In K. Miesenberger, J. Klaus, W. Zagler, & A. Karshmer (Eds.) Computers Helping People with Special Needs. 10th International Conference, ICCHP 2006 Linz, Austria, July 2006. Berlin: Springer.

Embodied Agents in Language Learning for Children with Language Challenges

Dominic W. Massaro[1]

[1] Department of Psychology
University of California, Santa Cruz
Santa Cruz, CA 95064 U.S.A.
Massaro@ucsc.edu
http://mambo.ucsc.edu/dwm

Abstract. Given the value of face-to-face interaction in communication and learning, our persistent goal has been to develop, evaluate, and apply animated agents to produce realistic and accurate speech. We have implemented these agents as computer-assisted speech and language tutors for hard of hearing and autistic children, and other children with language challenges. Our language-training program utilizes conversational agents, who guide students through a variety of exercises designed to teach vocabulary and grammar, to improve speech articulation, and to develop linguistic and phonological awareness. We report a new experiment showing its effectiveness for school children learning English as a new language. Some of the advantages of this pedagogy and technology include the popularity and effectiveness of computers and embodied conversational agents, the perpetual availability of the program, and individualized instruction. Animated tutors offer a promising approach to language learning, human-machine interaction, and education.

1 Introduction

Language challenges are pervasive in today's world. Given the maximum migration that our society has ever experienced, many individuals become members of a new linguistic community. Given the highly mobile society, individuals of all walks of life find themselves in situations in which successful education, business, and social interactions require use of a nonnative language. As an obvious example, English is becoming increasing necessary and desirable, and the number of people in the world who are learning English is increasing at a rapid rate. Learning a new language, however, is a significant challenge for all individuals, whether young or old. In addition, there are surprisingly many individuals who have language and speech disabilities, and these individuals like those learning a new language require additional instruction in language learning. Currently, however, these needs are not being met because there are not enough skilled teachers and professionals to give them the one on one attention that they need. So they resort to other resources, such as books or other media, but the problems with these are that they are not easily personalized to the students'

needs, they lack the engaging capability of a teacher, they are rather expensive, and they are relatively ineffective.

In this paper, we describe several language learning applications with a virtual tutor for language learning and speech training. We begin by describing research that demonstrates that our perception and understanding of language are influenced by a speaker's face and accompanying gestures, as well as the actual sound of the speech.

1.1 State of the Art: Value of the Face in Communication

Although traditionally speech has been viewed as solely an auditory phenomenon, speech as a multimodal phenomenon is supported by experiments indicating that our perception and understanding are enhanced by a speaker's face and accompanying gestures, as well as the actual sound of the speech [1,2]. There are several reasons why the use of auditory and visual information together is so successful. These include a) robustness of visual speech, b) complementarity of auditory and visual speech, and c) optimal integration of these two sources of information. Speechreading, or the ability to obtain speech information from the face, is robust in that perceivers are fairly good at speechreading even when they are not looking directly at the talker's lips [1]. Furthermore, accuracy is not dramatically reduced when the facial image is blurred (because of poor vision, for example), when the face is viewed from above, below, or in profile, or when there is a large distance between the talker and the viewer.

Complementarity of auditory and visual information simply means that one of the sources is strong when the other is weak. A distinction between two segments robustly conveyed in one modality tends to be relatively ambiguous in the other modality. Two complementary sources of information make their combined use much more informative than would be the case if the two sources were non-complementary [1]. Perceivers combine or integrate the auditory and visual sources of information in an optimally efficient manner. Many different empirical results have been accurately predicted by a Fuzzy Logical Model of Perception (FLMP) that describes an optimally efficient process of combination [1].

1.2 Embodied Conversational Agents

The value of visible speech in face-to-face communication was the primary motivation for the development of Baldi®, a 3-D computer-animated talking head. Baldi provides realistic visible speech that is almost as accurate as a natural speaker [1,3]. The goal of the visible speech synthesis carried out in the Perceptual Science Laboratory (PSL) has been to develop a polygon (wireframe) model with realistic motions (but not to duplicate the musculature of the face). Our animated face can be aligned with either the output of a speech synthesizer or natural auditory speech [3]. We have also developed the phoneme set and the corresponding target and coarticualtion values to allow synthesis of several other languages. These include Spanish (Baldero), Italian (Baldini), Mandarin (Bao), Arabic (Badr), French (Balduin), German

(Balthasar), and Russian (Balda). Baldi, and his various multilingual incarnations are seen at: http://mambo.ucsc.edu/psl/international.html.

In our facial animation algorithm, each segment is specified with a target value for each facial control parameter. Coarticulation, defined as changes in the articulation of a speech segment due to the influence of neighboring segments, is based on a model of speech production using rules that describe the relative dominance of the characteristics of the speech segments [1,3].

A central and somewhat unique quality of our work is the empirical evaluation of the visible speech synthesis, which is carried out hand-in-hand with its development. The quality and intelligibility of Baldi's visible speech has been repeatedly modified and evaluated to accurately simulate naturally talking humans [1,3]. The gold standard we use is how well Baldi compares to a real person. Given that viewing a natural face improves speech perception, we determine the extent to which Baldi provides a similar improvement. We repeatedly modify the control values of Baldi in order to meet this criterion. We modify some of the control values by hand and also use data from measurements of real people talking.

Several advantages of utilizing a computer-animated agent as a language tutor are clear, including the popularity of computers and embodied conversational agents. Computer-based instruction is emerging as a prevalent method to train and develop vocabulary knowledge for both native and second-language learners and individuals with special needs [4]. An incentive to employing computer-controlled applications for training is the ease in which automated practice, feedback, and branching can be programmed. Another valuable component is the potential to present multiple sources of information, such as text, sound, and images in parallel. Instruction is always available to the child, 24 hours a day, 365 days a year. Furthermore, instruction occurs in a one-on-one learning environment for the students. Applications with animated tutors perceived as supportive and likeable will engage foreign language and ESL learners, reading impaired, autistic and other children with special needs in face-to-face computerized lessons. We now review several different applications utilizing Baldi to carry out language tutoring, and then report a new study in the learning of vocabulary in a new language.

2 Pedagogy of language learning

Vocabulary knowledge is critical for understanding the world and for language competence in both spoken language and in reading. There is empirical evidence that very young children more easily form conceptual categories when category labels are available than when they are not [4]. Even children experiencing language delays because of specific language impairment benefit once this level of word knowledge is obtained. It is also well-known that vocabulary knowledge is positively correlated with both listening and reading comprehension [5]. It follows that increasing the pervasiveness and effectiveness of vocabulary learning offers a timely opportunity for improving conceptual knowledge and language competence for all individuals, whether or not they are disadvantaged because of sensory limitations, learning disabilities, or social condition.

Learning and retention are positively correlated with the time spent learning. Our technology offers a platform for unlimited instruction, which can be initiated whenever and wherever the child and/or mentor chooses. Instruction can be tailored exactly to the student's need, which is best implemented in a one-on-one learning environment for the students. Other benefits of our program include the ability to seamlessly meld spoken and written language, and provide a semblance of a game-playing experience while actually learning. Given that education research has shown that children can be taught new word meanings by using direct instruction methods [5], we implemented these basic features in an application to teach vocabulary and grammar.

2.1 Research and methodological approach

To test the effectiveness of vocabulary instruction using an embodied conversational agent as the instructor, we developed a series of lessons that encompass and instantiate the developments in the pedagogy of how language is learned, remembered, and used.

One of the principles of learning that we exploit most is the value of multiple sources of information in perception, recognition, learning, and retention. An interactive multimedia environment is ideally suited for learning [4]. Incorporating text and visual images of the vocabulary to be learned along with the actual definitions and sound of the vocabulary facilitates learning and improves memory for the target vocabulary and grammar. Many aspects of our lessons enhance and reinforce learning. For example, the existing program makes it possible for the students to 1) Observe the words being spoken by a realistic talking interlocutor, 2) Experience the word as spoken as well as written, 3) See visual images of referents of the words, 4) Click on or point to the referent or its spelling, 5) Hear themselves say the word, followed by a correct pronunciation, and 6) Spell the word by typing, and 7) Observe and respond to the word used in context.

2.2 Effectiveness for hearing loss

It is well known that children with hearing loss have significant deficits in both spoken and written vocabulary knowledge. To assess the learning of new vocabulary, we carried out an experiment based on a within-student multiple baseline design where certain words were continuously being tested while other words were being tested and trained [6]. Although the student's instructors and speech therapists agreed not to teach or use these words during our investigation, it is still possible that the words could be learned outside of the learning context. The single student multiple baseline design monitors this possibility by providing a continuous measure of the knowledge of words that are not being trained. Thus, any significant differences in performance on the trained words and untrained words can be attributed to the training program itself rather than some other factor.

We studied eight children with hearing loss, who needed help with their vocabulary building skills as suggested by their regular day teachers. The experimenter de-

veloped a set of lessons with a collection of vocabulary items that was individually composed for each student. Each collection of items was comprised of 24 items, broken down into 3 categories of 8 items each. Three lessons with 8 items each were made for each child. Images of the vocabulary items were presented on the screen next to Baldi as he spoke. Assessment was carried out on all of the items at the beginning of each lesson. It included identifying and producing the vocabulary item without feedback. Training on the appropriate word set followed this testing.

As expected, identification accuracy was always higher than production accuracy. This result is expected because a student would have to know the name of an item to pronounce it correctly. There was little knowledge of the test items without training, even though these items were repeatedly tested for many days. Once training began on a set of items, performance improved fairly quickly until asymptotic knowledge was obtained. This knowledge did not degrade after training on these words ended and training on other words took place. In addition, a reassessment test given about 4 weeks after completion of the experiment revealed that the students retained the items that were learned.

2.3 Effectiveness for autism

The tutoring application has also been used in evaluating vocabulary acquisition, retention and generalization in children with autism [7]. Although the etiology of autism is not known, individuals diagnosed with autism must exhibit a) delayed or deviant language and communication, b) impaired social and reciprocal social interactions, and 3) restricted interests and repetitive behaviors. The language and communicative deficits are particularly salient, with large individual variations in the degree to which autistic children develop the fundamental lexical, semantic, syntactic, phonological, and pragmatic components of language. Vocabulary lessons were constructed, consisting of over 84 unique lessons with vocabulary items selected from the curriculum of two schools. The participants were eight children diagnosed with autism, ranging in age from 7-11 years.

The results indicated that the children learned many new words, grammatical constructions and concepts, proving that the application provided a valuable learning environment for these children. In addition, a delayed test given more than 30 days after the learning sessions took place showed that the children retained over 85% of the words that they learned. This learning and retention of new vocabulary, grammar, and language use is a significant accomplishment for autistic children.

Although all of the children demonstrated learning from initial assessment to final reassessment, it is possible that the children were learning the words outside of our learning program (for example, from speech therapists or in their school curriculum). Furthermore, it is important to know whether the vocabulary knowledge would generalize to new pictorial instances of the words. To address these questions, a second investigation used the single subject multiple probe design, as was done in [6]. Once a student achieved 100% correct, generalization tests and training were carried out with novel images. The placement of the images relative to one another was also random in each lesson. Assessment and training continued until the student was able

to accurately identify at least 5 out of 6 vocabulary items across four unique sets of images.

Although performance varied dramatically across the children and across the word sets during the pre-training sessions, training was effective for all words sets for all children. Given training, all of the students attained our criterion for identification accuracy for each word set and were also able to generalize accurate identification to four instances of untrained images. The students identified significantly more words following implementation of training compared to pre-training performance, showing that the program was responsible for learning. Learning also generalized to new images in random locations, and to new interactions outside of the lesson envirronment. These results show that our learning program is effective for children with autism, as it is for children with hearing loss.

We were gratified to learn that the same application could be used successfully with both autistic children and children with hearing loss [4]. Specific interactions can be easily modified to accommodate group and individual differences. For example, autistic children are much more disrupted by negative feedback, and the lesson can be easily designed to instantiate errorless learning. Given these successes, we carried out a study with English language learners (ELL).

2.4 Effectiveness for English language learners (ELL)

This study was carried out with English Language Learners (ELL) and involved the use of a recently-released application, Timo Vocabulary (http://animatedspeech.com), which instantiated the pedagogy we found in our earlier research 6,4,7]. Nine children ranging in age from 6-7 years were tested in the summer before first grade. Almost all of the children spoke Spanish in the home. The children were pretested on lessons in the application in order to find three lessons with vocabulary that was unknown to the children. A session on a given day included a series of three test lessons, and on training days, a training lesson on one of the three sets of words. Different lessons were necessarily chosen for the different children because of their differences in vocabulary knowledge. As shown in Figure 1, the test session involved the presentation of the images of a given lesson on the screen with Timo's request to click on one of the items, e.g., Please click on the oven. No feedback was given to the child. Each item was tested once in two separate blocks to give 2 observations on each item. Three different lessons were tested, corresponding to the three sets of items used in the multiple baseline design. A training session on a given day consisted of just a single lesson in which the child was now given feedback on their response. Thus, if Timo requested the child to click on the dishwasher and the child clicked on the spice rack, Timo would say, "I asked for the dishwasher, you clicked on the spice rack. This is the dishwasher. The training session also included the Elicitation and Imitation sections in which the child was asked to repeat the word when it was highlighted and Timo said it, and the child was asked to say the item that was highlighted. Several days of pretesting were required to find lessons with unknown vocabulary. Once the 3 lessons were determined, the pretesting period was followed by the training days. Given that the children learned at different rates, the results were

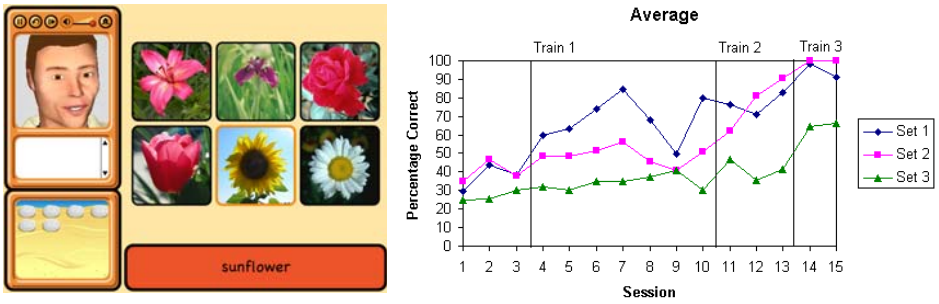

Fig. 1. The left panel shows a typical screen shot from the Timo Vocabulary application. The right panel gives the average percentage of correct identifications for each of the three sets of words. The three vertical bars indicate when training was initiated for each of the three sets, respectively. Performance improved on each set of words after training on that set was initiated.

averaged with respect to when the different training regimens were initiated. Thus, for example, the results at block 3 give performance before training on Set 1 words, and the results at block 4 give performance after one block of training on Set 1 words. As can be seen in Figure 1, training was effective in teaching new vocabulary. Some of the variability in the figure is due to having just a few subjects at some blocks: Blocks 9 and 15 have just 1 subject's data.

3. Future Plans and Conclusion

Animated Speech is releasing a Lesson Creator that allows easy creation of new language lessons. This user-friendly application allows the composition of lessons with minimal computer experience and instruction. Although it has many options, the program has wizard-like features that direct the coach to explore and choose among the alternative implementations in the creation of a lesson. The application will include a curriculum of thousands of vocabulary words, and can be implemented to teach both individual vocabulary words and metacognitive awareness of word categorization and generalized usage. This application will facilitate the specialization and individualization of vocabulary and grammar lessons by allowing teachers to create customized vocabulary lists from words already in the system or with new words. If a teacher is taking her class on a field trip to the local Aquarium, for example, she will be able to create lessons on the fish the children will see at the museum. A parent could prepare lessons on the child's relatives, her schoolmates, and teachers. Most importantly, lessons can easily be easily created for the child's most recent interest.

We found that the Timo Vocabulary is effective in teaching vocabulary to English Language Learners. This result replicates previous studies carried out on hearing-impaired and autistic children with Baldi as the animated conversational tutor. In other experiments, we have also observed that Baldi's unique characteristics allow a novel approach to training speech production to both children with hearing loss [8]

and adults learning a new language [9]. We look forward to new studies that continue to test the value of computer-animated tutors for language learning and related applications in education and human-machine interaction.

4. Acknowledgements

Baldi® is a trademark of Dominic W. Massaro. The research and writing of the paper were supported by the National Science Foun-dation (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. The author thanks Roger Kimbrough for testing the children and Jennifer Schmida and David Payne for supporting this research project.

References

1. Massaro, D. W. (1998). *Perceiving Talking Faces: From Speech Perception to a Behavioral Principle*. MIT Press: Cambridge, MA.
2. Granström B, House D & Beskow J (2002). Speech and gestures for talking faces inconversational dialogue systems. In B. Granstrom, D. House, & I. Karlsson (Eds.), *Multilmodality in language and speech systems* (pp. 209-241). The Netherlands: Kluwer Academic Publishers.
3. 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.
4. Massaro, D. W. (2004). *Symbiotic Value of an Embodied Agent in Language Learning.* In R.H. Sprague, Jr.(Ed.), Proceedings of 37th Annual Hawaii International Conference on System Sciences,(HICCS'04) (CD-ROM,10 pages). Los Alimitos, CA: IEEE Computer Society Press. Best paper in Emerging Technologies.
5. Beck, I. L., McKeown, M. G., & Kucan, L. (2002). *Bringing words to life: Robust Vocabulary Instruction*. New York: The Guilford Press.
6. Massaro, D.W., & Light, J. (2004). Improving the vocabulary of children with hearing loss. *Volta Review, 104*(3), 141-174.
7. 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.
8. Massaro, D.W., & Light, J. (2004). 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.
9. Massaro, D. W., & Light (2003, September). *Read my tongue movements: bimodal learning to perceive and produce non-native speech /r/ and /l/.* Eurospeech 2003-Switzerland (Interspeech). 8th European Conference on Speech Communication and Technology, Geneva, Switzerland.