Using Visual Speech for Training Chinese Pronunciation: An In-vivo Experiment

Ying Liu[1], Dominic W. Massaro[2], Trevor H. Chen[2], Derek Chan[1], and Charles Perfetti[1]
[1]Learning Research and Development Center
University of Pittsburgh, Pittsburgh, PA15260
[2]Department of Psychology, Perceptual Science Laboratory
University of California, Santa Cruz
Santa Cruz, CA 95060 U.S.A.

Abstract

Recent research showed that our perception and understanding of speech are influenced by a speaker's facial expressions 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. Baldi® is a three-dimensional animated talking head appropriately aligned with either synthesized or natural speech. The present in-vivo experiment used Bao, a Chinese version of Baldi, to teach Chinese syllables to adult English native speakers. The result showed that students trained with Baldi, improved more than students trained with ordinary speech. 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 technological edge of Baldi holds great promise in language learning, dialog, human-machine interaction, education, and edutainment.

1. Introduction

In recent years, studies have demonstrated that proper use of multi-modality input can facilitate learning [1]. Current use of technologies makes it possible for language learning to take advantage of the visual modality. The Science Laboratory (PSL-UCSC) aims 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 hearing-challenged 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 [2].

There have been several approaches to facial animation, including muscle models to simulate the muscle and tissues during talking [3], performance-based synthesis that tracks a live speaker[4], and image synthesis that combines together images of a real speaker [5][6]. The facial animation used in the current applications, however, is a descendant of Parke's software and his particular 3-D talking head [7]. Modifications include 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 facial subareas. Many of the facial shape parameters – such as cheek, neck, or forehead shape, and 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

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

phoneme is represented as a set of target values for the control parameters such as jaw rotation and mouth width. 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 how much weight its target value carries against those of neighboring segments independently for each control parameter over time [8]. In a test of several coarticulation models, Beskow [9] found that our model gave the best fit to the observed articulatory data.

We evaluate the accuracy and intelligibility of Baldi's synthetic visible speech by perceptual recognition tests given to human observers [10]. These experiments 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 short of a good natural talker [11].

The present study is part of a larger foreign language learning project associated with the Pittsburgh Science of Learning Center (PSLC). In a previous study, we carried out one experiment to test the effect of Baldi under laboratory environment [12]. The present experiment was carried out online, including both training and testing. We tested the effectiveness of Baldi vs. human face and voice only in an in-vivo experiment.

2. Method

2.1 Participants

One hundred and one students of an introductory Chinese course at Carnegie Mellon University participated this study. They logged into the web based learning program to participate in the study as a course requirement, which account for 5 percent of their final grade.

2.2 Procedure and Materials

Each participant was randomly assigned to one of the three training conditions. Each training condition consisted of two sessions of training and two sessions of testing. In the first session, all participants received audio only training by hearing the sound of a native Mandarin speaker (a female speaking a Beijing dialect) pronouncing 23 Mandarin syllables with the Pinyin spelling presented on the screen. Table 1 lists the 23 Mandarin syllables in pin-yin (the alphabetical system used in Mainland China). As shown in the table, some of the consonants and vowels are unique in Chinese. They either do not exist in English or are pronounced differently in Chinese and English. After this first session, participants immediately took the first testing session. After one week, participants logged back into the learning program to do the second training session, in which participants were randomly assigned to one of the three different training conditions. In condition 1, the training method simply repeated the audio only training in 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, which is a modified version of Baldi who has been modified to speak Mandarin (Figure 1). In the third training condition, Bao's visible articulators were shown, as can be seen in Figure 1.

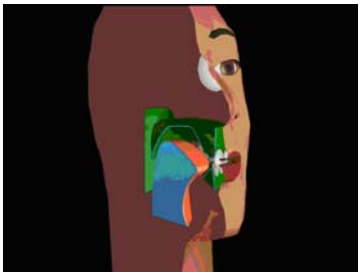

Figure 1

At the end of the second training session, participants did the second test. In both the first and second tests, their pronunciations of the 23 syllables were recorded without any time constraint. Two independent native Mandarin speakers coded blind the participants' pronunciation of the 23 syllables on initial onsets and final rimes respectively in a scale from 0 (incorrect), 1 (correct but not accurate) to 2 (accurate).

Table 1. *The syllables are all tone-1 Mandarin words (pin-yin) except those with the tones 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			

3. Results

The performance on initials and finals were separately analyzed by a logistic model [13], which modeled the probability of improvements of test 2 over test 1. In both analyses, the dependent variable was the number of improved initials/finals in test 2 divided by the number of initials/finals that did not receive code 2 (accurate) in test 1. This independent variable measured the percentage of improvement for all initial and finals that had the potential to improve. The independent variable was training condition (audio only, human face, vs. synthetic Baldi).

Least square means from the two logistic models are listed in Table 2. The mathematical mean is not calculated because the distribution of dependent variable is binomial instead of normal. Nevertheless, similar to mathematical means, a higher least square mean indicates a larger improvement.

Table 2. Least square mean percentages of improvement after session 2 based on a logistic model

	Initials	Finals
Audio only	53.1	34.2
Human face	54.5	39.6
Synthetic Baldi	51.5	46.4

It can be seen that all three conditions had more than 50% improvement on initial pronunciations, which showed that for all initials which are not pronounced accurately, more than half of them will be accurate after the second training session. However, the statistical tests on initials did not showed any significant training condition effect (χ^2 (2)=0.25, p=0.8815). Although there are small differences, the p value indicates that all three training methods yielded the same level of improvement.

The means of improvement on finals were less than initials and varied significantly across conditions. The analysis on finals means showed a significant condition effect (χ^2 (2)=7.39, p=0.025). Further pairwise comparisons showed that synthetic Baldi was significantly better than the audio only condition (χ^2(1)=7.36, p=0.0067), while the human face fell between Baldi and audio only conditions, but the differences did not reach significance (Baldi vs. Human face: p=0.1145; Human face vs. audio only: p=0.1542).

4. Discussion

All three methods improved the pronunciation accuracy after the second training session. Both the initials and finals were significantly improved. However, the patterns of initial and final improvements were different. For the initial, there was no difference between the three methods, but significant advantage of Baldi over audio only was observed in teaching final pronunciations.

Three possible factors might be responsible for the differences found between initial and final position. The first is that the initial onsets in Chinese are normally consonants, which are pronounced fairly quickly with perhaps minimal visual information available from either a human or animated talking head. Instead, the pronunciation of the final rimes are usually vowels, which are much longer and might provide more visual information for the learners to use.

The second factor is that the final rimes might have a higher similarity than the initial onsets in Chinese. For example, the finals “a”, “an”, “ao”, “ai”, “ang”, “ian”, and “iang” all share the “a” pronunciation. Thus, visual cues from an animated or human talking head provide more benefit for learning finals than initials.

The third factor is that the outside of the face might provide more information for the initial onsets whereas the inside of the mouth would provide more information for the final rimes. This would account for the finding that Bao was more helpful for the final rimes than was the human face.

We conclude that visual speech provides significant benefit for learners to improve their

pronunciation. 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 have 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 information from orally-delivered presentations. This is especially important for enhanced acquisition of speech reading 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 communication across linguistically and culturally diverse societies.

5. Acknowledgements

Baldi® is a trademark of Dominic W. Massaro. The research and writing of the paper were supported by the National Science Foundation (SBE-0354420, CDA-9726363, BCS-9905176, 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.

6. References

[1] Mayer, R., and Moreno, R. "Aids to computer-based multimedia learning", Learning and Instruction, 12(1):107-119, 2002.

[2] 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.

[3] Kähler, K., Haber, J., and Seidel, H. P. "Geometry-based Muscle Modeling for Facial Animation", Proc. Graphics Interface. 2001.

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

[5] Bregler, C., Covell, M., and Slaney, M. "Video rewrite: Driving visual speech with audio", Proc. of ACM SIGGRAPH 97, 1997.

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

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

[8] Cohen, M. M., and 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.

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

[10] Massaro, D. W. "Perceiving talking faces: From speech perception to a behavioral principle". Cambridge, Massachusetts: MIT Press, 1998.

[11] Jesse, A., Vrignaud, N., and Massaro, D. W. "The processing of information from multiple sources in simultaneous interpreting", Interpreting, 5:95-115, 2001.

[12] Massaro, D. W., Liu, Y., Chen, T. H., & Perfetti, C. A. "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, 2006

[13] Agresti, A. "Categorical data analysis", 165-266, New York: Wiley-Interscience, 2002.

[14] Massaro, D. W., and 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. "Improving the vocabulary of children with hearing loss". Volta Review, 104(3): 141-174, 2004.

[16] Bosseler, A., and Massaro, D. W. "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, 2003.