

# Internationalization of a Talking Head

Slim Ouni, Dominic W. Massaro, Michael M. Cohen, Karl Young and Alexandra Jesse

Perceptual Science Laboratory – University of California – Santa Cruz

Santa Cruz, California, USA

E-mail: { slim, massaro, karly, alex }@fuzzy.ucsc.edu, mmcohen@ranx.ucsc.edu

## ABSTRACT

In this paper we describe a general scheme for internationalization of our talking head, Baldi, to speak other languages. We describe the modular structure of the auditory/visual synthesis software. As an example, we have created a synthetic Arabic talker, which is evaluated using a noisy word recognition task comparing this talker with a natural one.

## 1. INTRODUCTION

One goal of the Perceptual Science Laboratory (PSL) has been to create computer-animated agents that produce accurate auditory and visible speech, as well as realistic facial expressions, emotions, and gestures [1]. 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 [2]. Baldi has also been used for speech training of both hard of hearing children [3] and adults learning a second language [4]. The animated characters that we are developing have also been used to train autistic children “read” visible speech [5]. These applications illustrate how synthetic talking heads can facilitate face-to-face communication and access to online information presented orally by either real or lifelike computer characters.

Baldi produces reasonably accurate visible English speech, as well as realistic facial expressions, emotions, and gestures and is composed of a 3D wireframe with parametrically controlled movements, controllable texture mapping, and internal structures (hard palate, velum, teeth and tongue). The development of the internal structures is important as it improves the visible speech quality and accuracy, and is of great value for pedagogically illustrating correct articulation. Baldi can be Viterbi aligned with natural auditory speech or can be aligned with speech from a text-to-speech synthesizer (TTS), as long as the TTS provides the phonemes and their durations before the utterance is said. Target values of the control parameters and their dynamic behavior are specified for each phoneme and the animation is made fluent and realistic by using a coarticulation scheme [1].

It is, of course, valuable to add new languages to Baldi’s speaking repertoire. This is important for extending visible

and bimodal speech research and for creating applications in other languages. The previous method used to introduce new languages like Spanish and Italian [6] was to add the implementation of that module within the talking head system. This solution limits the possibility of having many different languages and the talking system is hard to maintain for further addition. Recently, we have been working to extend the capabilities of Baldi to speak other languages than English (Spanish, Italian, French, German, Mandarin, and Arabic).

To evaluate our improved system, we use empirical evaluation of the visible speech synthesis recognition by human subjects, which is carried out hand-in-hand with its development. These experiments are aimed at evaluating the realism of our speech synthesis relative to natural speech. The goal of the evaluation is to learn how our synthetic visual talker falls short of natural talkers and to modify the synthesis accordingly to bring it more in line with natural visible speech.

## 2. GENERAL SCHEME FOR A MULTILINGUAL TALKING HEAD

To have Baldi speak a new language accurately, new phonemes and parametric specifications for these phonemes must be defined. For this purpose, we have developed a set of graphical editing tools to view and adjust the control parameters and their dynamic behavior (Figure 1). These tools allow us to define new visual language-specific phonemes using either real physical measurements from a speech corpus to train the face [7] or more qualitative analyses of the articulations of the new language.

In addition, we improved the Baldi software by reorganizing the system and making it more modular. The improved system uses client/server architecture. The client software is currently an application built around the talking head constructed within the CSLU Toolkit [8], but could be some other standalone application. The server module provides a standard interface to various TTS systems (Figure 2).

### *The server*

The most important component of the server is the TTS. After receiving the text to synthesize, the minimum requirements are that a TTS has to provide for the face the phonemes and their durations, pitch values, and the

synthesized speech. We can have many servers: each server supports a TTS for the one or more languages (for example, a server supporting Chinese using Lucent TTS [9], a server supporting English using Festival TTS [10] or AT&T natural voice, a server supporting Arabic, German and French using Euler/Mbrola TTS [11], etc.). Thus it is now fairly easy to add a new language if we have its corresponding TTS.

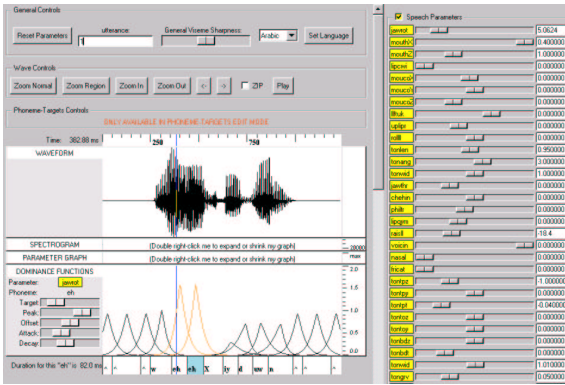


Figure 1: A screenshot of the tool used to edit and redefine the articulation of Baldi.

### The client

The client sends to the server the text to be uttered and the server sends back all the information needed by the client. The client mainly needs the phonemes and their duration to produce the visual phonemes and synchronize them with the synthesized speech. Adding new languages requires the introduction of new visual phonemes and the definition of their articulation and coarticulation. For the improved talking head, we define the list of the phonemes and their definitions (articulation and coarticulation) in separate files so we can add as many languages as we want. To define the articulation and the coarticulation we use software that helps to define the articulation of each phoneme by adjusting each facial (and also vocal tract) control parameters and the coarticulation by giving the changes of articulation over time (Figure1).

### The client/server system

The new architecture offers many advantages. We can have a server for each TTS system. The server can deal with many clients, even via the Internet. This general scheme allows the client) to run without requiring the different TTS servers to run together on the same PC.

## 3. APPLICATION OF THE NEW SCHEME: INTERNATIONALIZATION OF BALDI

We applied this new architecture to extend the capabilities of Baldi to speak three of the new languages: Arabic, German and French. The chosen TTS supporting these languages is based on Euler/Mbrola TTS [11]. We improved the existing Euler version and we added extra modules but we used Mbrola as it is. The main improvements were made for the Arabic module because

this language was most extensively studied and also presented many problems in the accuracy of the phoneme generation and the prosody. In fact, some phonemes provided by Euler to Mbrola were wrong and there were some grapheme-to-phoneme rules missing.

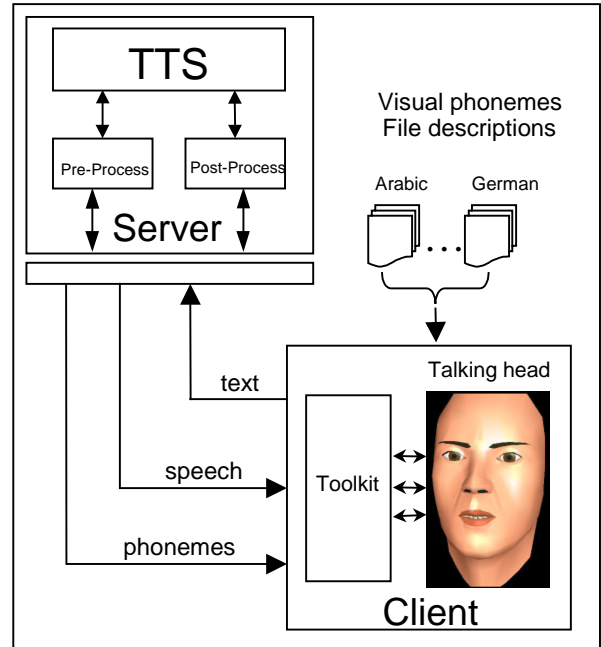


Figure 2: Client/Server architecture system

An additional module was made to support the German language. This server based on Euler/Mbrola is operational. The client side was also improved to support any new language by simply by preparing the definition files for each new language. We worked on the three languages but developed Arabic the most. Thus, we describe our results for this language.

In a preliminary study, we began by defining the Arabic phonemes the same as the most similar English ones. Phonemes that occurred only in Arabic were defined appropriately. We performed a perceptual evaluation of the system and used the analysis of the preliminary evaluation along with the literature on Arabic phonetics and articulation [12,13,14] to improve the definitions of the Arabic phonemes.

The Arabic visual phonemes were specified on both the visible part (the face) and partially visible (tongue and inner articulators). The data and information found in literature are mainly midsagittal views of the vocal tract (positions of the tongue relatively to the palate and teeth), and facial views of some phonemes. We also made a video recording of an Arabic speaker (SO) reading a list of words, which we used to compare and improve the articulation. There is also some information about the coarticulation that we utilized [12,13,14]. As some phonemes were not described, we used the results for English.

#### 4. EXPERIMENTAL EVALUATION

Our evaluation experiments help to determine how easily perceivers can speechread the face and how much the face adds to intelligibility of auditory speech presented in noise. In one experiment, participants were asked to recognize noisy auditory phrases presented alone, or in one of two bimodal conditions with the same noisy auditory input aligned with the computer animation or with the video of the natural face. The auditory input was natural speech. The same auditory was used for the bimodal conditions. For the synthetic face, the visual phonemes were viterbi aligned and then manually adjusted and corrected. We supported this solution instead of using the TTS speech (which is already aligned with the visual phonemes) to prevent that our results will be biased by the errors that may be introduced by the TTS. The stimuli were a set of 300 words, presented in a series of three words on each trial. Thus, the number of trials for each condition was 100. Figure 3 shows an example of the stimuli on a trial. The same list of words was presented in each modality condition but in a different random order. There were 19 participants, all native Arabic speakers living in Tunisia.

حَدِيقَةٌ	قَالَ	السَّرِيرُ
/hadi:qatun/	/qa:la/	/assari:ru/
(garden)	(he said)	(the bed)

Figure 3: Example test trial words (top) with phonetic transcription and English translations.

Subject #	Auditory	Synthetic Face	Natural Face	$C_V^r$
11	29%	47%	69%	44%
17	41%	62%	84%	50%
8	18%	40%	61%	50%
16	31%	54%	73%	55%
14	25%	37%	46%	59%
6	21%	34%	42%	60%
15	26%	47%	61%	61%
9	48%	71%	85%	62%
2	33%	57%	72%	63%
18	25%	52%	68%	63%
7	37%	59%	71%	63%
10	37%	67%	84%	65%
5	31%	54%	66%	66%
12	25%	49%	61%	66%
3	21%	59%	78%	66%
13	35%	60%	73%	68%
1	32%	62%	76%	68%
19	34%	64%	77%	70%
4	27%	55%	65%	73%
<b>Mean</b>	<b>30%</b>	<b>54%</b>	<b>69%</b>	<b>62%</b>

Table 1. The percentage of correct words for the different subjects in the three modalities (unimodal auditory, bimodal synthetic and natural faces). The subjects are arranged in ascending order of the relative visual contribution  $C_V^r$ .

In Table 1, we present the overall results of the experiment. We present the proportion of correct words recognized by each participant under one of the three modalities conditions: unimodal auditory, bimodal synthetic face and bimodal natural face. Under the unimodal auditory, the average of recognized words was 30%. We have a great improvement for the recognized words in bimodal synthetic face (54%), which is close to the performance of the natural face (69%). As we can notice clearly, there is a significant advantage over the unimodal auditory condition when the audio is paired with either the synthetic or natural visual information. Moreover, the performance for the synthetic face compares well with that for the natural face. Figure 3 presents graphically the same results as the table, where the results are sorted using a relative visual contribution, as explained in the next section.

#### 4. NEW INTELLIGIBILITY MEASURE

To better understand the evaluation results it is important to have an accurate measure of the intelligibility of the visible speech. Simply having the highest score in word recognition in bimodal synthetic face does not mean that we have the best visual contribution if the unimodal auditory score is already high.

Sumbly and Pollack [15] proposed a formula to measure the visual contribution to the missing auditory information in a given S/N condition:

$$C_v = \frac{C_{AV} - C_A}{1 - C_A}, \quad (1)$$

where  $C_{AV}$  and  $C_A$  are the bimodal audiovisual and unimodal auditory intelligibility scores. This measure was used later to measure synthetic talking faces intelligibility [16]. However, equation (1) considers that the aim is to retrieve 100% of the information. We know, for example, that degraded auditory speech cannot be completely informative even when presented with natural face [15]. For the intelligibility measure, one goal is to have the synthetic scores equal to the natural scores. Thus, it is better to consider the intelligibility score of the natural face as the maximum of the information that we can retrieve in the noisy bimodal condition. This is what we call the *relative visual contribution*:

$$C_V^r = \frac{C_S - C_A}{C_N - C_A}, \quad (2)$$

where  $C_S$ ,  $C_A$  and  $C_N$  are bimodal synthetic face, unimodal auditory and bimodal natural face intelligibility scores. *The relative visual contribution* in equation (2) is the contribution of the synthetic face relative to the natural face.

The results, in Figure 4 (and Table 1), are sorted using the relative visual contribution  $C_V^r$ . As we can see, for subject 6, the synthetic face helps recognition more than it does for subject 17, even though the latter has a higher level of performance overall. This is because the remaining contribution to reach the natural face result for subject 6 (8%) is smaller than that for subject 17 (21%).

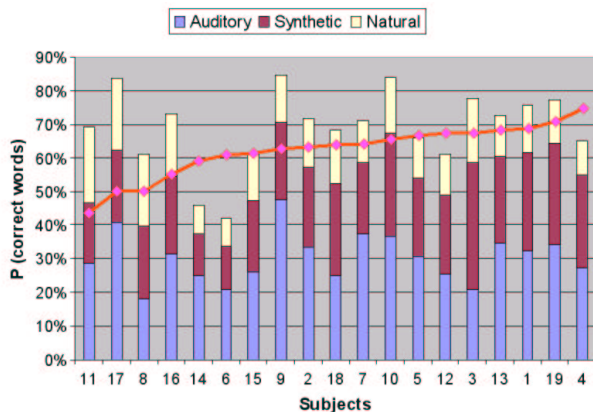


Figure 4: Percentage of correct words in the three modalities presented for each subject (bars) and relative visual contribution (curve). The subjects are arranged in ascending order of the relative visual contribution.

## 5. CONCLUDING REMARKS

The evaluation test shows that our Arabic visible speech synthesis is respectable. We have named this talking head Badr (بدر), meaning full moon. Given our research and development platform, Baldi can be made to speak a new language in a quick and efficient procedure. The evaluation paradigm also provides immediate feedback on the quality of the visible speech in the new language.

## ACKNOWLEDGEMENTS

This research was supported by grants from the National Science Foundation (Grant No. CDA-9726363, Grant No. BCS-9905176, Grant No. IIS-0086107) and the Public Health Service (Grant No. PHS R01 DC00236). Thanks also to Kais Ouni and Nouriddine Ellouze at Ecole Nationale d'Ingenieur de Tunis, Tunisia for running the experiment and providing the test facilities.

## REFERENCES

[1] D.W. Massaro, *Perceiving Talking Faces, From Speech Perception to a Behavioral Principle*, MIT Press, 1998.

[2] A. Bosseler and D.W. Massaro, "Development and Evaluation of a Computer-Animated Tutor for Vocabulary and Language Learning for Children with Autism", *Journal of Autism and Developmental Disorders*, in press, 2002.

[3] D.W. Massaro and J. Light, "Using Visible Speech for Training Perception and Production of Speech for Hard of Hearing Individuals", *Journal of Speech, Language, and Hearing Research*, submitted, 2002.

[4] J. Light and D.W. Massaro, "Learning to perceive and produce non-native speech", unpublished paper, 2002.

[5] D.W. Massaro and A. Bosseler, "Perceiving Speech by Ear and Eye: Multimodal Integration by Children with Autism", unpublished paper, 2002.

[6] P. Cosi, M.M. Cohen, and D.W. Massaro, "Baldini: Baldi speaks Italian", *ICSLP 2002, 7th International Conference on Spoken Language Processing*. Denver, Colorado, 2002.

[7] M.M. Cohen, D.W. Massaro and R. Clark, "Training a talking head", *ICMI'02, IEEE 4th Int. Conf. on Multimodal Interfaces*. Pittsburgh Pennsylvania, 2002.

[8] R. Cole, T. Carmell, P. Connors, M. Macon, J. Wouters, J. de Villiers, A. Tarachow, D.W. Massaro, M.M. Cohen, J. Beskow, J. Yang, U. Meier, A. Waibel, P. Stone, G. Fortier, A. Davis, C. Soland, "Intelligent animated agents for interactive language training", In *STiLL: ESCA Workshop on Speech Technology in Language Learning*, p163-166, Stockholm, Sweden, 1998.

[9] R. Sproat, *Multilingual Text-to-Speech Synthesis: The Bell Labs Approach*, Sproat, R., (ed.), Kluwer Academic Publishers, 1997.

[10] FESTIVAL. A.W. Black, P. Taylor, R. Caley and R. Clark, Centre for Speech Technology - University of Edinburgh. <http://www.cstr.ed.ac.uk/projects/festival>

[11] M. Bagein, T. Dutoit, F. Malfrere, V. Pagel, A. Ruelle, N. Tounsi, D. Wynsberghe, "The EULER Project: an Open, Generic, Multi-lingual and Multi-Platform Text-To-Speech System", *Proc. ProRISC'2000*, pp.193-197, Veldhoven, 2000.

[12] S. Ghazeli, *Back consonants and backing coarticulation in Arabic*, Ph.D. dissertation, university of Texas at Austin, 1977.

[13] S.H. Al-Ani, *Arabic Phonology: An acoustical and physiological Investigation*, Mouton, 1970.

[14] W.H.T. Gairdner, *Phonetics of Arabic*, Humphrey Milford, Oxford University Press, 1925.

[15] W.H. Sumby and I. Pollack, "Visual contribution to speech intelligibility in noise", *J. Acoust. Soc. Am.*, 26, 212-215, 1954.

[16] B. LeGoff, T. Guiard-Marigny, M.M. Cohen and C. Benoît, "Real-Time Analysis-Synthesis and Intelligibility of Talking Faces", *2nd Conf. On Speech Synthesis*, Newark, 1994.