Massaro, D.W., Ouni, Slim, Cohen, Michael M., & Clark, Rashid (2005). A Multilingual Embodied Conversational Agent. IEEE Proceedings of 38th Annual Hawaii International Conference on System Sciences (HICCS'05) (CD-ROM, 8 pages). Los Alimitos, CA: IEEE Computer Society

A Multilingual Embodied Conversational Agent

Dominic W. Massaro, Slim Ouni, Michael M. Cohen, Rashid Clark
Department of Psychology, Perceptual Science Laboratory
University of California, Santa Cruz
Santa Cruz, CA 95060 U.S.A.
http://mambo.ucsc.edu/dwm
Massaro@fuzzy.ucsc.edu; Slim@fuzzy.ucsc.edu; Mmcohen@ranx.ucsc.edu; Rashid@fuzzy.ucsc.edu

Abstract

In our paper last year [1]*, we described our language-training program, which utilizes Baldi as a tutor, 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. In this paper, we describe the technology of Baldi in more detail, and how it has been enhanced along a number of dimensions. First, Baldi is now multilingual and is capable of acquiring new languages fairly easily. Second, Baldi has grown a body to allow the communication of communicative gesture.*

1. Introduction

One goal of the Perceptual Science Laboratory (PSL) has been 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, 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].

Baldi is an expert English speaker. It is, of course, valuable to add new languages to Baldi's speaking repertoire to extend speech and language research and applications to other languages. Our previous method used to introduce a new language was to add its implementation within the talking head application. This procedure requires a large amount of software development to add new languages. To overcome these limitations, we have implemented a client/server system to extend the capabilities of Baldi to speak other languages than English (Brazilian Portuguese, Spanish, Italian, French, German, Arabic Mandarin, Swedish, and Danish).

In order to increase the quality of the interaction with the animated agent, we have augmented Baldi by adding a body, hands, arms, shoulders, and legs to extend communication through gesture.

2. 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 over the last 12 years have included increased resolution of the underlying wireframe model, additional and modified control parameters, a realistic tongue, coarticulation, paralinguistic information and affect in the face, alignment with natural speech, text-to-speech synthesis, and bimodal (auditory/visual) synthesis. 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 shape, and also some affect parameters such as smiling – use interpolation.

Baldi's facial animation can be aligned with either natural auditory speech or the output of a speech synthesizer. Phonemes are used as the unit of 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 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.

3. Extension of the Technology of Taking Heads: Accurate visual animation

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 [10].

A Cyberware 3D laser scanning system is used to enroll new citizens in our gallery of talking heads. To illustrate this procedure, we describe how a Cyberware laser scan of the head of the senior author (DWM) was made, how Baldi's generic morphology was mapped into the form of DWM, how this head was trained on real data, and how the quality of its speech was evaluated.

A laser scan of a new target head produces a very high polygon count representation. Rather than trying to animate this high-resolution head (which is impossible to do in real-time with current hardware), our software uses these data to reshape our canonical head to take on the shape of the new target head. A human operator marks corresponding facial landmarks, such as the corner of the nose, on both the laser scan head and the generic Baldi head. The Baldi head is then warped until it assumes as closely as possible the shape of the target head, with the additional constraint that the landmarks of the Baldi face move to positions corresponding to those on the target face. Then, for the main portion of the head, we first translate all vertices so that the center point of the model coincides with the coordinate system origin. We then move the vertices so that they are at a unit distance from the origin. At this point, the vertices of the triangles making up the model are on the surface of the unit sphere. This is done to both Baldi's source head and the Cyberware laser scan target head. The landmarks are then connected into a mesh of their own. As these landmarks are moved into their new positions, the non-landmark points contained in triangles defined by the landmark points are moved to keep their relative positions within the landmark triangles. Then, for each of these source vertices we determine the location on the target model to which a given source vertex projects. This gives us a homeomorphic mapping (1 to 1 and onto) between the source and target datasets, and we can thereby determine the morph coordinate of each source vertex as a barycentric coordinate of the target triangle to which it maps. This mapping guides the final morph between the source and target datasets.

We implemented independent control parameters for each side of the face. There is evidence that people show an asymmetry in their facial movements for both speech and affect [11]. For example, some talkers tend to open the right side of their mouths faster and wider than the left side. This difference for the two sides results in a problematic rip along the facial midline. To repair the rip, algorithms compromise the different sides of the face by adjusting the Y and Z values of each point within a certain threshold distance of the midline. Asymmetry during speech synthesis can be implemented by having steeper dominance functions for the more dominant side, which will cause less coarticulation and therefore more extreme articulation. Another approach is to simply have different segment definitions for the two sides.

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 [12][13]. 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. For our initial measurements, we recorded a large bimodal (auditory-visual) speech database by placing markers on the face of DWM at important locations that move significantly during speaking.

Fitting of these dynamic data occurred in several stages. To begin, we assigned points on the surface of the synthetic model that best correspond to the Optotrak measurement points. There were 19 marker points on the face in addition to 4 points on the top of the head that were used to remove head motion from these 19 points. Two of the 19 points (on the eyebrows) were not used in the current study. The other 17 points were used to train the synthetic face. In the training, the Optotrak data were scaled to best match the corresponding points marked on the synthetic face.

The data collected for the training consisted of 100 CID sentences recorded by DWM speaking in a fairly natural manner. In the first stage fit, for each time frame (30 fps) we automatically and iteratively adjusted 11 facial control parameters of the face to get the best fit (the least sum of squared distances) between the Optotrak measurements and the corresponding point locations on the synthetic face. A single jaw rotation parameter was used, but the other 10 parameters were fit independently for the two sides of the face. This yielded 21 best-fitting parameter tracks that were the inputs to the second stage fit.

In the second stage fit, the goal was to tune the segment definitions (parameter targets, dominance function strengths, attack and decay rates, and peak strength time offsets) used in our coarticulation algorithm [7] to get the best fit with the parameter tracks obtained in the first stage fit. We first used Viterbi and hand alignment of the acoustic speech data to obtain the phonemes and their durations. Given the phonemes and durations, we used the then current default parameters in our standard parametric phoneme synthesis and coarticulation algorithm to synthesize the parameter tracks for all 100 sentences [10]. These were compared with the parameter tracks obtained from the first stage fit, the error computed, and the control parameters adjusted until the best fit was achieved. The root mean squared (RMS) error for this fit was 12%, less than ½ of the error of the original TtS. We have also used phoneme definitions conditional on the following phoneme. In the 100 sentences there were 509 such diphones. These definitions gave an RMS error of 6%. Thus, 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.

Perceptual Evaluation of Improved Animation

We carried out a perceptual recognition experiment with human subjects to evaluate the how well our synthetic talker conveyed speech information relative to the real talker [10]. To do this we presented the 100 CID sentences in three conditions: auditory alone, auditory + synthetic talker, and auditory + real talker. In all cases there was white (speech band) noise added to the audio channel. For the synthetic talker, the 21 best-fitting parameters from the first stage fit were used to drive the face. The auditory signal was analyzed using Viterbi alignment to derive the phonemes and durations for using our standard TtS for the speech parameters not influenced by the Optotrak data. For example the TtS was used to drive the tongue, because the Optotrak data does not measure tongue motion.

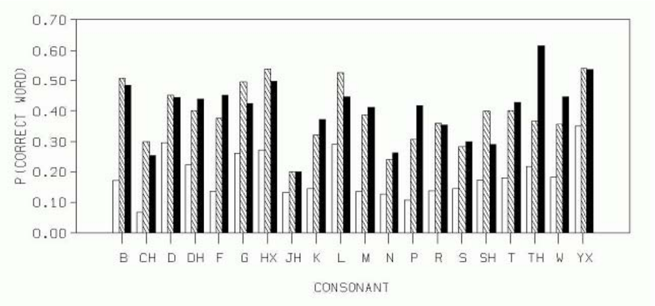


Figure 1. Percentage of correct words as a function of the auditory (white), synthetic (striped), and natural face (black) conditions.

Figure 1 gives the percentage of correct words as a function of the auditory (white), Baldi (striped), and Natural Face (black) conditions. Overall, the proportion of correctly reported words for the three conditions was 0.22 auditory, 0.43 synthetic face, and 0.42 with the real face. The two visual conditions did not differ significantly from each other (F(1,9) = 0.66), showing that the intelligibility improvement given by Baldi was equivalent to that given by a natural talker.

4. Extension of the Technology of Taking Heads: Toward a multilingual talking head

To have multilingual talking head there are several requirements to be achieved at the auditory level and visual level:

Auditory: For any new language to be added to the Talking head system, a text-to-speech (TTS) engine capable of that language is required. Often there are several TTS engines available for a given language. Within one TTS engine, many voices (female and male voices, for instance) might also be available. Existing TTS engines are developed by different groups, they do not provide exactly the same output, and they may require different input formats. This means that a unique interface definition is required to deal with different TTS input and output.

Visual: Baldi speaks English. To have Baldi speak a new language accurately, new phonemes will be probably required as well as revised definition for the existing phonemes. For each phoneme, an articulatory prototype should be provided with the articulator target values and coarticulation properties of that phoneme. For this purpose, we have developed a set of graphical editing tools to view and adjust the control parameters and their dynamic behavior (Figure 2). These tools allow us to define new visual language-specific phonemes using either real physical measurements from a speech corpus to train the face [8] or more qualitative analyses of the articulation of the new language.

We improved the existing Baldi software by reorganizing the system and making it more modular. The improved system uses a client/server architecture; the server deals with auditory level, and the client deals with the visual and animation level. Such architecture makes the system flexible, distributable and usable via a network. The client software can be any application built around the talking head. The server module provides a standard interface to various TTS systems (Figure 3).

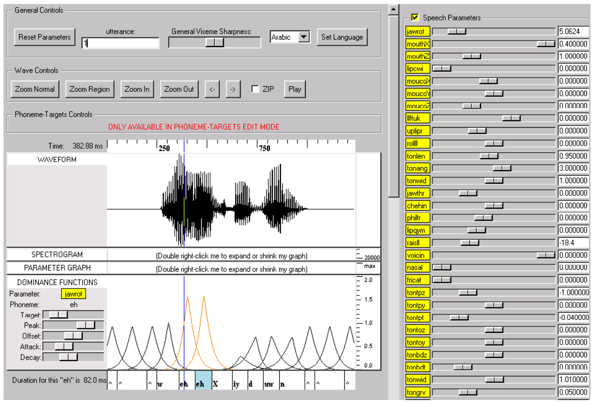
The server

To synthesize many languages, we require a variety of different TTS engines. Our TTS server utilizes a consistent interface between the different TTS engines and the talking head. A TTS engine has to provide the phonemes, their durations, and the synthesized auditory speech. The server can support many TTS engines for one or more languages. Thus it is now fairly easy to add a new language if we have its corresponding TTS engine. The server preprocesses the text to be synthesized in the appropriate format that each TTS engine expects, and then it post-processes the symbolic and waveform results for the client. This scheme keeps the client and the server fairly independent. It is also possible to send inquiries to the server (for example: what are the installed TTS engines, what are the installed voices, is it female or male voice, change the current voice, change the speech rate, etc.). In addition, the server handles timestamp bookkeeping for any commands embedded in the text on request and sends this information to the client. When the server is started, it installs in memory the required TTS engines (that exist on the machine), and then sends to the client the list of TTS engines, languages, voices and other information related to each voice.

The client

The client is an application built around the talking head. This can be a simple graphic user interface (GUI) that allows the user to simply enter text or include more detailed control of the talking head and of the speech (slowing down the speech, adding emotion and gesture, etc.). Basically, the client sends the text to be uttered to the server and the server sends back the information required by the client.

It is not necessary that the TTS server and the face client run on the same machine. This is very advantageous if the machines have limited resources. These machines can run the client and send the inquiries (including the utterance to synthesize) via a network to another more powerful machine running the server. The resources required by the server depends on the number of TTS engines and their respective footprints (high quality requires a huge footprint).



utterance: 1

Arabic

Time: 382.88 ms

Parameter: jawrot

Phoneme: eh

Duration for this "eh" is 82.0 ms

jawrot	5.0624
mouthX	0.400000
mouthZ	1.000000
lipcwi	0.000000
moucoX	0.000000
moucoY	0.000000
moucoZ	0.000000
llftuk	0.000000
uplipr	0.000000
rollll	0.000000
tonlen	0.950000
tonang	3.000000
tonwid	1.000000
jawthr	0.000000
chehin	0.000000
philtr	0.000000
lipqym	0.000000
raisll	-18.4
voicin	0.000000
nasal	0.000000
fricat	0.000000
tontpz	-1.000000
tontpy	0.000000
tontpt	-0.040000
tontoz	0.000000
tontoy	0.000000
tonbdz	0.000000
tonbdt	0.000000
tonwid	1.010000
tongrv	0.050000

Figure 2.: A screenshot of the tool used to edit and redefine the articulation of Baldi.

We applied this new architecture to extend the capabilities of Baldi to speak new languages. In addition to English, we extended Baldi to speak Arabic, German, French, Chinese Mandarin, Italian, Castilian Spanish, Mexican Spanish, Brazilian Portuguese, Swedish and Danish. The supported TTS engines are Euler/Mbrola, AT&T Natural voices, Lucent Articulator, NeoSpeech and any TTS compliant with Microsoft SAPI4 or SAPI5. Our system can be extended to include any language used by these TTS engines or new engines can be added to the server. There is no modification to be done on the client when a new TTS engine for languages already existing on the client is added to the server.

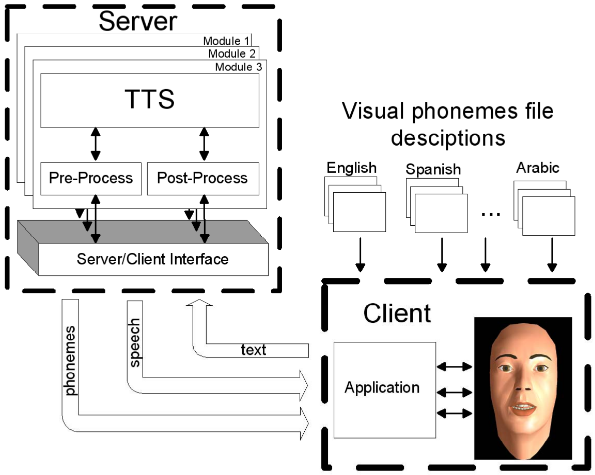


Figure 3. Client/Server architecture system.

5. Potential Applications for the multilingual talking head

A multilingual talking head will help to make many languages accessible for research and practical applications. By extending the research in auditory/visible speech performed for English to many other languages, it will be possible to study language differences and similarities.

Improving the accuracy of our talking head is particularly important because of its recent applications in language learning. It is well known that hard of hearing children fall behind in both spoken and written vocabulary knowledge. One reason is that these children tend not to overhear other conversations because of their limited hearing and are thus shut off from an opportunity to learn vocabulary. A series of experiments demonstrated that these children can learn new vocabulary with a Language Wizard/Player incorporating Baldi as a tutor [14]. The language and communicative challenges faced by autistic children are also particularly salient, and Baldi has proven effective in teaching them new vocabulary and grammar [15].

Baldi has also been used as a speech and listening tutor for hard of hearing children [16]. Some of the distinctions in spoken language cannot be heard with degraded hearing, even when the hearing loss has been compensated for by hearing aids or cochlear implants. In this case, Baldi's visible speech can provide guided instruction in speech perception and production. Other potential applications include teaching phonological awareness in learning to read and the learning of new languages. We look forward to progress in the application of animated speech to communication and human machine interaction.

6. Developing A Body and Communicative Gesture

In order to increase the quality of the interaction with the animated agent, we have augmented Baldi by adding a body, hands, arms, shoulders, and legs to extend communication through gesture. Including body language capabilities creates another source of information for the listener to use, and thus can make communication more successful. We studied the integration of a pointing gesture with audible speech [17]. Preschool and fourth-grade children were presented with gesture, speech, and both sources of information together. Both auditory speech and gestures influenced performance, and the form of the results was essentially identical to those found in experiments with audible and visible speech. Each source of information presented alone had some influence, and their joint influence followed the predictions of the fuzzy logical model of perception, an optimal model for combining several sources of information [9]. A second advantage to including a body is that it can help a character become a more general-purpose avatar. Arms, hands, tractor treads, cherub wings, extension cords, and most anything else imaginable can be made a part of the avatar's body, giving more potential with which to manipulate, interact with, and blend into a changing environment.

A basic, general means of handling bodies for Baldi is an animation engine which supports inserting and transitioning between arbitrary gestures in a script format. Scripts are activated and generated from the recognition systems, which would access various scripts appropriate to the educational setting. A useful means to tailor script building for such a situation is to define each gesture as having a temporal "hot spot". This is the point in the gesture animation at which it is to be synchronized to a specific time. For example, consider the gesture in which Yoda emphases a word by poking his walking cane into Luke Skywalker when the word is spoken [18]. Part of that character's emphasis gesture would be to lift the cane into position, in front of himself and pointed towards Luke, and then tap the cane forward at the proper moment. In this case, the frame of animation in which Yoda's cane is prodded fully into Skywalker would be defined as that gesture's "hot spot". The application can simply specify the time at which the emphasized word occurs, and the script builder would make certain to arrange the full gesture in the following manner. The lead-in animation (occurring before the cane makes contact with Luke) would be timed to occur before the point of emphasis, resulting in a beautiful synchronization of Yoda's tap without the text-to-gesture application needing to know the dynamics of the chosen gesture.

The body itself is animated via a technique known as forward kinematics. This is the process of arranging a body into a hierarchical structure of parts, which are linked to one another. For instance, the shoulder is linked to the upper arm, which is in turn linked to the forearm, which is linked to the hand, and so on. Any motion applied to a body part is to also be applied to those parts, which are linked higher to it in the hierarchy. That is, if the upper arm is moved, then the forearm and hand will be moved along with it. By collectively building upon the motions of the body parts as the hierarchy is traversed, the final pose is formed.

The animation engine that has been created for the body is also very flexible. It can support multiple bodies and multiple scripts, and display as many of them on the screen at once as hardware will allow. Computational resources can be conserved by rendering less detail about bodies that are more distant from the viewer. The animation engine therefore allows for the possibility of utilizing more than one agent (See Figure 4). They can speak different languages and each can make different gestures. In addition each agent can express different emotions. Each body displayed on the screen keeps track of its own body parts, which in turn keep track of their own parameters (such as color, on/off state, etc).

An individual body also learns of body parts that it does not have but which other, arbitrary bodies happen to have. Thus, if one of the bodies that is loaded has wings, then a wing-less body can store information about wings. This allows for a degree of flexibility which can be useful in future situations where an avatar may be asked to change to different shapes, and back again, without losing any of the settings used in the body parts unique to the shape it returns to. Scripts are also interchangeable among bodies. A body will find and utilize those motions applicable to it; thus a winged body can use gestures defined for a wingless body and vice versa.

Although solid body parts rotating about each other is a fast and adequate method for handling animation (it is the most popular method used in the computer game market), a single, stretching “skin” has been developed. Doing so not only would increase the quality of the body, but it is also be a better visual match to Baldi's head, which also uses a skin stretching technique. There are various approaches to creating a single mesh for a 3D structure, which vary in quality and speed. The approach described here is to work from the basics up, to begin with a relatively simple skin algorithm and to proceed to add quality improvements as they seem necessary, while testing the impact these additions have upon performance.

When performing interpolations between gestures, it is possible in a simple, mathematical transition from one gesture to another that body parts will pass through one another. For example, if gesture1 consists of holding the arms behind the figure’s back, and gesture2 consists of holding the arms in front of the figure, then a straight transition between the two would move the arms through the figure’s body. To algorithmically calculate a path of travel that avoids any collisions and looks natural is a problem that is complex in scope, and no doubt one which may be prone to exceptions and errors. Thus, a proposed method of solving this is to provide a means for the animator to test the transitions to and from any newly created gestures against the pre-existing gestures. If any of the transitions are judged unnatural, the animator may then opt to create her/his own transition, or in many cases supply a single in-between key frame, which the scripting code will utilize for those special cases. This approach would result in the most visually natural solution to the problem, and the option to only supply a single, “universal” key frame would reduce the animator’s workload for a gesture that has a common collision problem with many other gestures.

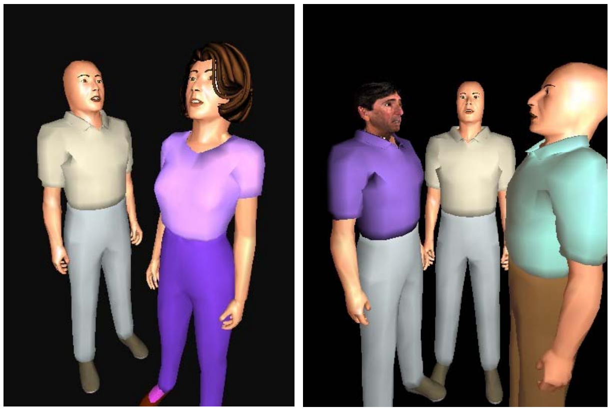


Figure 4. Example of three conversational agents with personalized bodies and heads.

Our conversational agent has real-time performance, in the sense that the user is not kept waiting an intolerably long time for a response. However, it is far from natural in the sense of being able to converse at the natural pace so evident in human-human interaction. The problem is not merely one of the limitations of today's speech recognition algorithms and hardware power. Even if a computer could understand words instantaneously, it would often be inappropriate for it to produce a response as soon as it could. In human-human conversation, people, except when being deliberately rude, obey certain conventions for when to speak and when to listen. Baldi will have to acquire understanding and rules of conversational dialogue to mimic a human interlocutor. Ideally, we should also specify a mood or personality for Baldi, such as the businesslike Baldi, the sympathetic-listener Baldi, the patient-but-firm-tutor Baldi, the entertaining Baldi, and so on.

Along with this enhanced conversational ability, Baldi should also be capable of performing new actions. One is the ability to nod in the direction of and/or to point at various objects in the environment and also manipulating objects (e.g. pick something up). Our research to date has provided a foundation for allowing effective signaling by our conversational agent. The emotional states of the talking face and the menu of gestures of the head and body allow Baldi to initiate the type of nonlinguistic displays expected from believable interlocutors. Signals that provide symbolic meaning have been called emblems. Table 1 lists common examples that are meaningful in North America.

Table 1. Examples of visual and auditory gestures and their meanings in North America.

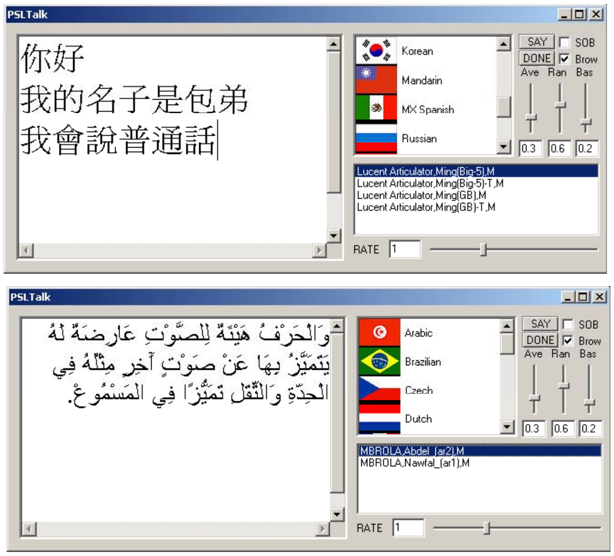
Visual Gesture	**Meaning**
Head nod	Yes
Thumb up	I approve
Greeting wave	Hello
Shoulder shrug	I don't know
Head shake	No
Thumb down	I reject
Farewell wave	Good-bye
Wink	I'm kidding
Auditory Gesture	**Meaning**
Clap	I approve
Wolf whistle	How beautiful
Raspberry	I dishonor you
Hiss	I disapprove
Rise-fall whistle	How surprising
Tongue-click	Shame on you

In addition to these more blatant visual gestures, there are more subtle gestures such as the many illustrators that accompany narrative, as studied by McNeill [19]. There are also fleeting and subtle gestures in typical conversations. The eyes look to the upper left for an instant, the mouth tenses for a moment, the eyebrows go up, or the mouth opens slightly. One also observes a head toss, a noisy in breath, or bringing the hand closer to the face. These gestures can have communicative significance, as you can tell by watching any two people talking on the bus. We expect that these facial gestures correlate with various aspects of the speaker's pragmatic functions, and also how they correlate with subsequent responses of various types by the participant.

7. Toward a multilingual embodied conversational agent

As an example of the use of the new system, we developed a client application that allows the user to enter text, choose a language, a TTS engine, and a speaker to utter the text. Since this application is dealing with multilingual agent, the graphical user interface (GUI) supports different scripting allowing non-western languages to be displayed correctly using UNICODE characters. For example, it is very easy to display Mandarin Chinese or Arabic words, and to switch text entry from right-to-left or left-to-right (See Figure 5).

A configuration interface was also added to specify different agents and their configurations. This interface allows the control of various parts of the agent: changing the color and the shape of the head, teeth, eyes, skin, etc (Figure 6). It allows editing of the control parameters, modification of the phoneme definitions, dominance functions, and viewing characteristics. It also allows the control of multiple characters (see Figure 4) the use textured-map head of a real speaker (see section 2), and the creation of movies or later playback.



你好
我的名子是包弟
我會說普通話

Lucent Articulator,Ming(Big-5),M
Lucent Articulator,Ming(Big-5)-T,M
Lucent Articulator,Ming(GB),M
Lucent Articulator,Ming(GB)-T,M

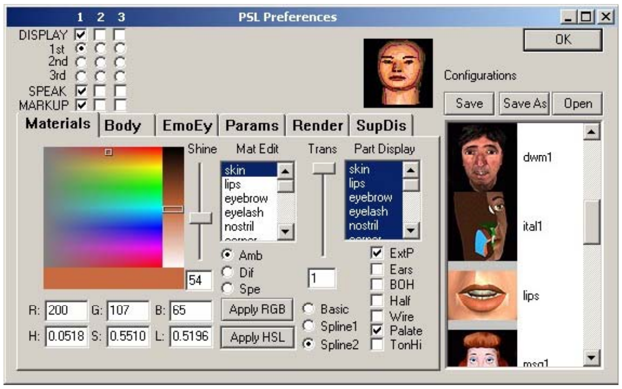
RATE 1

وَالْحَرْفُ هَيْئَةٌ لِلصَّوْتِ عَارِضَةٌ لَهُ
يَتَمَيَّزُ بِهَا عَنْ صَوْتٍ آخَرِ مِثْلِهُ فِي
الْحِدَّةِ وَالثِّقَل تَمَيُّزًا فِي المَسْمُوعْ.

MBROLA,Abdel_(ar2),M
MBROLA,Nawfal_(ar1),M

RATE 1

Figure 5: Multilingual text examples with Chinese in the upper panel and Arabic (which is read from right to left) in the lower panel.



R: 200 G: 107 B: 65
H: 0.0518 S: 0.5510 L: 0.5196

Figure 6. Configuration GUI that allows control of display characteristics of the agent as well as the editing of the visible speech animation.

Figure 7 illustrates a GUI that governs the control of facial expression and gesture while talking. Emotions can be inserted in the text entry so that Baldi can express his anger, happiness, sadness, etc. Various hand and body gestures can also be inserted. It is possible to add some body gestures independently of whether or not the agent is speaking. For example one agent can nod his head while a second agent is speaking. Given that TtS engines produce

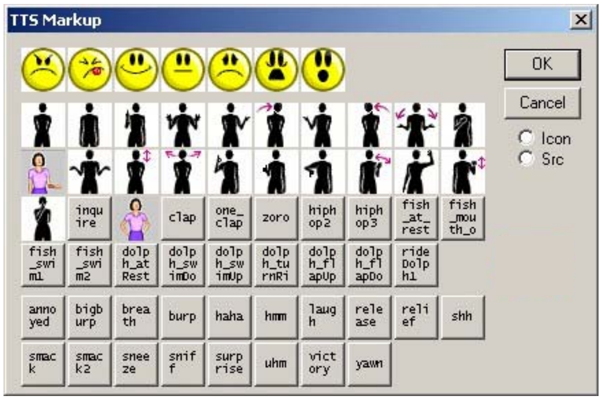


Figure 7. GUI to add emotion and gesture.

only speech, we have created a set of non-speech sounds (laugh, smack, etc.) that can also be inserted.

More recently, we extended the current system toward a multilingual embodied conversational agent. In addition of having a multilingual talking head, we are able to display multiple heads and bodies. Currently we can display up to three agents with personalized heads and bodies, as was shown in Figure 4.

8. Acknowledgement

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, and the University of California, Santa Cruz. All of the software and development environments described in this paper were created in the Perceptual Science Laboratory at the University of California, Santa Cruz.

9. 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] Cohen, M.M. and Massaro, D.W. & Clark R. *Training a talking head*. In D.C. Martin (Ed.), Proc. of the IEEE Fourth Intl Conf. on Multimodal Interfaces, (ICMI'02)(pp. 499-510). PiTtSburgh, PA., 2002.

[11] Graves, R. & Potter, S.M. Speaking from two sides of the mouth. Visible language, 22, 129-137. 1998.

[12] Cohen, M. M., Beskow, J. & Massaro, D. W. *Recent developments in facial animation: An inside view.* Proc. of Auditory Visual Speech Perception '98. Terrigal-Sydney Australia, 1998.

[13] Cohen, M.M., Clark, R. & Massaro, D.W. *Animated speech: Research progress and applications*. In D.W. Massaro, J. Light & K. Geraci (Eds.), Proc. of Auditory-Visual Speech Processing, Aalborg, Denmark, 2001.

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

[15] Bosseler, A. & 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.

[16] 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.

[17] Thompson, L.A., & Massaro, D.W. (1994). Perceptual development. In V.S. Ramachandran (Ed.) Encyclopedia of Human Behavior, Orlando, FL: Academic Press, 3, 441-451, 1994

[18] Lucas, G. (1973). Star Wars. Lucas Films.

[19] McNeill, D. *So you think gestures are nonverbal?* Psychological Review, 92, 350-371. 1985.