Available online at www.sciencedirect.com
SCIENCE DIRECT®
Speech Communication 45 (2005) 115–137
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

Training Baldi to be multilingual: A case study for an Arabic Badr

Slim Ouni *, Michael M. Cohen, Dominic W. Massaro

Perceptual Science Laboratory, University of California at Santa Cruz, CA, USA

Received 31 January 2004; received in revised form 14 October 2004; accepted 8 November 2004

Abstract

In this paper, we describe research to extend the capability of an existing talking head, Baldi, to be multilingual. We use parsimonious client/server architecture to impose autonomy in the functioning of an auditory speech module and a visual speech synthesis module. This scheme enables the implementation and the joint application of text-to-speech synthesis and facial animation in many languages simultaneously. Additional languages can be added to the system by defining a unique phoneme set and unique phoneme definitions for the visible speech for each language. The accuracy of these definitions is tested in perceptual experiments in which human observers identify auditory speech in noise presented alone or paired with the synthetic versus a comparable natural face. We illustrate the development of an Arabic talking head, Badr, and demonstrate how the empirical evaluation enabled the improvement of the visible speech synthesis from one version to another.
© 2005 Elsevier B.V. All rights reserved.

Keywords: Talking head; Avatar; Visible and visual speech synthesis; Text-to-speech; Auditory; Arabic; Multilingual

1. Introduction

Research during the past several decades proves that the face presents visual information during speech that supports effective communication. Movements of the lips, tongue and jaw enhance intelligibility of the acoustic stimulus, particularly when the auditory signal is noisy (Jesse et al., 2000; Sumby and Pollack, 1954). Given this important dimension of speech, our persistent goal has been to develop and evaluate an animated agent Baldi® (Fig. 1) to produce accurate visible speech (Massaro, 1998). Baldi has a promising potential to benefit virtually all individuals, but especially those with hearing problems (28,000,000 in the USA alone), including the millions of people who acquire age-related hearing loss every year

* Corresponding author. Address: LORIA, Speech Group, 615 rue du jardin botanique, 54600 Villers-lès-Nancy, France. Tel.: +33 3 83 59 20 22; fax.: +33 3 83 27 83 19.
E-mail address: slim@fuzzy.ucsc.edu (S. Ouni).

0167-6393/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2004.11.008

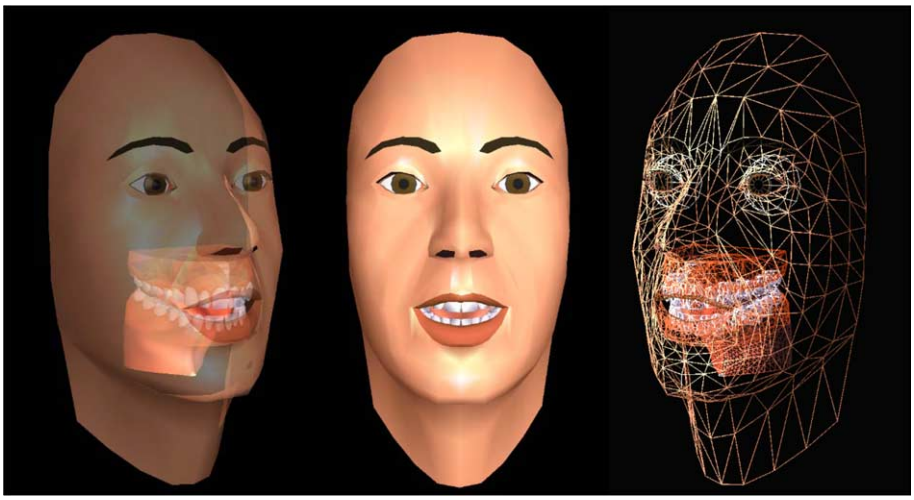

Fig. 1. Talking head Baldi. Three different views. In the middle, the standard Baldi; to the left, semi-transparent Baldi (which allows to see the inner articulation: tongue, palate and teeth); to the right, the wire frame.

(http://www.nidcd.nih.gov/health/statistics/hearing.asp), and for whom visible speech takes on increasing importance. One of many applications allows the training of individuals with hearing loss to "read" visible speech, and thus facilitate face-to-face oral communication in many situations (educational, social, work-related, etc). Baldi can also function effectively as a spoken language tutor, a reading tutor, or personal agent in human machine interaction.

For the past 10 years, the Perceptual Science Laboratory (PSL) has been improving the accuracy of visible speech produced by Baldi (e.g., Cohen et al., 2002). Research has also shown that language learning exercises featuring Baldi can improve both speech perception and production of hard of hearing children (Massaro and Light, 2004b). Baldi has also been used effectively to teach vocabulary to children with hearing loss (Barker, 2003; Massaro et al., 2003; Massaro and Light, 2004a). The same pedagogy and technology has been employed for language learning with autistic children (Bosseler and Massaro, 2003).

While Baldi's facial and tongue animation probably represent the state of the art in real-time visible speech synthesis, experiments have shown that Baldi's visible speech is not quite as effective as its human counterpart is (Massaro, 1998, Chapter 13). Preliminary observations strongly suggest that the specific segmental and prosodic characteristics are not defined optimally. One of our continual goals, therefore, is to significantly improve Baldi's communicative effectiveness. In this paper, we present our work to extend the capability of Baldi to be multilingual. We begin with an overview of facial animation and visible speech synthesis. Then, we present the general scheme to make Baldi multilingual using a client/server architecture. Finally, we present our perceptual evaluation of the Arabic version of the multilingual talking head and how this evaluation was used to improve its articulation.

2. Facial animation and visible speech synthesis

Visible speech synthesis is a sub-field of the general areas of speech synthesis and computer facial animation (Massaro, 1998, Chapter 12, organizes the representative work that has been done in this area). The goal of the visible speech synthesis at PSL has been to develop a polygon (wireframe) model with realistic motions (but not to duplicate the musculature of the face to control this mask). We call this technique, terminal analogue synthesis because its goal is to simply duplicate the observable articulation of speech production (rather than illustrate the

underlying physiological mechanisms that produce it). This method of synthesizing visible speech has also proven successful in the parameter formant synthesis of audible speech (Klatt, 1987). One advantage of the terminal analogue synthesis is that calculations of the changing surface shapes in the polygon models can be carried out much faster than those for muscle and tissue simulations (e.g., Kähler et al., 2001). For example, our software can generate a talking face in real time on a commodity PC, whereas muscle and tissue simulations are usually too computationally intensive to perform in real time (Massaro, 1998, 2004; Massaro et al., in press). More recently, image synthesis, which joins together video images of a real speaker, has been gaining in popularity because of the realism that it provides (Blanz et al., 2003; Bregler et al., 1997; Ezzat and Poggio, 1998; Ezzat et al., 2002). However, these systems are also computationally intensive and they do not permit viewing from a new perspective or to view the inside of the mouth. Finally, performance-based synthesis that tracks a live talker (e.g., Guenter et al., 1998) does not have the flexibility of saying anything at any time in real time, as does our text-to-speech system.

Our own current software (Cohen and Massaro, 1993; Cohen et al., 1996, 2002; Massaro, 1998) is a descendant of Parke's software and his particular 3-D talking head (Parke, 1975). We have increased the resolution of the model, modified and added additional control parameters, allowed asymmetric facial movements, trained a complex tongue (which was lacking in Parke's model), implemented a coarticulation algorithm, and added controls for paralinguistic information and affect in the face. Baldi can either be aligned with natural speech or controlled by text-to-speech synthesis to generate bimodal (auditory/visual) speech. Most of the control parameters move vertices (and the neighboring 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 sub-areas. 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 synthetic tongue is constructed of a polygon surface defined by sagittal and coronal b-spline curves (see Fig. 1). The control points of these b-spline curves are controlled singly and in pairs (see Fig. 2) by speech articulation control parameters. For example, the tip of the tongue can be extended by moving the E and F pair of points to the right, and the tongue tip thickness can be modified by adjusting the distance between these two points. There are now 9 sagittal and 3 sets of 7 coronal parameters (for tongue front middle and back) that are modified to mimic natural tongue movements. The tongue, teeth, and palate interactions during speaking require an algorithm to prevent the tongue from going into rather than colliding with the teeth and palate. To ensure this, we have developed a fast collision detection method to instantiate the appropriate interactions. Two sets of observations taken from real talkers have been used to inform the appropriate movements of the tongue. These include (1) three-dimensional ultrasound measurements of upper tongue surfaces and (2) EPG data collected from a natural talker using a plastic palate insert that incorporates a grid of about a hundred electrodes that detect contact between the tongue and palate at a fast rate (e.g. a full set of measurements 100 times per second). Maureen Stone, at John Hop-

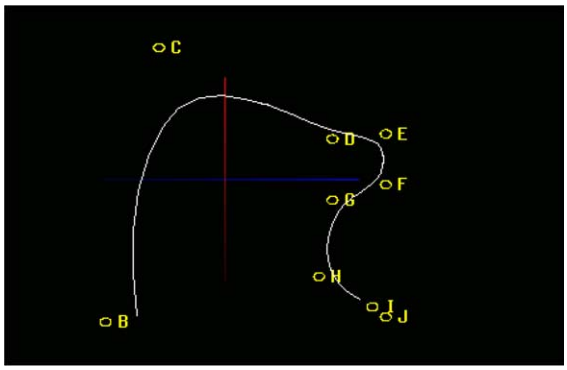

Fig. 2. Mid-sagittal b-spline curve of the synthetic tongue. The lettered circles give the locations of the curve control points.

kins University, provided these measurements. Optimization routines are used to create animated tongue movements that mimic the observed tongue movements by minimizing the difference between real and synthetic geometries (Cohen et al., 1998). One possible application of a realistic tongue is for second language learning, where presenting the inner articulation for phonemes of the second language compared with the native language phoneme may help to improve the production of those phonemes (Massaro and Light, 2003, 2004b).

We have used phonemes as the basic unit of speech synthesis. In this scheme, 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, rounding, 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 by a process called coarticulation. Coarticulation is based on a model of speech production using rules that describe the relative dominance of the characteristics of the speech segments (Cohen and Massaro, 1993; Massaro, 1998, Chapter 12). For each facial control parameter of a phoneme, there are also temporal dominance functions dictating the influence of that phoneme over the control parameter. These dominance functions determine independently for each control parameter how much weight its target value carries against those of neighboring phonemes, which will in turn determine the final parameter control values.

3. Multilingual talking head

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. For instance, we can ask how does the articulation of the same phoneme differ across different languages, how informative and influential is visible speech across different languages, and how important is the visible information for perception of non-native speech? On the application side, we can ask how visible speech can facilitate the learning of new languages.

Our previous method used to introduce new languages like Spanish and Italian implemented them within the talking head system (Massaro et al., 2000; Cosi et al., 2002). Because our original goals did not include multilingual synthesis, the software that was written had several limitations. For example, there were only a limited number of symbol entries available for storing unique phonemes and their corresponding control parameters and dominance functions. This is a serious limitation if we deal with many different languages, which would require an almost unlimited number of phoneme definitions. In addition, the software became excessively complex as new languages were added. In this paper, we present a new platform to extend the capabilities of Baldi to speak a variety of other languages than English. We use parsimonious client/server architecture to impose an independent implementation, modification, and application of the auditory speech module and the visual speech synthesis module. This scheme enables an efficient extension of text-to-speech (TtS) synthesis and facial animation to many different languages with minimal development effort. For example, no modification of the client is necessary when a new TtS engine is added to the server.

The new software simplifies the extension to other languages. This allows the development of the phoneme set and the corresponding target and coarticulation values to allow synthesis of several other languages. These include Spanish (Baldero), Italian (Baldini, Cosi et al., 2002), Mandarin (Bao), Arabic (Badr), French (Baladin), and German (Balthasar). (The 10 most frequently spoken languages in the world are in descending order Chinese Mandarin, English, Spanish, Bengali, Hindi, Portuguese, Russian, Arabic, Japanese, and German.)

A defining characteristic of our research is the empirical evaluation of the visible speech synthesis based on recognition by human observers. These experiments are aimed at evaluating the intelligibility 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. In this paper, we present the evaluation results for Arabic Badr, and compare them to analogous results for English Baldi. In the following section, we present the general scheme for our multilingual talking head.

3.1. General scheme for a multilingual talking head

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

3.1.1. Auditory

For any new language to be added to the talking head system, a text-to-speech (TtS) engine capable of producing that language is required (or a database of natural speech that has been segmented and phonemically labeled). 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. Because TtS engines are developed by different groups, they do not provide exactly the same output phonemes, they may require different input formats, and they may use different output formats (even if they use a standardized interface as SAPI architecture (Huang, 1998), for example). This means that a unique interface definition is required to deal with different TtS inputs and outputs. It is also possible to have a corpus of natural auditory speech that has been segmented and phonetically labeled. Given that the facial animation system requires the phonemes and their durations, either the TtS system or the natural speech corpus can drive the animation and be aligned with it.

3.1.2. Visual

By default, Baldi speaks English. To have Baldi speak a new language accurately, new phonemes will be required as well as revised definitions for the existing phonemes (Lindau and Ladefoged, 1983). For each phoneme in a specific language, target values must be defined for each of the animation control parameters and the dominance functions (coarticulation parameters) must be specified. For this purpose, we have developed a set of graphical editing tools that allow the user to view and adjust the control parameters and their dynamic behavior (Fig. 3). These tools allow us to define new visual language-specific parameters for the phonemes and their corresponding dominance functions. The changes in the parameters can be based on either real physical measurements from a speech corpus to train the face (Cohen et al., 2002) or more qualitative analyses of the articulation of the target language (Massaro, 1998, Chapter 13).

We improved the existing Baldi software by reorganizing the system and making it more modular. The improved system uses a client/server architecture in which the server handles the auditory speech dimension and the client deals with the visual animation dimension. In addition to the common benefits of modular software, the architecture allows the system to be 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 or to a natural speech corpus (Fig. 4).

3.1.3. The server

To synthesize many languages, we require a variety of different TtS engines. The *PSL TtS Server* is a unique interface between the different TtS engines and the talking head. To include a TtS engine in our system, the minimum set of requirements is that the TtS engine provide the phonemes, their durations, and the synthesized auditory speech. The server can support many TtS engines for one or more languages (for example, Mandarin Chinese using Lucent TtS (Sproat, 1997); English using Festival TtS (Taylor and Black, 1997) or AT&T natural voices (Beutnagel et al., 1999); Arabic, German and French using Euler/Mbrola TtS (Bagein et al., 2000) etc.). 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

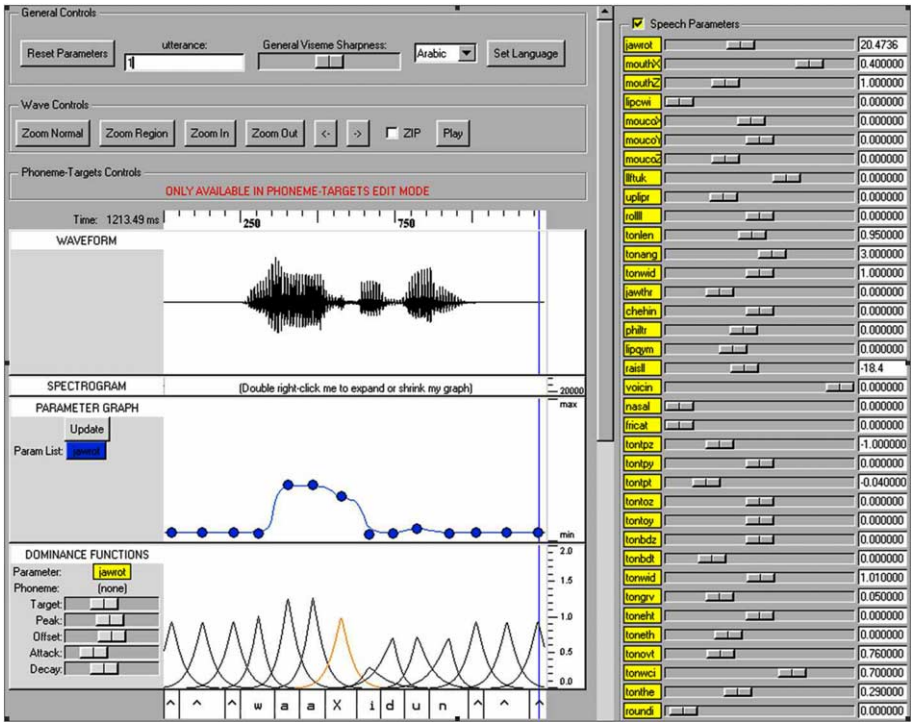

Fig. 3. A screenshot of the tool used to edit and redefine the articulation of the talking head. To the right, we have the control parameters panel to adjust the position of each articulatory control parameter. At the bottom left we have the controls for the refining the phoneme definition in terms of target values, dominance strength, dominance peak time offset, and attack and decay rate of a dominance function. These settings are independent for each control parameter and phoneme. We also display the speech waveforms, parameter tracks over time, and the dominance functions that produce the parameter tracks.

independent. It is also possible to send inquiries and instructions to the server (for example, what are the installed TtS engines, what are the installed voices, is a voice female or male, 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. For example, the text might contain a command for changing the emotion of the face or the rate of speaking, and this information would be sent 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 server can handle many clients running at the same time on the same machine or on remote machines. Any client connected to the server receives the phonemes, their durations, and the speech wave file corresponding to the text sent by the client to the server.

3.1.4. The client

The client is an application built around the talking head. This application can be a simple graphic user interface (GUI) that allows the user to simply enter a text or include more detailed control of the talking head and of the speech (slowing down the speech, specifying a face configuration and viewpoint, adding emotion and gesture, etc.). Basically, the client sends the text to be uttered to the server and the server sends back all of the information required by the client. At a minimum, the information includes the phonemes and their durations to animate the visual phonemes and synchronize them with the auditory speech. If it is provided by the TtS, the server can send word boundaries, pitch information and any

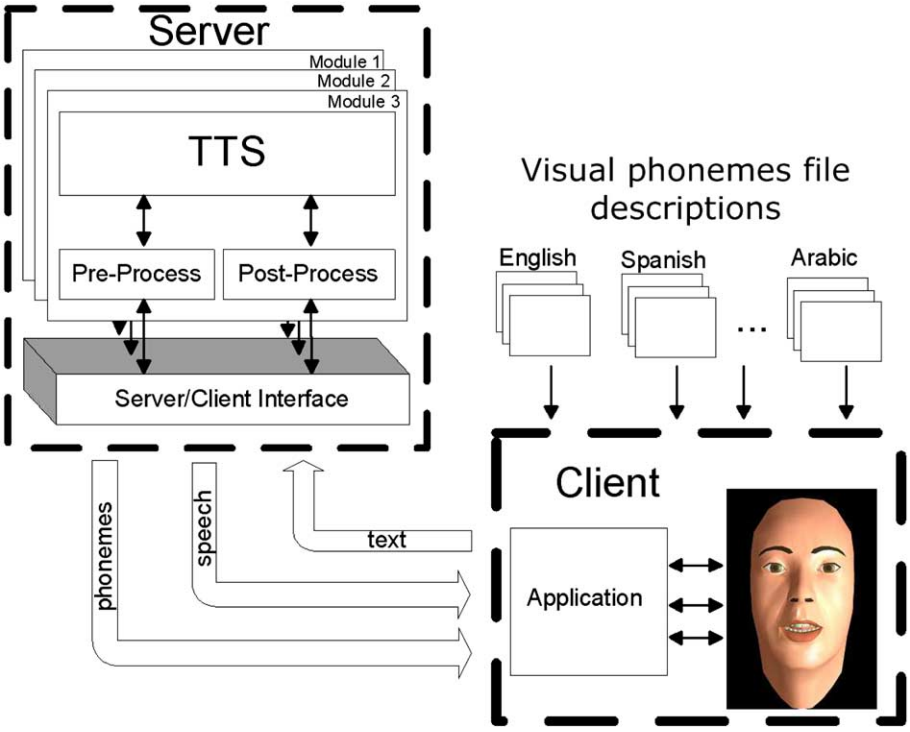

Fig. 4. Client/server architecture system. The server is a unique interface between different TtSs and the client (talking head). Each TtS has a corresponding module in the server that pre-processes the input and post-processes the output. The client is an application built around the talking head. The visual phonemes used for the talking head are defined in a description file. Client sends the text to the server, and in return, the server sends back the speech corresponding to the text and the phonemes with their duration.

XML-markup information to the client. This information can be used to control eyebrow movement (based on pitch information) and to control eye blinking (based on word boundaries).

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 server itself does not need a lot of resources, but depending on the number of TtS engines installed in memory, it may require a large memory (some TtS engines have a huge footprint).

3.2. Application of the new scheme

We applied this new architecture to extend the capabilities of Baldi to speak new languages. In addition to English, we have extended Baldi to speak Arabic, German, French, Chinese Mandarin, Italian, Castilian Spanish, Mexican Spanish, Brazilian, Portuguese and Swedish. 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. No modification of the client is required when a new TtS engine is added to the server.

In addition to *PSL TtS server*, we developed client software *Bapi* that shows a typical use of the client/server system. The client has a simple user interface. The user selects a language from a menu of languages. A list of available voices is then displayed for that language, from which the user can select a particular voice. The user can type some text or an existing file can be entered, and the

talking head will say it. *Bapi* has many other useful features that are not germane to the goals of this paper.

Our goal is to create realistic and intelligible visible speech for all languages and to date, the most complete recent work has been done for Arabic. In the following sections, we illustrate our method to define the articulation and coarticulation for each phoneme for this language.

4. An Arabic talking head: Badr

4.1. Description of Arabic articulation and coarticulation

There are 34 phonemes in Arabic language: 28 consonant phonemes and 6 vowels (see Table 1). Three vowels are short in duration (/æ/,/i/ and /u/) and three are long (/ææ/, /ii/ and /uu/).

An important articulatory feature of Arabic is the presence of pharyngealized and pharyngeal phonemes (Fig. 1). There are two pharyngeal fricatives (/ħ/ and /ʕ/). These phonemes are characterized by the constriction formed between the tongue and the lower pharynx in addition to the rising of the larynx. There are three uvulars (/x/, /ʁ/ and /q/) characterized by a constriction formed between the tongue and the upper pharynx for /x/ and /ʁ/ and a complete closure for /q/ at the same level. These five consonants are considered pharyngeal phonemes. In addition, there are four pharyngealized or emphatic phonemes: /sˤ/, /dˤ/, /tˤ/ and /ðˤ/. These phonemes are a pharyngealized version of the oral dental consonants /s/, /d/, /t/ and /ð/.

The main characteristic of the pharyngealization is the rearward movement of the back of the tongue. Thus, the vocal tract shape presents an increased oral cavity and a reduced pharyngeal cavity because of the retraction of the body and the root of the tongue toward the back wall of the pharynx (Al-Ani, 1970; Ghazali, 1977; Jakobson, 1962). The pharyngealized consonants also induce a considerable backing gesture in neighboring segments, which occurs primarily for the adjacent vowels (the pharyngealized consonants affect the neighboring vowels in such a way that they will be transformed in their pharyngealized version. For instance: /æ/ → /ɑ/, /i/ → /ɨ/, /u/ → /ʉ/ and for the long vowels: /ææ/ → /ɑɑ/, /ii/ → /ɨɨ/, /uu/ → /ʉʉ/).

Table 1
Panel (a): The 28 consonant phonemes of the Arabic language (The first line of each cell presents the IPA symbol and the second line presents its grapheme.); Panel (b): IPA symbols for the six vowels of the Arabic language and their generally used graphemes

Panel (a)

ʔ أ	b ب	t ت	θ ث	ʒ ج
ħ ح	x خ	d د	ð ذ	r ر
z ز	s س	ʃ ش	sˤ ص	dˤ ض
tˤ ط	ðˤ ظ	ʕ ع	ʁ غ	f ف
q ق	k ك	l ل	m م	n ن
h ه	w و	j ي		

Panel (b)

æ ـَ	ææ ـَا	i ـِ	ii ـِي	u ـُ	uu ـُو

Pharyngealization is both an intrasyllabic and intersyllabic phenomenon (Ali and Daniloff, 1972; Ghazali, 1977). The coarticulatory effect of pharyngeal and pharyngealized consonants can affect just a single syllable or several. It is not easy to determine the extent of the coarticulation effect of the pharyngealized and pharyngeal phonemes on their neighboring consonants and vowels. For example, in the word مطار (*airport*) /mɑtˤɑɑrun/, /m/ is pharyngealized because of the following vowel /ɑ/ which was pharyngealized because of the anticipatory coarticulation effect of /tˤ/. The pharyngealization is observed in the pharyngealized

consonants /sˤ/, /dˤ/, /tˤ/ and /ðˤ/ in all of the Arabic dialects. However, for pharyngeal phonemes, pharyngealization varies from one dialect to another. The pharyngealization may also affect /l/, /r/ and /j/ in certain instances. As expected, researchers are not unanimous about the properties of these pharyngeal and pharyngealized phonemes in Arabic and its various dialects, their effects on other segments, and the mechanism used for pharyngeal consonant production (Al-Ani, 1970; Elgendy, 2001; Ghazali, 1977).

4.2. Description of the Arabic TtS

For the Arabic language, we used the non-commercial multilingual text-to-speech system Euler that uses Mbrola for the output stage (Bagein et al., 2000). The latter uses 34 + 6 phonemes: 28 consonants and 6 + 6 (their pharyngeal counterparts) vowels. The pharyngealized consonants (/sˤ/, /dˤ/, /tˤ/ and /ðˤ/) are usually followed by pharyngealized vowels (/ɑˤ/, /iˤ/, /uˤ/, /ɑɑˤ/, /iiˤ/ and /uuˤ/). As a result, the pharyngealized consonant in a CV sequence is followed by a pharyngealized vowel. This was the only kind of pharyngealization phenomenon that was implemented in the Arabic module for Euler. Although the prosody does not sound entirely natural, Euler/Mbrola generally provides fairly good quality Arabic speech. Similar to other synthesizers, however, the Arabic Euler/Mbrola module presented some phonemes lacking of naturalness as in /r/ and /q/. This unnaturalness was also noticed by some of the participants (see Section 4.3).

4.3. Modeling Arabic visual phonemes

In the absence of detailed measurements of the Arabic articulation and coarticulation, we used a variety of resources to define the visual phonemes. We present two successive developments of the talking head: a preliminary version (*version 1*) and an improved one (*version 2*). For version 1, we roughly mapped the Arabic phonemes into their English closest ones. This preliminary version was used to identify which phonemes required more extensive modifications than others based on the first perceptual evaluation study. We then used the results of the evaluation made on version 1 along with other resources to create version 2, as is explained in the following section.

4.3.1. Version 1: Badr's visual phonemes Baldi's English visual phonemes

In this preliminary study, we began by mapping the Arabic phonemes into similar English ones, as the latter have already been defined and evaluated for English (Cohen et al., 2002). For some other phonemes, we used the English phoneme that was closest visually as presented in Table 2. For the phonemes /b/, /t/, /θ/, /ʒ/, /d/, /ð/, /r/, /z/, /s/, /ʃ/, /f/, /l/, /m/, /n/, /h/, /w/ and /j/, we used the corresponding English phonemes. For the phonemes without a corresponding one in English (/ʕ/, /ħ/, /x/, /ɣ/ and /ʁ/), we used roughly the closest English viseme which is /h/. For the pharyngealized phonemes we used the non-pharyngealized version of those phonemes: /sˤ/ → /s/, /dˤ/ → /d/, /tˤ/ → /t/, /ðˤ/ → /ð/. Finally, for vowels we used the three Arabic vowels, short and long: /æ/, /i/, /u/, /ææ/, /ii/ and /uu/ based on the visual phoneme definition of the corresponding English vowels (/æ/, /i/ and /u/). The same vowels were used for the pharyngealized as for the non-pharyngealized consonants. In conclusion, our first version is basically mapping the Arabic visual phonemes into corresponding English visual phonemes (see Table 3).

This rough and fast definition of the Arabic visual phonemes was carried out to establish a performance baseline of how much a synthetic talking head can facilitate the recognition of Arabic speech in noise. Although this method is not ideal, it reduced the amount of time required to implement a working system. We could then conduct a perceptual experiment and then use the results to improve the visible speech synthesis. Improving the synthesis based on perceptual

Table 2
Pharyngeal and pharyngealized phonemes in Arabic

Pharyngeal fricatives	/ħ/ , /ʕ/
Uvulars	/x/, /ʁ/, /q/
Pharyngealized phonemes	/sˤ/ , /dˤ/ , /tˤ/ , /ðˤ/

Table 3
The mapping between Arabic and English phonemes used in the version 1 of the Arabic talking head, Badr

Arabic phonemes	English phonemes
/b/, /t/, /θ/, /ʒ/, /d/, /ð/, /r/, /z/, /s/, /ʃ/, /f/, /l/, /m/, /n/, /h/, /w/ and /j/	/b/, /t/, /θ/, /ʒ/, /d/, /ð/, /r/, /z/, /s/, /ʃ/, /f/, /l/, /m/, /n/, /h/, /w/ and /j/
/ʔ/, /ħ/, /x/, /ʕ/ and /ʁ/	/h/
/sˤ/, /dˤ/, /tˤ/ and /ðˤ/	/s/, /d/, /t/ and /ð/
/ɑ/, /i/, /u/, /ɑɑ/, /ii/, /uu/	/æ/, /i/, /u/, /ææ/, /ii/, /uu/

experiments is a strategy that has been used successfully in English (Massaro, 1998, Chapter 13) and can be productive especially when objective measures of natural articulation are not available. The results of the recognition experiment will highlight which phonemes need improvement. Furthermore, successive perceptual experiments and modifications ideally allow the eventual evolution of completely accurate synthetic visible speech.

To evaluate Version 1 of Badr, the first perceptual experiment used the multilingual client/server system to generate a set of 100 Arabic utterances consisting of three words each. The words were 2–5 syllables in length. A typical example utterance is shown in Fig. 12. These common and familiar words were spoken by the TtS synthesizer in classical Arabic and aligned with Badr. Arabic text of each utterance was sent to the TtS server, which sent back to the client a sequence of phonemes and their corresponding durations, in addition to a wave file of the auditory speech. The auditory speech was produced by the Euler/Mbrola TtS, using the voice AR2 (Arabic Male). The sequence of phonemes and their corresponding durations in addition to the speech wave file were used by the client to produce the synthetic face animation. The same wave files were used for both unimodal auditory and bimodal conditions. White noise was added to these utterances to make the words somewhat difficult to recognize.

There were 12 participants in the experiment, all native Arabic speakers: 3 Moroccans, 1 Syrian and 8 Tunisians. All participants were graduate students living in France, but they had lived and finished their undergraduate degrees in their native countries.

Experimental participants were presented with each of these 100 utterances under two different conditions: auditory-only or in a bimodal condition aligned with Badr. The 100 sentences were randomly presented under the auditory and bimodal conditions for a total of 200 trials. Participants had a break of 5 min after 100 trials. The instructions to the participants were to write down as many words as they could for each utterance on each trial. After finishing the experiment, the participants' answers were scored by the senior author.

The participants viewed a 15 in. flat LCD screen of a PC equipped with a P-III 450MHZ, NVIDIA Geoforce 2 graphics card. The auditory speech embedded in white noise was presented over headphones at a normal listening level. The white noise was generated in advance and saved in a wave file. Then it was played simultaneously while playing the speech in both conditions.

4.3.2. Results of experiment 1

Fig. 5 and Table 4 give the percentage of correct word answers for the unimodal (auditory alone) and bimodal (audiovisual synthetic face) condition for the twelve participants. The unimodal condition averaged 34% correct, which increased 16 percentage points to 50% when the synthetic face was added.

The goal of this experiment was to use the results to improve the Arabic visible speech. For this purpose, a more detailed phoneme analysis was performed to assess performance on particular speech segment. All of the phonemes in a word were scored as correct if the word was reported correctly, whereas all the phonemes were scored

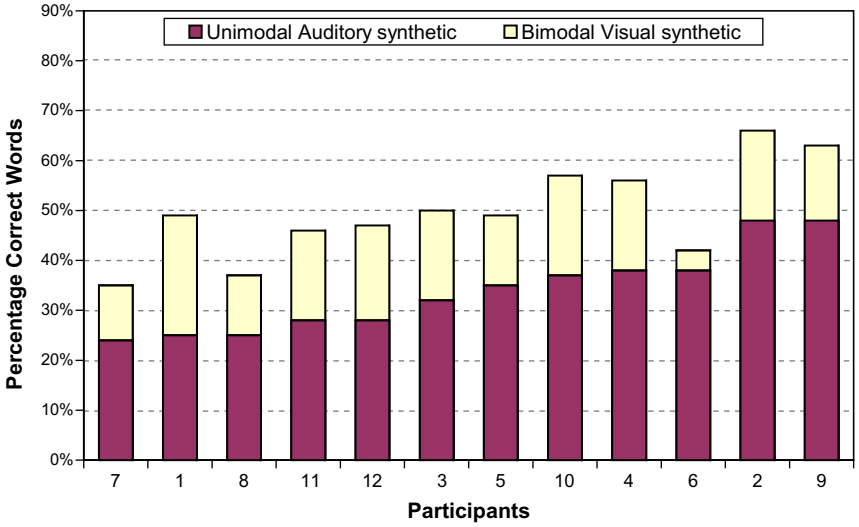

Fig. 5. The percentage of correct word recognition as a function of the unimodal auditory and bimodal conditions. Results from Experiment 1.

Table 4
The individual correct word recognition results of Experiment 1 (version 1) as a function of the unimodal auditory condition and the bimodal synthetic face condition

Participant#	Auditory (%)	Synthetic face (%)
1	25	49
2	48	66
3	32	50
4	38	56
5	35	49
6	38	42
7	24	35
8	25	37
9	48	63
10	37	57
11	28	46
12	28	47
Mean	34	50

as incorrect if the word was not reported. Fig. 6 presents these phoneme recognition results for the unimodal and bimodal conditions. As can be seen in the figure, there is a large range in performance across the different phonemes but that the synthetic face usually improved recognition performance. Fig. 7 presents percentage of improvement added by the visual facial information to the auditory condition for each phoneme.

Fig. 7 shows that the amount of improvement given by the face varied significantly across the different phonemes. For six phonemes (/ɑɑ/, /ɨɨ/, /θ/, /k/, /z/ and /ʒ/), we have an improvement that exceeds 30%. The nine phonemes: /l/, /h/, /s/, /m/, /i/, /x/, /f/, /j/ and /dˤ/ present an improvement between 20% and 30%. These relatively large improvements might be best explained by the fact that these phonemes are visually close to their English counterparts. The 11 phonemes: /ææ/, /w/, /u/, /ʃ/, /d/, /æ/, /n/, /b/, /ʁ/, /ʔ/ and /ð/ have an improvement between 10% and 20%. For the remaining phonemes, the improvement does not exceed 10%.

Four phonemes did not show any improvement at all, but actually decreased recognition when presented in the bimodal condition: /ʉ/, /ɨ/, /ɑ/, /r/. The reason is that the visible speech of the English phonemes was a poor match to that appropriate for Arabic. In the English /r/, the tongue is not as visible as it is in Arabic. In /ɑ/, the mouth is not as open as much as it should have been for Arabic. The poor results of all the varieties of the sound /u/ (/u/, /uu/, /ʉ/, /ʉʉ/) might be explained by the fact that the lip rounding for /u/ is more prominent in Arabic than it is in English.

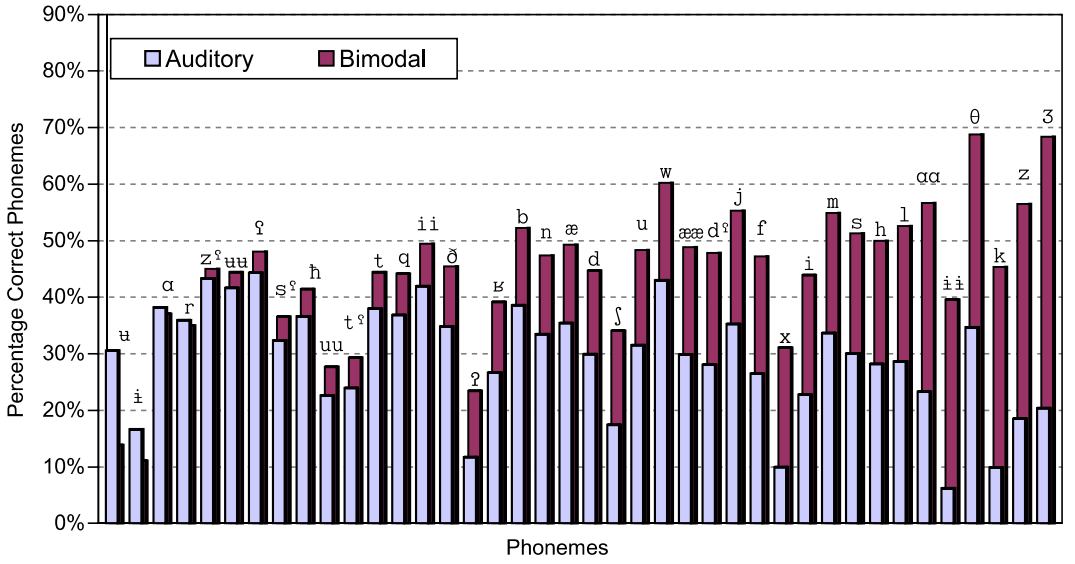

Fig. 6. The correct phoneme recognition results of Experiment 1. For each phoneme, we present the results in the unimodal auditory condition and in the bimodal synthetic face condition.

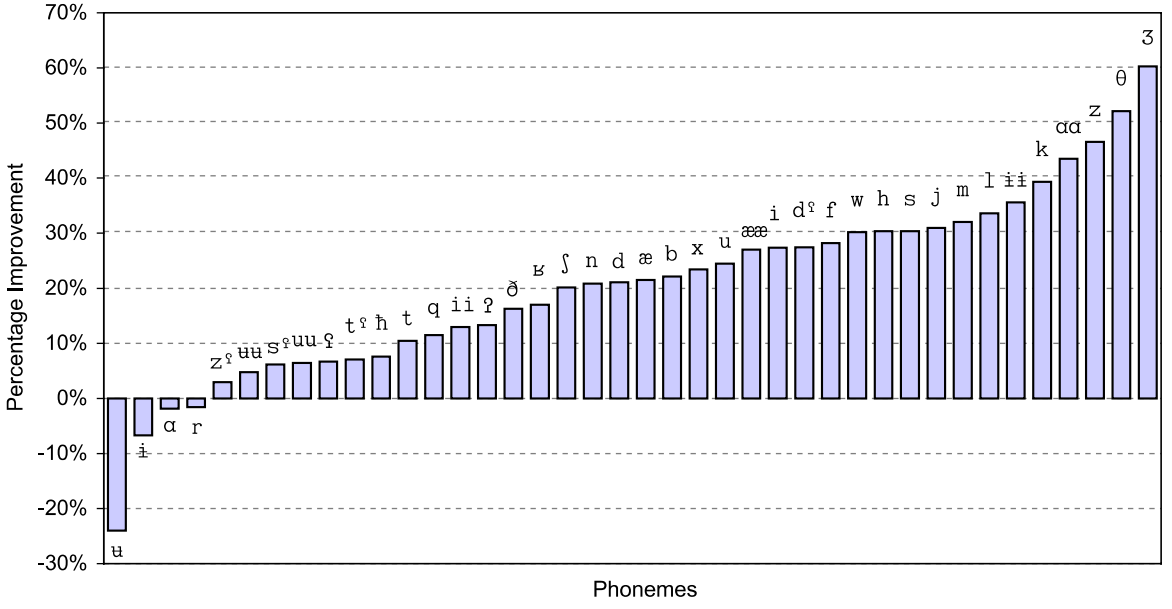

Fig. 7. Improvement of phoneme recognition, given by difference between the accuracy on the bimodal synthetic face and the accuracy on the unimodal auditory condition for Experiment 1.

Given these results, the next step was to improve the articulation of each phoneme. We gave particular attention to the phonemes that gave less than 10% improvement: /ʉ/, /ɨ/, /ɑ/, /r/, /ðˤ/, /ʉʉ/, /ʕ/, /sˤ/, /ħ/, /uu/, /tˤ/, /t/, /q/, /ii/. Many of these phonemes are specific to the Arabic language, and only roughly approximated by their English equivalents. In addition, we were expect-

ing that the pharyngeal and pharyngealized phonemes would provide poor results because their main characteristic, the backing of the tongue that is probably partially visible, was not synthesized. Notwithstanding these limitations, the visible speech was still somewhat effective showing that the partially seen articulation may provide informative clues. As one possible application of this multilingual talking head is speech training (Massaro and Light, 2003, 2004b), we paid attention to the articulation that takes place in the back of the oral cavity as in /ʕ/, /q/ and /ħ/ even though it might not be normally visible. However, the skin of a synthetic face can be made partially transparent or eliminated completely to show the inner articulation. Viewing this inner articulation might be very helpful in learning speech production. For second language learners, for example, one important difficulty is that it is hard for them to produce some phonemes correctly by just listening to the speech. Having the opportunity to see how production occurs inside the mouth can be very helpful to improve their pronunciation (Massaro and Light, 2003, 2004b). There were 4 phonemes (/ʁ/, /ɨ/, /ɑ/ and /r/) that needed particular attention for improvement, since the visual information did not improve performance relative to just the auditory condition.

4.3.3. Version 2: Improvement of the Arabic visual phonemes

We used the analysis of experiment 1 and the results found in the literature on Arabic phonetics and articulation (Al-Ani, 1970; Gairdner, 1925; Ghazali, 1977) to improve the visible speech animation of the Arabic phonemes. We used our graphical editing tool to define the phonemes. This tool allowed the user to view and adjust the control parameters and their dynamic behavior. In addition, it provides control of each articulatory parameter and the opportunity to see the face and inner articulation from different views (for example, it is possible to focus on the modeling of the front view, or on the tongue by making the skin semitransparent. It is also possible to see a top view of the mouth to define the contact between tongue and palate).

The Arabic visual phonemes were characterized by both the visible face and partially visible tongue. The literature provided mainly midsagittal views of the vocal tract (positions of the tongue relatively to the palate and teeth), and frontal views of some phonemes. Figs. 8–10 show examples from the resources we used to define the phoneme prototypes. Some of the data were tracings of X-ray images (Fig. 8) and facial photos (Fig.

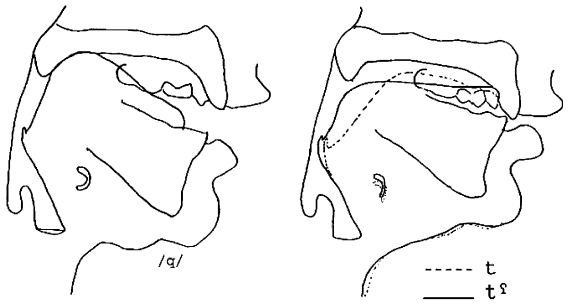

Fig. 8. Two examples of data used to define the Arabic phonemes. These are tracings of X-ray data (Ghazali, 1977).

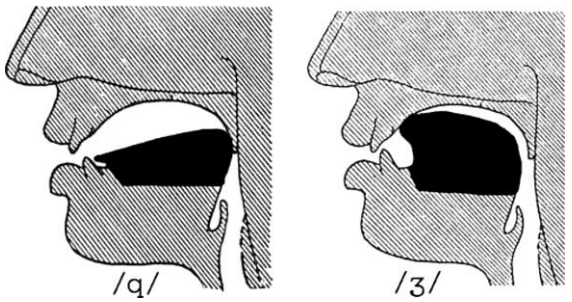

Fig. 9. Two examples of the images used to define some of the Arabic phonemes. A prototype for /q/ and for /z/ (Gairdner, 1925).

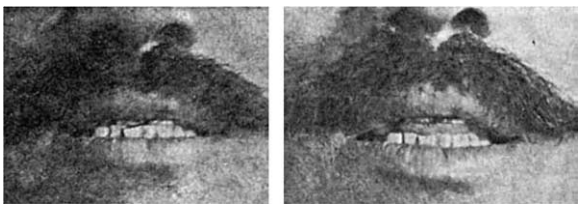

Fig. 10. Facial images used for defining some phonemes. To the left we have, the phoneme /s/ and to the right the phoneme /t/ (Gairdner, 1925).

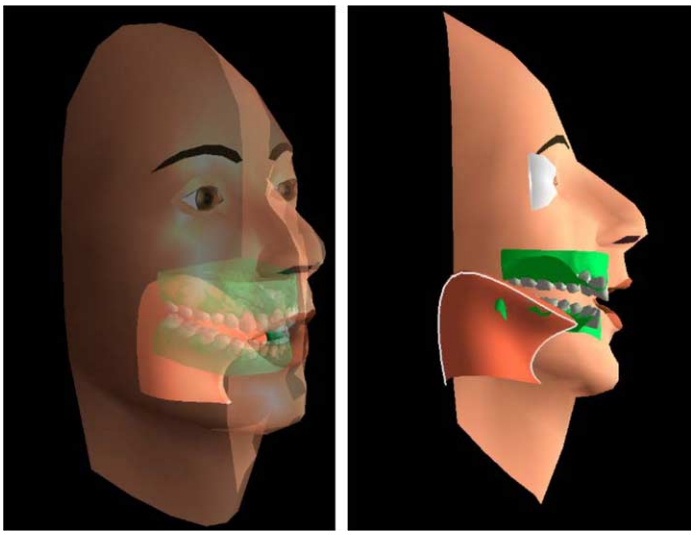

Fig. 11. Semi-transparent view and mid-sagittal view of the talking head. This is an example of a phoneme prototype defined for the Arabic phoneme /q/, based on some of the data found in literature.

10). We used 18 tracings of X-ray images found in (Ghazali, 1977). We also made a video recording of a native Arabic speaker reading a list of words that we analyzed to improve the articulation of the visible speech synthesis. For the phonemes that gave low performance, we chose several utterances containing those phonemes, and we compared the natural speaker side by side with the synthetic talking head Badr. Badr's control parameters for these phonemes were successively modified to give a good match the natural face articulation. This video proved helpful in refining the articulatory definition of the lips (lip rounding in /w/ and /u/ for example) and improving the articulation of the partially seen tongue (for example, the video showed that /t/ and /d/ are dental plosives: that is to say, the tip of the tongue touches the upper teeth themselves, which can often be seen through the teeth).

We also used some information about coarticulation (Al-Ani, 1970; Elgendy, 2001; Ghazali, 1977), which basically involved several aspects of pharyngealization. In this version, we modified the pharyngealization influence of the consonants /ħ/, /ʕ/, /x/, /ʁ/, /q/ and /r/ in addition to /sˤ/, /dˤ/, /tˤ/ and /ðˤ/. When these phonemes occur in Arabic, the following vowel is replaced by its pharyngealized counterpart (see Section 4.1). The dominance functions of the tongue control parameters were modified to implement this influence for these consonants.

Fig. 11 shows an example of the newly defined phoneme /q/, which is based on the /q/ presented in Figs. 8 and 9. The phoneme /q/ is a voiceless uvular stop that has a complete closure between the tongue dorsum and the uvular. The alveolar trill /r/ is defined as a rapid succession of taps of the tip of the tongue against the teeth ridge. In this version of the talking head, we used two taps of the tongue for the Arabic /r/.

5. Evaluation of the Arabic talking head

The recognition 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. This experiment was very similar to the first experiment. One major difference between the first experiment and this experiment is that we added the natural face as an additional test. This control condition is valuable because we can also compare the improved performance given the talking head to that given

by the natural face. 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, which was also used for both bimodal conditions. To align the natural auditory speech to the synthetic face, the visual phonemes of each utterance were determined and then their phoneme boundaries were manually adjusted to give an accurate alignment to the waveform. We chose natural auditory speech instead of the TtS speech (which is already aligned with the visual phonemes) because the TtS cannot be accurately aligned with the natural face. In addition, this solution eliminated any errors that might be produced by the TtS. Some participants in the first experiment noticed that some auditory phonemes were not pronounced accurately (as in /q/ and /r/, for example).

The stimuli were the same set of 100 three-word utterances in classic Arabic used in experiment 1. The number of trials for each presentation condition was 100. Fig. 12 shows an example from one trial. The 100 utterances and the three presentation conditions were randomly presented for a total of 300 trials. Participants viewed the visible speech on a 17 in. screen with a visual angle of 18 degrees, generated by a P-4 and Integrated Intel Extreme Graphic card. The natural speech utterances were made into video files and the client/server system was used to generate the synthetic visible speech utterances that were also converted into video files. Thus, for both bimodal conditions, we played video files. The noise was played simultaneously while playing the auditory in unimodal and bimodal conditions. There were 19 participants tested in the experiment, all native Arabic speakers living in Tunis, Tunisia.

Fig. 13 and Table 5 present the percentage of correct words recognized under each of the three conditions: unimodal auditory, bimodal synthetic face and bimodal natural face. For the unimodal auditory condition, the average accuracy of recognized words was 30%. There was a significant

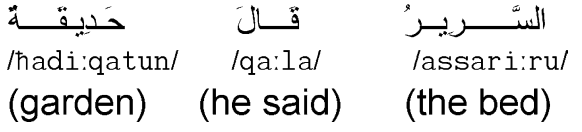

Fig. 12. Example test trial words (top) with phonetic transcription and English translations.

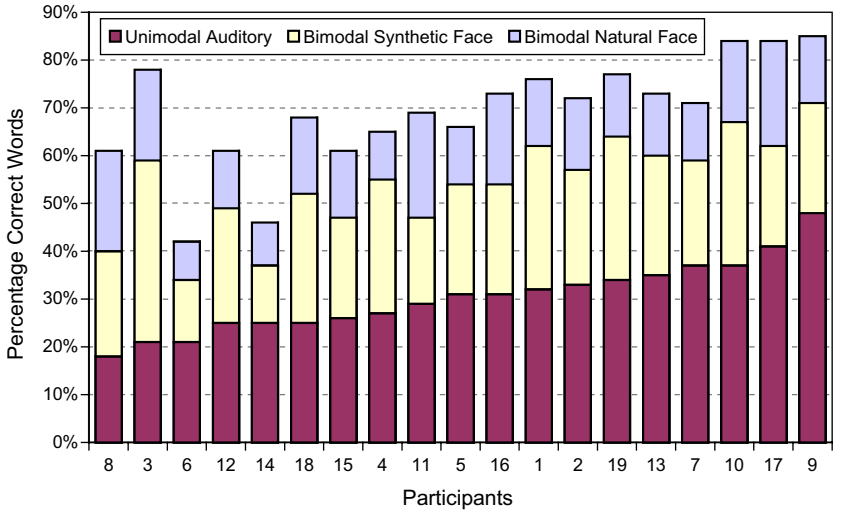

Fig. 13. The word recognition results of Experiment 2. For each of the 19 participants, we present the results in the unimodal auditory condition, bimodal synthetic face condition and in the bimodal natural face condition.

Table 5
The individual correct word recognition results of the Experiment 2 (version2) as a function of the unimodal auditory condition, the bimodal synthetic face condition, and the bimodal natural face condition

Participant #	Auditory (%)	Synthetic face (%)	Natural face (%)
1	32	62	76
2	33	57	72
3	21	59	78
4	27	55	65
5	31	54	66
6	21	34	42
7	37	59	71
8	18	40	61
9	48	71	85
10	37	67	84
11	29	47	69
12	25	49	61
13	35	60	73
14	25	37	46
15	26	47	61
16	31	54	73
17	41	62	84
18	25	52	68
19	34	64	77
Mean	30	54	69

advantage over this unimodal auditory condition when the audio is paired with either the synthetic or natural visual information. Moreover, the synthetic face showed a larger improvement relative to the first experiment. (24% versus 16%). As expected, the improvement of 24% for the recognized words in bimodal synthetic face (54%) was lower than the performance with the natural face (69%).

Comparing the results of the bimodal synthetic face to the bimodal natural face, instead of just comparing it the unimodal auditory condition is very informative. In the absence of results of the natural face, we do not have the upper bound of how much the face can contribute to recognition. For example, as shown in Table 5, participant #6 recognized 34% of the words in bimodal synthetic face condition, which can be considered very low compared to participant #9 who has the best result among all participants with 71% correct in the same condition. However, participants #6 and #9 recognized 42% and 85% of the words in bimodal natural face condition. We can therefore conclude both participants benefited from the synthetic face and both performed better with the natural face. Showing the results of the bimodal natural face defines the amount of improvement that is possible for the synthetic visible speech. Of course, we could attempt to make our synthetic face super-realistic and produce performance that is even better than the natural face. Although there are potentially valuable applications for super-realism, our immediate goal is to simulate natural speech for realistic speech synthesis.

A more detailed analysis is summarized in Figs. 14 and 15 for the results at the phoneme level. Fig. 14 presents the phoneme recognition results. Fig. 15 presents the improvement of bimodal synthetic face versus unimodal auditory. To be able to compare the improvement made in the first version with the improvement made in the second version, we used a normalized measure presented by (Sumby and Pollack, 1954) that allows the comparison of the improvement even though auditory performance in each experiment was not exactly the same. This measure is the ***R*** ratio:

$$\boldsymbol{R} = (\text{Bimodal result} - \text{Auditory result})/(1 - \text{Auditory result}) \quad (1)$$

Figs. 16 and 17 present the improvement in the two experiments using this measure, and Fig. 18 gives the results together to show the relative degree to which the synthetic speech was improved. Almost all the pharyngeal and pharyngealized phonemes have improved relative to the previous version. Some of the phonemes (/ɑɑ/, /ɨɨ/, /θ/, /ɻ/, /z/, /ʒ/, /x/, /ɦ/ and /dˤ/), however, did not improve relative to the first version (even though in the bimodal synthetic face modality these phonemes presented good results compared to unimodal auditory modality). The small differences for the phonemes /ɑɑ/, /ɨɨ/, /θ/, /ɻ/, /z/, /ʒ/ can be explained by the fact that performance was already highly accurate in version 1 and both experiments do not have exactly the same environment (synthetic speech vs. natural speech, for example). However, as can be seen in Fig. 18, /ɦ/ and /dˤ/ need further improvement.

Fig. 14 provides additional information for comparing results of the synthetic and the natural face. This is helpful as our goal is to reach the per-

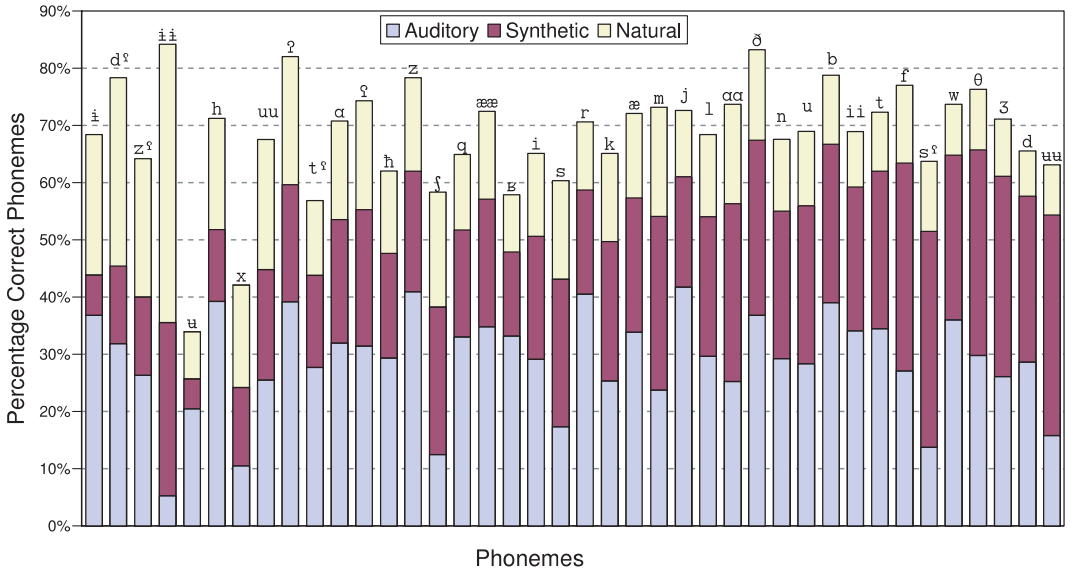

Fig. 14. The correct phoneme recognition results of the Experiment 2. For each phoneme, we present the results in the unimodal auditory condition, bimodal synthetic face condition and bimodal natural face condition.

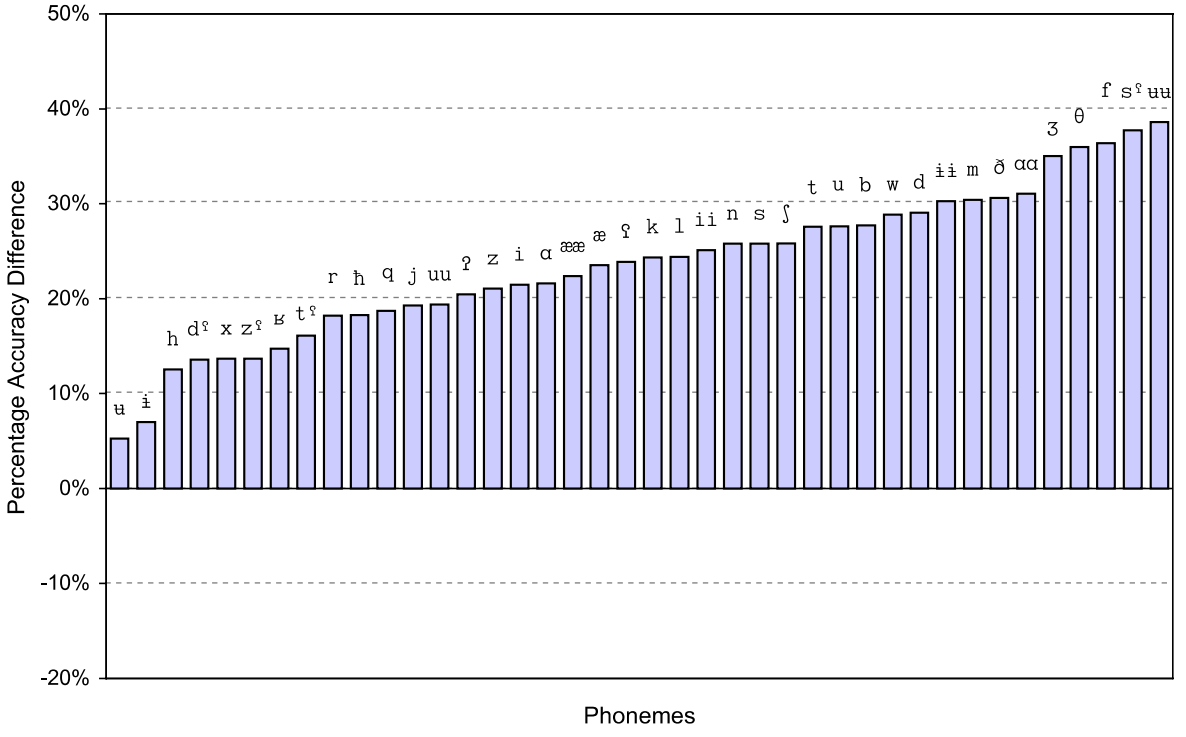

Fig. 15. Improvement of phoneme recognition, given by difference between the accuracy on the bimodal synthetic face and the accuracy on the unimodal auditory condition for Experiment 2.

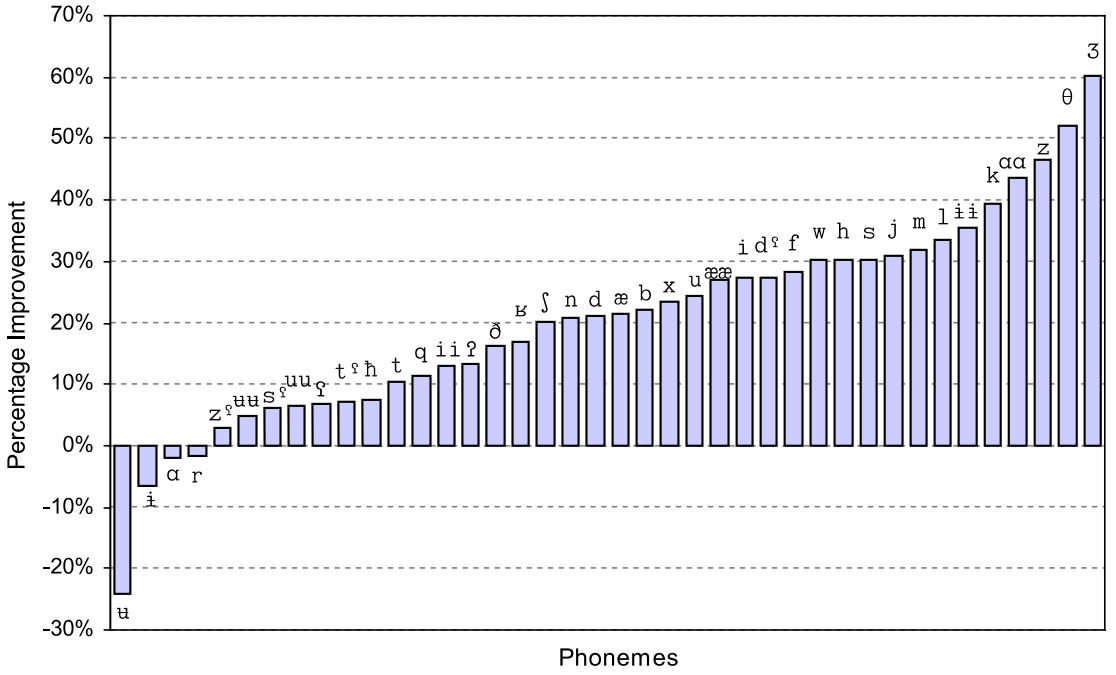

Fig. 16. Improvement of phoneme recognition independently of the performance in the auditory condition, for the Experiment 1.

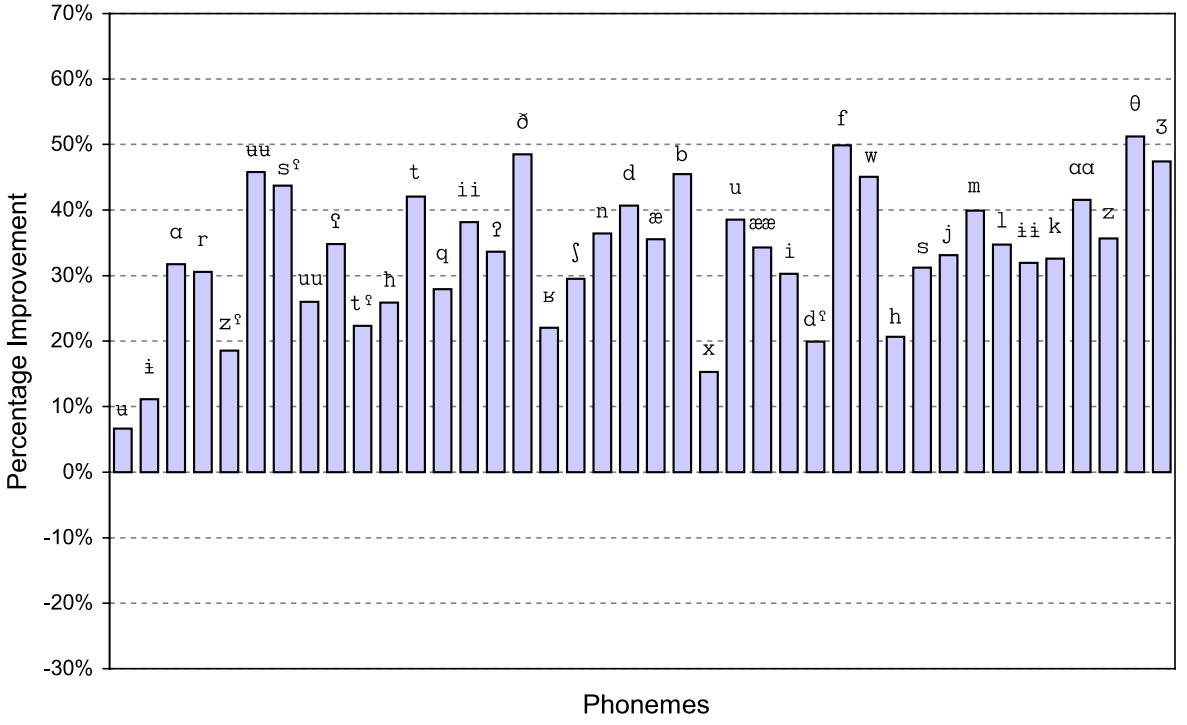

Fig. 17. The improvement of phoneme recognition provided by the synthetic face independently of the performance in the auditory condition, for Experiment 2, using the measure given in Eq. (1).

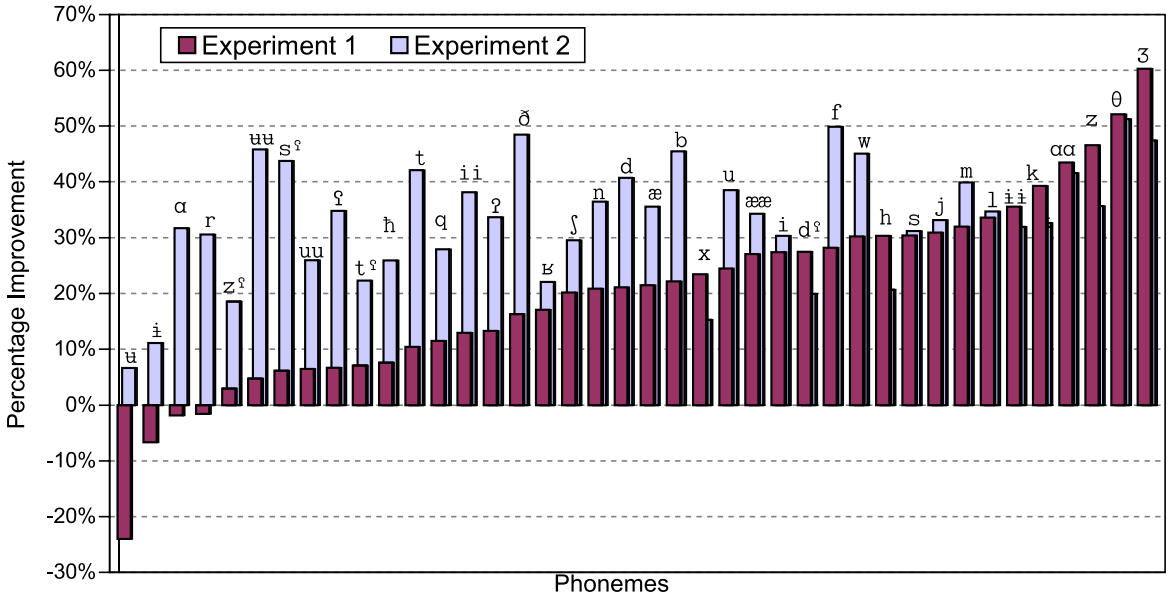

Fig. 18. Comparison of the improvement of phoneme recognition independent of performance in the auditory condition between Experiments 1 and 2.

formance of a natural face. The gap between natural and synthetic performance was reduced for many of the phonemes. For example, for /ʉ/ even though the recognition results are low for the synthetic face, we cannot expect to do much because the natural face performance is also low. In general, additional work remains to improve the visible speech synthesis. For example, synthetic /ɨ/, /ɨɨ/, /ðˠ/ and /dˠ/ presented particularly poor results relative to the natural face. Our goal is to continue improving visible speech for Arabic Badr.

6. Comparison of the performance of Baldi Vs. Badr

In this section, we compare the accuracy of Badr to Baldi to give some more objective measure of the quality of the Arabic visible speech synthesis, and whether it would be useful for applications. Baldi has been shown to be sufficiently accurate to improve the vocabulary and grammar of children with language challenges because of hearing loss or autism (Massaro et al., 2003). In addition, Baldi has been effective in improving speech production for both hard-of-hearing children (Massaro and Light, 2004b) and Japanese college students speaking English as a second language (Massaro and Light, 2003).

6.1. Evaluation of sentence intelligibility with Baldi

To give a standard for the improvement that we obtained for the Arabic talking head, Badr, we now give the method and results for an analogous experiment with Baldi, our English talking head. In this experiment carried out several years earlier (Massaro and Cohen, unpublished) college students were asked to report the words of short sentences presented in noise. On some trials, only the auditory sentence was presented. On other trials, the auditory sentence was accompanied by either the video of the original talker articulating the sentence or accompanied by Baldi aligned with the natural auditory speech.

Fifteen introductory psychology students served as participants. The test items consisted of 65 meaningful sentences from the CID set (Davis and Silverman, 1978), e.g., "We will eat lunch out" or "Pick up the pencil." The sentences were 3, 4, and 5 syllables in length. Two talkers were used. First, a professional speaker recorded on video disk (Bernstein and Eberhardt, 1986), and

second, Baldi, our synthetic talker. The audio from the video disk was used for an auditory alone condition and for both the bimodal human and synthetic talkers. For the synthetic talker, the natural speech was analyzed with viterbi forced alignment to determine the durations of each phoneme. The resulting phoneme and duration data was used to drive the visual speech synthesis.

The video was played on a SONY LDP-1500 laserdisc player while the synthetic talker was generated in real time at 30 frames/s on an SGI Crimson-Reality Engine. All experimental events and data collection were controlled by the SGI computer. On each bimodal trial, one of the two talkers and one of the 65 sentences were selected for presentation. The audio/video signal from the computer or the laser disk was selected by a PANASONIC MX-50 mixer under computer control, and presented on JVC TM-131SU 13 in. monitors. The auditory speech was mixed with speech noise (Grason Stadler noise generator). The participants were asked to type in as many words as they could for each sentence, via TVI-950 terminals connected to the computer.

There were three presentation (auditory-alone or the same auditory with two talkers) conditions times 65 sentences. There were 260 trials (one complete randomized design of 65 sentences * 3 conditions plus an additional 65 sentences randomly selected from the complete design), occurring in two 20 min sessions of 130 trials each, with a short break in between.

For each participant, for each of the presentation conditions, the results were scored in terms

Table 6
The individual correct word recognition results of an earlier experiment on the English Talking head Baldi as a function of the unimodal auditory condition, the bimodal synthetic face condition, and the bimodal natural face condition

Participant #	Auditory (%)	Synthetic face (%)	Natural face (%)
1	48	63	79
2	54	63	79
3	47	52	64
4	41	62	74
5	41	60	83
6	34	56	68
7	45	67	75
8	57	71	86
9	47	66	82
10	52	64	78
11	45	59	72
12	58	74	86
13	58	72	80
14	55	68	77
15	23	47	69
Mean	47	63	77

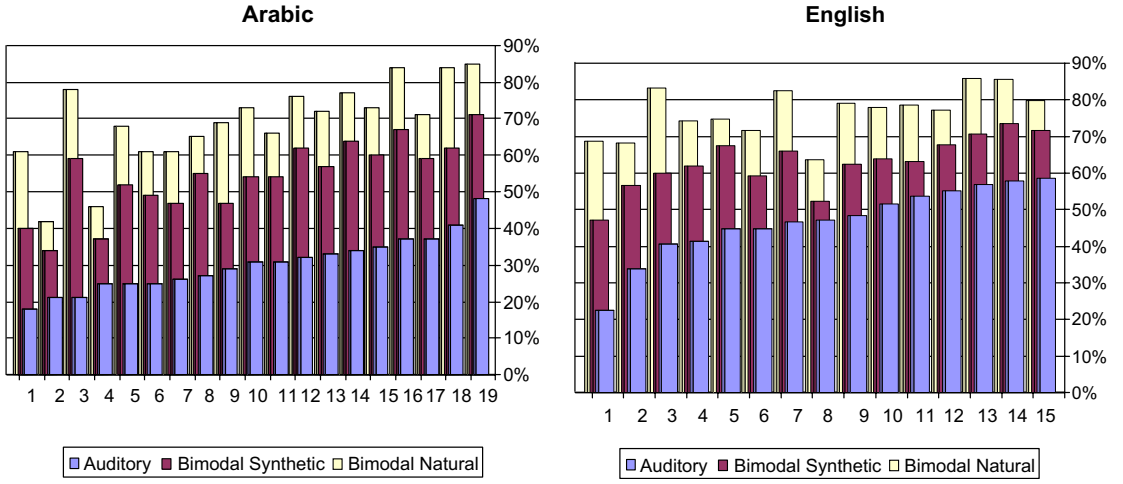

Fig. 19. The percentage of correct word recognition of each of the 19 participants in the Arabic experiment and the 15 participants in the English experiment as a function of the unimodal auditory condition, bimodal synthetic face condition, and the bimodal natural face condition.

of the proportion of correct word identifications. In this analysis, the score was determined by the proportion of words in the stimulus sentences that occurred in the response sentences. The mean performance scores were 45.0%, 61.5% and 75.3% for auditory alone, synthetic and natural bimodal respectively. The results are presented for each participant in Table 6.

6.2. Comparison

Fig. 19 shows the results of the English and Arabic experiments. As we mentioned earlier it is hard to compare two experiments if they have different auditory levels. Thus, Fig. 20 gives a measure given by Equation (1) that allows the comparison of the improvement even though auditory performance in each experiment was not exactly the same (Sumby and Pollack, 1954). We can notice clearly that both talking heads provide roughly the same amount of, with a small advantage to the Arabic talking head. Thus, we can conclude that we were successful in improving the Arabic visible speech to a level comparable to the improvement found in English and one that has been shown to be effective in facilitating language learning. It should be noted that further improvements have already been made in English by training Baldi on objective measures of articulation (Cohen et al., 2002; see also Chuang et al., 2002). We also plan to record objective measures for Arabic and to use these data to train Badr (Table 4–6).

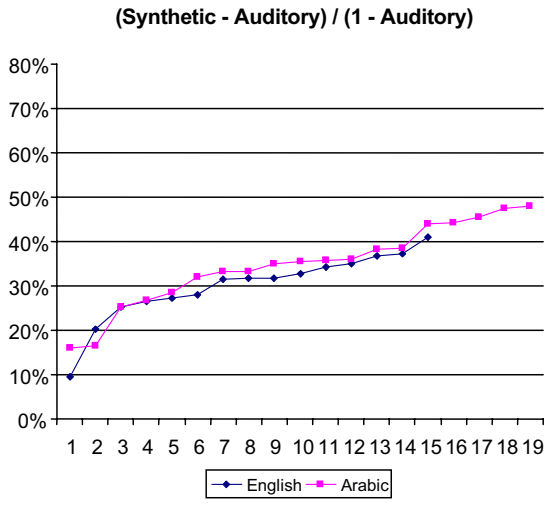

Fig. 20. The improvement of phoneme recognition provided by the synthetic face independently of the performance in the auditory condition, for Arabic and English participants, using the measure given in Eq. (1).

7. Potential applications

A multilingual talking head will help to make many languages accessible for research and other applications. By extending the research in 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 (Breslaw et al., 1981; Holt et al., 1997). 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 (Massaro and Light, 2004a). 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 (Bosseler and Massaro, 2003).

Baldi has also been used as a speech and listening tutor for hard of hearing children (Massaro and Light, 2004b). 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 are constantly working in improving the articulation of our talking head and adding new parts that are valuable to modeling accurately speech. For

example, our latest work in progress is the addition of a velum and the back of the vocal tract to Baldi. This should allow a better design of the vocal tract and provide more information for language learners to distinguish nasals from non-nasal sounds. We look forward to progress in the application of animated speech to communication and human machine interaction.

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. Thanks to all the participants in the different experiments ran in France and in Tunisia. Thanks also to Kais Ouni and Noureddine ELLOUZE at Ecole Nationale d'Ingenieur de Tunis, Tunisia for running the experiment and providing the test facilities.

References

Al-Ani, S., 1970. Arabic Phonology, An Acoustical and Physiological Investigation. Mouton, The Hague.

Ali, L., Daniloff, R., 1972. A contrastive cinefluorographic investigation of the articulation of emphatic/non-emphatic cognate consonants. Studia Linguistica 26 (2), 81–105.

Bagein, M., Dutoit, T., Malfrere, F., Pagel, V., Ruelle, A., Tounsi, N., 2000. The EULER Project: an Open, Generic, Multi-lingual and Multi-Platform Text-To-Speech System. Paper presented at the ProRISC'2000, Veldhoven.

Barker, L.J., 2003. Computer-assisted vocabulary acquisition: The CSLU vocabulary tutor in oral–deaf education. Journal of Deaf Studies and Deaf Education 8, 187–198.

Bernstein, L.E., Eberhardt, S.P., 1986. Johns Hopkins Lipreading Corpus Videodisk Set. The Johns Hopkins University, Baltimore, MD.

Beutnagel, M., Conkie, A., Schroeter, J., Stylianou, Y., Syrdal, A., 1999. The AT&T Next Gen TTS System. In: Joint Meeting of ASA, EAA and DAGA. http://www.naturalvoices.att.com/.

Blanz, V., Basso, C., Poggio, T., Vetter, T., 2003. Reanimating faces in images and video. In: Brunet, P., Fellner, D. (Eds.), Proceedings of EUROGRAPHICS 2003, Granada, Spain, 2003.

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.

Bregler, C., Covell, M., Slaney, M., 1997. Video rewrite: driving visual speech with audio. In: Proceedings of ACM SIGGRAPH 97.

Breslaw, P.I., Griffiths, A.J., Wood, D.J., Howarth, C.I., 1981. The referential communication skills of deaf children from different educational environments. Journal of Child Psychology 22, 269–282.

Chuang, E.S., Deshpande, H., Bregler, C., 2002. Facial Expression Space Learning. In: Proceedings of Pacific Graphics, pp. 68–76.

Cohen, M.M., Beskow, J., Massaro, D.W., 1998. Recent developments in facial animation: an inside view. In: ETRW on Auditory–Visual Speech Processing. Terrigal-Sydney, Australia, pp. 201–206.

Cohen, M.M., Massaro, D.W., 1993. Modeling coarticulation in synthetic visual speech. In: Models and Techniques. In: Thalmann, D., Magnenat-Thalmann, N. (Eds.), Computer Animation. Springer-Verlag, Tokyo, pp. 141–155.

Cohen, M.M., Massaro, D.W., Clark, R., 2002. Training a talking head. In: Proceedings of ICMI'02, IEEE Fourth International Conference on Multimodal Interfaces. 14–16 October, Pittsburgh, Pennsylvania.

Cohen, M.M., Walker, R.L., Massaro, D.W., 1996. Perception of synthetic visual speech. In: Stork, D.G., Hennecke, M.E. (Eds.), Speechreading by Humans and Machines: Models, Systems, and Applications. Springer, Germany, pp. 53–168.

Cosi, P., Cohen, M.M., Massaro, D.W., 2002. Baldini: Baldi speaks Italian. Seventh International Conference on Spoken Language Processing (ICSLP 2002), Denver, Colorado.

Davis, H., Silverman, S.R., 1978. Hearing and Deafness, fourth ed. Holt, Rinehart and Winston, New York.

Elgendy, A.M., 2001. Aspects of Pharyngeal Coarticulation. University of Amsterdam, The Netherlands, Amsterdam.

Ezzat, T., Poggio, T., 1998. MikeTalk: A Talking Facial Display Based on Morphing Visemes, Tony Ezzat and Tomaso Poggio, Proceedings of the Computer Animation Conference. Philadelphia, PA, June 1998.

Ezzat, T., Geiger, G., Poggio, T., 2002. Trainable videorealistic speech animation. ACM Transactions on Graphics 21 (3), 388–398.

Gairdner, W.H.T., 1925. The Phonetics of Arabic: A Phonetic Inquiry and Practical Manual for the Pronounciation of Classical Arabic and of One Colloquial: (the Egyptian) Humphrey Milford. Oxford University Press.

Ghazali, S., 1977. Back Consonants and Backing Coarticulation in Arabic. The university of Texas at Austin, Austin.

Guenter, B., Grimm, C., Wood, D., Malvar, H., Pighin, F., 1998. Making faces. SIGGRAPH, Orlando—USA, 55–67.

Holt, J.A., Traxler, C.B., Allen, T.E., 1997. Interpreting the Scores: A User's Guide to the Ninth Edition Stanford Achievement Test for Educators of Deaf and Hard-of-hearing Students. Gallaudet Research Institute, Washington, DC.

Huang, X.D., 1998. Spoken language technology research at microsoft. Journal of the Acoustical Society of America 103 (5), 2815–2816.

Jakobson, R., 1962. "Mofaxxama", the emphatic phonemes in Arabic. Jakobson, R., Selected Writings, Vol. 1. Mouton and Co, The Hague, pp. 510–522.

Jesse, A., Vrignaud, N., Cohen, M.M., Massaro, D.W., 2000. The processing of information from multiple sources in simultaneous interpreting. Interpreting 5 (2), 95–115.

Kähler, K., Haber, J., Seidel, H.-P., 2001. Geometry-based Muscle Modeling for Facial Animation. Proceedings of the Graphics Interface 2001 (7–9), 37–46.

Klatt, D., 1987. Journal of the Acoustical Society of America (Review of text-to-speech conversion) 82, 737–793.

Lindau, M., Ladefoged, P., 1983. Variability of Feature Specifications. In: Perkell, J., Klatt, D. (Eds.), Invariance and Variability in Speech Processes. Lawrence Erlbaum Associates, Inc., New Jersey, pp. 464–479.

Massaro, D.W., 1998. Perceiving Talking Faces: From Speech Perception to a Behavioral Principle. MIT Press, Cambridge, MA.

Massaro, D.W., 2004. Symbiotic value of an embodied agent in language learning. In: Proceedings of the Thirty-Seventh Annual Hawaii International Conference on System Sciences. IEEE Computer Society.

Massaro, D.W., Bosseler, A., Light, J., 2003. Development and evaluation of a computer-animated tutor for language and vocabulary learning. Fifteenth International Congress of Phonetic Sciences (ICPhS '03), Barcelona, Spain.

Massaro, D.W., Cohen, M.M., Beskow, J., 2000. Developing and evaluating conversational agents. In: Cassell, J., Sullivan, E., Prevost, S., Churchill, E. (Eds.), Embodied Conversational Agents. MIT Press, Cambridge, MA, pp. 286–318.

Massaro, D.W., Cohen, M.M., Tabain, M., Beskow, J., Clark, R., (in press). Animated speech: Research progress and applications. In: Vatiokis-Bateson, E., Bailly, G., Perrier, P. (Eds.), Audiovisual Speech Processing. MIT Press, Massachusetts.

Massaro, D.W., Light, J., 2003. Read my tongue movements: Bimodal learning to perceive and produce non-native speech /r/ and /l/. Eurospeech 2003—Switzerland (Interspeech). Eighth European Conference on Speech Communication and Technology, Geneva, Switzerland.

Massaro, D.W., Light, J., 2004a. Improving the vocabulary of children with hearing loss. Volta Review 104 (3), 141–174.

Massaro, D.W., Light, J., 2004b. 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.

Parke, F.I., 1975. A model for human faces that allows speech synchronized animation. Journal of Computers and Graphics 1 (1), 1–4.

Sproat, R., 1997. Multilingual Text-to-Speech Synthesis: The Bell Labs Approach. Kluwer Academic Publishers, Boston.

Sumby, W.H., Pollack, I., 1954. Visual contribution to speech intelligibility in noise. Journal of Acoustic Society of America 26, 212–215.

Taylor, P., Black, A.W., 1997. The Festival Speech Synthesis System: System Documentation. Human Communication Research Centre, University of Edinburgh, Scotland, UK.