vertices) on the face by geometric functions such as rotation (e.g. jaw rotation) or translation of the vertices in one or more dimensions (e.g., lower and upper lip height, mouth widening). Other parameters work by scaling and interpolating different face subareas. Many of the face shape parameters – such as cheek, neck, or forehead

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

shape, and also some affect parameters such as smiling – use interpolation.

Phonemes are used as the unit of visible speech synthesis. Any utterance can be represented as a string of successive phonemes, and each phoneme is represented as a set of target values for the control parameters such as jaw rotation, mouth width, etc. Because speech production is a continuous process involving movements of different articulators (e.g., tongue, lips, jaw) having both mass and inertia, phoneme utterances are influenced by the context in which they occur. This so-called coarticulation is implemented in the synthesis by dominance functions, which determine independently for each control parameter over time how much weight its target value carries against those of neighboring segments [7]. In a test of several coarticulation models, Beskow [8] found that our model gave the best fit to observed articulatory data.

We evaluate the accuracy and intelligibility of Baldi's synthetic visible speech by perceptual recognition tests given to human observers [9]. These experiments are aimed at evaluating the speech intelligibility of the visible speech synthesis relative to natural speech. The goal of the evaluation is to learn how the synthetic visual talker falls short of natural talkers and to modify the synthesis accordingly to bring it more in line with natural visible speech. The intelligibility of Baldi's visible speech has been successively improved across a number of studies, although overall it still falls somewhat short of a good natural talker [10].

An important goal for the application of talking heads is to have a large gallery of possible agents and to have highly intelligible and realistic synthetic visible speech. Our development of visible speech synthesis is based on facial animation of a single canonical face, Baldi. Although the synthesis, parameter control, coarticulation scheme, and rendering engine are specific to Baldi, we have developed software to reshape our canonical face to match various target facial models [11][12].

To achieve realistic and accurate synthesis, we use measurements of facial, lip, and tongue movements during speech production to optimize both the static and dynamic accuracy of the visible speech. This optimization process is called minimization because we seek to minimize the error between the empirical observations of real human speech and the speech produced by our synthetic talker [11][12]. To improve the intelligibility of our talking heads, we have developed software for using dynamic 3D optical measurements (Optotrak) of points on a real face while talking. The new control parameters obtained from the optical measurements were about 4 times more accurate in reproducing the natural speech than the previous generation of control parameters.

2. Development and Evaluation of a Speech Training Tutor

We have extended previous approaches by viewing speech learning as a multisensory experience [9]. We test the idea that Baldi can function effectively as a language tutor to teach speech perception and production. As mentioned, 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 [11]. Although previous approaches have used palatometry and electropalatography as a form of visual feedback, Baldi can display a more representative view of the actual articulation. Baldi can demonstrate articulation by illustrating a midsagittal view, or the skin on the face can be made transparent to reveal the internal articulators. In addition to simply showing the actual articulation, the area of contact between the tongue and palate and teeth can be highlighted during the articulation.

To optimize the learning experience, the illustrated articulation can be slowed down and the orientation of the face can be changed to display different viewpoints, such as a side view, or a view from the back of the head. 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. Although the hypothesis remains untested, 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 many people use a map. They often orient the map in the direction they are headed to make it easier to follow (e.g. turning right on the map is equivalent to turning right in reality).

In addition to illustrating appropriate articulations that are usually hidden by the face, Baldi can be made even more informative by embellishing of the visible speech with added characteristics that do not occur in natural speech. The primary motivation for these characteristics is to visually distinguish phonemes that have similar visible articulations. As examples, the difference between voiced and voiceless segments can be indicated by vibrating the neck during the voicing period; nasal sounds can be marked by making the nasal opening red; and turbulent airflow during frication can be characterized by lines emanating from the mouth during their articulation. These embellished speech cues potentially make the face more informative than it normally is. In addition to guiding speech production, a body of reading research implies that these additional visible cues would heighten the child's phonological awareness of the articulation of these segments and promote learning to read.

2.1 Training Children with Hearing Loss

Children with hearing loss require guided instruction in speech perception and production. Some of the distinctions in spoken language cannot be heard with degraded hearing—even when the hearing loss has been compensated by hearing aids or cochlear implants. To test whether the Baldi technology has the potential to help individuals with hearing loss, we carried out a speech training study with seven students (2 male and 5 female ages 8 to 13), from the Jackson Hearing Center and JLS Middle School in Los Altos, California participating in the study [14]. The students were trained to discriminate minimal pairs of words bimodally (auditorily and visually), and were also trained to produce various speech segments by visual information about how the inside oral articulators work during speech production. 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. Each student completed approximately 45 minutes of 8 training lessons for a total of 6 hours of training.

The students' ability to accurately perceive and produce words involving the trained segments improved from pre-test to post-test. Intelligibility ratings of the post-test productions were significantly higher than pre-test productions, indicating significant learning. It is always possible that some of this learning occurred independently of our program or was simply based on routine practice. To test this possibility, we assessed the students' productions six weeks after training was completed. Although these productions were still rated as more intelligible than the pre-test productions, they were significantly lower than post-test ratings, indicating some decrement due to lack of continued participation in the training program. This is evidence that at least some of the improvement must have been due to the program.

2.2 Training Adults Learning Mandarin Chinese

To implement multilingual agents, we have developed a client/server architecture system [12]. The client is the application controlling Baldi. It sends text from a variety of languages including Arabic, Mandarin, Russian, and many European languages as well as English to a general speech synthesis server. The server generates the appropriate phonemes in the appropriate language with all the information needed by the client (phonemes, duration, pitches, word boundaries, etc.) and the acoustic speech waveform, and then it sends them back to the client. Using this information, the client generates the appropriate language-specific visible phonemes synchronized with the synthesized speech.

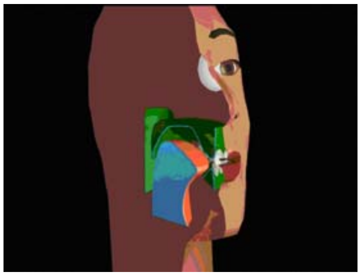

Figure 1. *Picture of a synthetic visual syllable in Condition 3.*

We just completed a very short experiment assessing visual speech training and the learning of pronunciation in a new language. Forty-five English-speaking students with no experience in Mandarin Chinese at the University of Pittsburgh were recruited and randomly assigned to one of three conditions (15 in each condition). Each condition consisted of two sessions of training followed by two sessions of testing, and participants were instructed to spend 15 minutes in each training session. In the first session, all of the participants heard the sound of a native Mandarin speaker (Beijing female) pronouncing 23 Mandarin syllables. Table 1 lists the 23 Mandarin syllables in pin-yin (the alphabetical system used in Mainland China). After this first session, participants continued to do the second session, in which participants in condition 1 simply repeated session 1. Participants in the other two conditions had visible speech added to the same sounds that were presented in the first session. Those subjects in condition 2 learned from the sound and face of the same speaker, and those subjects in condition 3 learned from the same sounds but now aligned with Bao, who is a modified version of Baldi who has been modified to speak Mandarin. In this training condition. Bao spoke at twice the normal duration of the natural Mandarin speaker and his visible articulators were shown, as can be seen in Figure 1. At the end of the second training session, participants' did the first testing session. Their pronunciations of the 23 syllables were recorded without a speed requirement for pronunciation. Then they did the second testing session, in which they were asked to pronounce the Pinyins presented on a computer screen as fast and accurately as possible. Both reaction time and accuracy were recorded. An independent native Mandarin speaker judged the participants' pronunciation of the 23 syllables in a scale from 1 (totally wrong) to 5 (excellent).

Table 1. *The syllables are all tone 1 Mandarin words (pin-yin) except those with the tones 2 and 3 indicated in parentheses. UC = unique consonants; NUC = Non-unique consonants; NUS = Non-unique syllables; US = unique syllables; UV = unique vowels*

UC	NUC	NUS	US	UV
ji	Pi	bao	ju	ge
qie	Nie	dao	qu	he
xian	Tian	gao	xu	ke
zhen	Fen			e(2)
chuan	kuan			u(3)
sha	La			

An analysis of variance (ANOVA) was carried out on the pronunciation ratings with two factors: Condition (3 levels), and Syllable (23 levels). The dependent variable (DV) was the pronunciation rating averaged across the two testing sessions. The training condition had no effect, with the average ratings of 4.032, 4.129, and 4.267 for conditions 1-3, respectively, [F(2,35) = .26, p = .77].

Syllable had a significant main-effect [F(22, 770) = 44.57, p < .001]. Evaluating the matched-pairs of ji-pi, qie-nie, xian-tian, zhen-fen, chuan-kuan, sha-la, the Mandarin-unique consonants that do not occur in English yielded significantly lower ratings than non-unique consonants that do occur. For example, this was evident in the pin-yin syllable "qie", which had a significantly lower average rating than all other Mandarin syllables. This result is reasonable because the closest English segment [tʃ] differs from the Mandarin [tɕ^{h}].

Syllables with the Mandarin-unique vowel [ɤ] yielded significantly lower ratings of their pronunciation than those with the vowel [aʊ], which also occurs in English.

Finally, although being instructed to spend about 15 minutes for each session, the participants did not all spend the same time in training. Therefore, another ANOVA assessed whether the training times differed across the three training conditions. The same factors were used, but the DV was the individual training time. In this analysis, however, there were no significant main effects for Condition. Overall, the participants spent roughly equal training times across the three conditions.

In summary, there seemed to be some overall themes: 1) Mandarin-unique consonants and vowels tended to yield lower ratings than non-unique Mandarin consonants and vowels, and 2) The pronunciation ratings did not differ across the three Conditions (audio-only, audio with real visual, audio with synthetic visual). However, subjects had only a single training session with visible speech, and it is impressive that viewing the fairly complex movements of the visible articulators pronouncing a new language did not impede pronunciation learning.

3. Discussion and Conclusion

Although training with visible speech showing the tongue and palate during pronunciation did not facilitate pronunciation learning relative to just auditory speech, subjects had only a single training session of 15 minutes. We expect from previous research that a more prolonged training period would show an advantage of visible speech [13]. Baldi has achieved an impressive degree of initial success as a language tutor with hard-of-hearing children [14][15]. The same pedagogy and technology has been employed for language learning with autistic children [16]. The improvements obtained from measures of real talking faces and documented in the evaluation testing have been codified, incorporated and implemented in current uses of the visible speech technology. Ultimately, improved visible speech in computer-controlled animated agents will allow all users to extract visible speech information from orally-delivered presentations. This is especially important for enhanced acquisition of speechreading in newly-deafened adults, language acquisition together with word enunciation in children with hearing loss, and those learning a new language.

We look forward to research and applications in the use of embodied conversational agents for language learning. The field offers a potentially significant technology and pedagogy that can facilitate language learning and thereby improve communications among linguistically and culturally diverse societies.

4. Acknowledgements

The research and writing of the paper were supported by the National Science Foundation (Grant No. CDA-9726363, Grant No. BCS-9905176, Grant No. IIS-0086107), Public Health Service (Grant No. PHS R01 DC00236), a Cure Autism Now Foundation Innovative Technology Award, the National Alliance for Autism Research, and the University of California, Santa Cruz, and an NSF project grant SBE-0354420 to the University of Pittsburgh.

5. References

[1] Massaro, D. W. *Symbiotic Value of an Embodied Agent in Language Learning*. In Sprague, R.H., Jr. (Ed.), IEEE Proc. of 37th Annual Hawaii Intl. Conference on System Sciences, 2004.

[2] Kähler, K., J. Haber, H.-P. Seidel: *Geometry-based Muscle Modeling for Facial Animation*, Proc. Graphics Interface. 2001.

[3] Guenter, B., C. Grimm, D. Wood, H. Malvar and F. Pighin. *Making faces*. SIGGRAPH, Orlando - USA 55-67. 1998.

[4] Bregler, C., Covell, M. & Slaney, M. *Video rewrite: Driving visual speech with audio* in Proc. of ACM SIGGRAPH 97, 1997.

[5] Ezzat, T., G. Geiger and T. Poggio. *Trainable videorealistic speech animation* ACM Trans. on Graphics, 21(3): 388-398. 2002.

[6] Parke, F.I. *A model for human faces that allows speech synchronized animation*. Journal of Computers and Graphics, 1(1). 1975.

[7] Cohen, M. M., & Massaro, D. W. *Modeling coarticulation in synthetic visual speech*. In N. M. Thalmann & D. Thalmann (Eds.) Models and Techniques in Computer Animation. Springer-Verlag, Tokyo, 1993

[8] Beskow, J. *Trainable Articulatory Control Models for Visual Speech Synthesis*. Journal of Speech Technology, 7(4), to appear.

[9] Massaro, D. W. *Perceiving talking faces: From speech perception to a behavioral principle.* Cambridge, MassachuseTtS: MIT Press, 1998

[10] Jesse, A., Vrignaud, N., & Massaro, D. W. (2001/01). The processing of information from multiple sources in simultaneous interpreting. Interpreting 5, 95-115.

[11] Cohen, M.M. and Massaro, D.W. & Clark R. (2002). Training a talking head. In D.C. Martin (Ed.), Proceedings of the IEEE Fourth International Conference on Multimodal Interfaces, (ICMI'02)(pp. 499-510). Pittsburgh, PA.

[12] Massaro, D.W., Ouni, S., Cohen, M.M., & Clark, R. (2005). A Multilingual Embodied Conversational Agent. In R.H. Sprague (Ed.), Proceedings of 38th Annual Hawaii International Conference on System Sciences, (HICCS'05) (CD-ROM,10 pages). Los Alimitos, CA: IEEE Computer Society Press.

[13] Bernsen, N.O., Hansen, T.K., Kiilerich, S. and Madsen, T.K.: Field Evaluation of a Single-Word Pronunciation Training System. Proceedings of The Fifth International Conference on Language Resources and Evaluation, LREC 2006, Genova, Italy, May, 2006, 2068-2073.

[14] Massaro, D.W. & Light, J. Using visible speech for training perception and production of speech for hard of hearing individuals. Journal of Speech, Language, and Hearing Research, 47(2), 304-320, 2004.

[15] Massaro, D.W., & Light, J. (2004). Improving the vocabulary of children with hearing loss. Volta Review, 104(3), 141-174.

[16] 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(6),653-672.