14th European Signal Processing Conference (EUSIPCO 2006), Florence, Italy, September 4-8, 2006, copyright by EURASIP

EMOTIONAL AUDIO VISUAL ARABIC TEXT TO SPEECH

*M. Abou Zliekha , S. Al-Moubayed**

*O. Al-Dakkak and, N. Ghneim***,

*Damascus University/Faculty of Information Technology
email: mhd-it@scs-net.org ; email: kamal@scs-net.org

** Higher Institute of Applied Science and Technology (HIAST)
P.O. Box 31983, Damascus, SYRIA
phone: + (963-11) 5120547, fax: + (963-11) 2237710. email: odakkak@hiast.edu.sy ; email: n_ghneim@netcourrier.com

ABSTRACT

*The goal of this paper is to present an emotional audio-visual Text to speech system for the Arabic Language. The system is based on two entities: un **emotional audio text to speech** system which generates speech depending on the input text and the desired emotion type, and **un emotional Visual model** which generates the talking heads, by forming the corresponding visemes. The phonemes to visemes mapping, and the emotion shaping use a 3-paramertic face model, based on the Abstract Muscle Model. We have thirteen viseme models and five emotions as parameters to the face model. The TTS produces the phonemes corresponding to the input text, the speech with the suitable prosody to include the prescribed emotion. In parallel the system generates the visemes and sends the controls to the facial model to get the animation of the talking head in real time.*

1. INTRODUCTION

Speech and face expressions are the two basic human communication ports to the external world and other humans. When emotions are expressed through both speech and face, they give a considerable added value to the meaning of the speech. The information extracted from the speaker face is very valuable to the perception of speech.

The visual information of the talking head is a very useful in both speech synthesizing and recognition systems. The collaboration between the audio phonemes and the visual vismes helps to improve speech perception and remove the ambiguity that may occur, on some phonemes (like the ambiguity between phonemes /m/ and /n/), especially in noisy environments, in addition to increasing the non-verbal communicative signals like prosody and emotions. These advantages are especially relevant when communicating with people hearing difficulties. The audio-visual speech synthesis could be used in various domains like 1) automatic news and weather broadcasting, 2) educational applications (virtual lectures) 3) aid to hearing impaired people.

Our Objective is to build an Audio Visual Arabic text to speech system, depending on the following components

- Emotional Arabic text to speech system.
- Emotional Face Model and Control to build visemes

2. THE AUDIO VISUAL SYSTEM

There are several projects and works, done on audio-visual speech. MIRALab, University of Geneva worked on the emotional aspects [8], other researches were done on audio-visual TTS: Department of computer Science at the Sheffield university on English [9], the Chinese University of Hong Kong on Chinese [4], INPG/ENSERG-Université Stendhal on French [2], and Tokyo Institute of Technology on Japanese [6]. Our work is a contribution for the Arabic language.

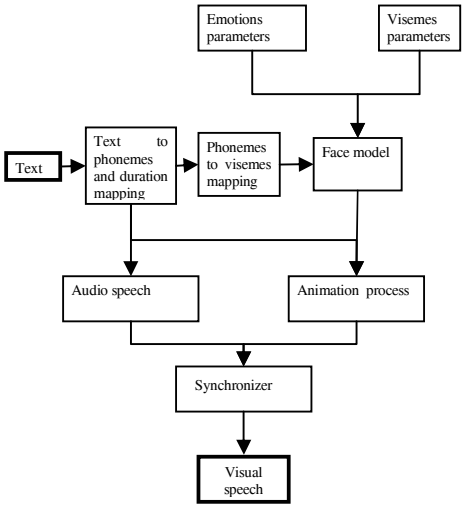

Figure 1: The system components

Figure 1 shows the components of our system. A text and an emotion choice is entered, and an animated face together with the corresponding speech is produced.

The first bloc includes the emotional TTS which generates the phonemes, with the appropriate prosody to give the desired speech. The output of this bloc controls the second bloc that generates the animated talking face. The emotion gives the general facial expressions, and the phonemes give the mouth movements. These parameters control the positions of the face muscles, during speech productions.

3. THE EMOTIONAL ARABIC TEXT TO SPEECH

In a previous paper, we presented the inclusion of emotions in an Arabic Text-to-Speech. In this section we remind of our TTS system in HIAST and the inclusion of emotions in it. We also recall our results regarding this issue

3.1 The Arabic TTS in HIAST

With the objective of building a complete system of standard spoken Arabic with a high speech quality, we defined the following steps to achieve this goal (1) the definition of the phonemes' set used in standard Arabic, we have 38 phonemes, 28 consonants and 5 vowels (/a/, /u/, /i/ and the opened vowels /o/ and /e/ with 5 emphatic vowels[3a], (2) the establishment of the Arabic text-to-phonemes rules using TOPH (Orthographic-PHonetic Transliteration) [5] after its adaptation to Arabic Language, and (3) the definition of the acoustic units; the semi-syllables, and the corpus from which these units are to be extracted.

As the Arabic syllables are only of 4 forms: V, CV, CVC, CVCC, the semi-syllables are of 5 forms: #CV, VC#, VCC# (# is silence), and in continuous speech we have VCV and VCCV; hence the logatoms from which those semi-syllables are extracted are respectively: Cvsasa, satVC, satVC$_1$C$_2$, tV$_1$CV$_2$sa, tV$_1$C$_1$C$_2$V$_2$sa. Where the small letters are pronounced as they are, V, V$_1$, V$_2$ scans all the vowels and C, C$_1$, C$_2$ scan all the consonants. Some combinations never occur in the language, they are excluded. This corpus is being recorded (not finished yet) It will be segmented and analyzed using PSOLA techniques [10], and in parallel (5) the incorporation of prosodic features in the syntactic speech.

3.2 Prosody and emotion generation

The output of the third step of our TTS is converted according to MBROLA transcription. MBROLA system allows control on pitch contour and duration for each phoneme. To control the amplitude we built an additional module. Therefore we were able to test our prosody and emotion synthesis.

The automatic prosody generation in our TTS enables the hearer to distinguish assertions, interrogations, exclamations, or negations. Our methodology was as follows: We began by recording groups of sentences (assertions, interrogations,…), with different length (short medium and long sentences). We extracted the prosodic features, then we modeled the different curves of prosody (F0, intension, durations). For each group, we have different models according to sentence length. A fuzzy logic controls the application of the different models. We perform a set of rules to produce the prosodic parameters automatically, depending on punctuations. Subjective tests showed the improvement in speech quality with the generated prosody.

In another study [11], we developed rules to modify the prosodic parameters to synthesize emotions (Joy, sadness, fear, surprise and anger). The automated tool we've developed for emotional Arabic synthesis proved to be successful, especially in conversational contexts. The emotion recognition rates ranged from 67% to 80%.

To improve this recognition rate, we developed a multimodal synthesis, incorporating face animation with emotional speech.

4. THE AUDIO-VISUAL ARABIC TEXT TO SPEECH

In the present section, we develop our approach to build the emotional audio visual TTS on Arabic. First, we define a set of visual models corresponding to phonemes clusters, then we present the face model and control methods to produce the animated faces

4.1. The Phonemes Visemes mapping

Since many phonemes can not be differentiated based on audio signals (such as voiced, voiceless or nasals), Fischer (1968) has introduced the notion of visual phonemes (visemes), where phonemes are clustered based on their visual similarity [Pelachaud]. Visemes play an important role in the selection of the suitable video segments of animated faces.

A viseme could correspond to more than one phoneme, because some phonemes are visually alike (ex. /s/ and /z/).

For our Arabic audio-visual speech synthesis system, we've built an inventory of Arabic visemes classes and developped a phoneme-viseme mapping. (See Table 1)

4.2. Face modelling and control

There are several methods to model faces, such as: 1) Constructive Solid Geometry (CSG), which describes the geometry of complex scenes by applying a set of operations to primitive objects, 2) voxels, which are volumic pixels (by adding depth to an image), 3) parametric surfaces, which represents the surfaces using functions and ranges, or polygons mesh. 4) polygons mesh, which is the mostly used method, as it takes less time to render, and can be efficiently manipulated by graphics hardware.

The polygon mesh is the model we adopt for our system. In fact, it is easy to convert any other designed model to this model. Figure 2 shows a face model with polygons mesh (right), and the same model covered with skin (left).

Arabic trascription	phoneme	Viseme number
ا	aa	(1)
ق	q	(1)
ب	b	(2)
م	m	(2)
ت	t	(3)
د	d	(3)
ز	z	(3)
س	s	(3)
ص	s.	(3)
ض	d.	(3)
ط	t.	(3)
ك	k	(3)
ن	n	(3)
ث	T	(4)
ذ	D	(4)
ظ	z.	(4)
ج	Z	(5)
ش	S	(5)
ح	X	(6)
خ	x	(6)
ه	h	(6)
ر	r	(7)
ع	H	(8)
غ	G	(8)
ف	f	(9)
ل	l	(10)
و	w	(11)
ضمة ممالة	O	(11)
ي	j	(12)
كسرة ممالة	E	(12)
	Pause #	(13)

Table 1: phonemes-visemes mapping (MBROLA notation)

Face control is a very complex process. Real face consists of bones, muscles and skin; any small change on any of these parameters may cause the perception of different emotions and may generate different visemes.

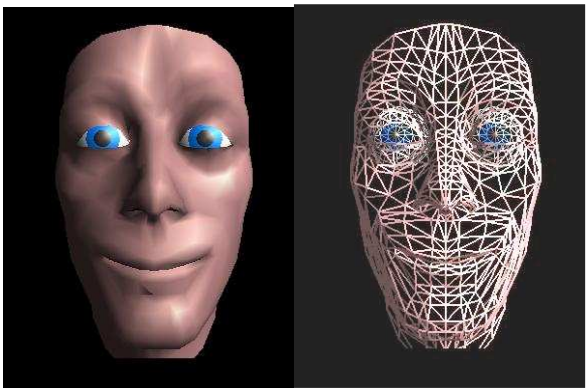

Figure 2 polygon mesh face model

There are several methods to control a face model: 1) key-frame morphing method, which animates a graphical object by creating smooth transitions between various models. This method is simple and fast, but needs the design of all the viesemes for each desired emotions (n visemes and m emotion requires n*m face designs), 2) Parametric control method [7], which divides the face into small controlled areas (like jaw rotation, mouth opening, etc). The parameters of this model (amount of area movement) are not simply related to the mesh, and any change in the mesh will influence the areas boundaries, and therefore needs a redefinition of the control areas, 3) Abstract muscle-based model [1], in which t he face consists of two categories of muscles: linear muscles and sphincter muscles. Linear muscles pull in the mesh, and are represented by an attachment point and a vector. The sphincter ones represent the muscle squeezing by a change of a corresponding ellipse. Neither of theses muscle kinds is connected to the polygon mesh, their action is limited to an influence zone.

Our approach will depend on two face control models: abstract muscles-based model and the key framing model. The abstract muscles based model is used to produce the shape of a face model for each of the desired emotion, with the corresponding visemes. The interpolation between two visemes is done by key frame modelling. The big advantage of the abstract muscles-based model is that the system of mesh modification is independent of the topology of the face. Further more, if we have n visemes and m emotions, the model requires only n+m parameter sets, but this is done on the expense of more processing time than in the other models.

4.3. Emotional Visemes animation

The production and animation of the different visemes is done as follows:

- Applying the emotion morphing on the mesh to get the desired emotion. This mesh will be the starting mesh of the animation process
- Applying the visemes morphing for each two successive visemes, using key frame models.

The application of visemes morphing on the mesh after applying the emotion morphing will produce the correct viseme shapes, because the abstract muscles-based model is independent of the topology of the face. This fact justifies taking the emotional face as the basic face mesh for the following visemes. This process produces the shape of successive visemes corresponding to the output phonemes of the TTS for the desired emotion.

The animation process between two consecutive visemes can be done, using the same approach (abstract muscle-based model) by extracting the difference between the values of the parameters and applying the morphing step by step. However this takes a lot of execution time, and does not produce a real-time animation.

We adopted to use key-frame animation process to produce animation between two visemes, while considering the emotional viseme shape is temporary static. To pass from one viseme shape to another, we used linear functions. The animation timing is taken from the phonemes durations given by the emotional TTS. The frame updating time is adaptive, depending on phonemes duration.

5. RESULTS

To study the influence of the visual components on the emotional speech intelligibility, we produced the five audio-visual sentences for each emotion. These sentences we exposed to 10 people. Each individual was asked to give the perceived emotion for each sentence. Table 2 shows the results of this test.

Identified / synthesized	*Anger*	*Joy*	*Sadness*	*Fear*	*Surprise*	*Others*
Anger	***92%***	0%	0%	0%	0%	8%
Joy	0%	***87%***	0%	0%	8%	5%
Sadness	0%	0%	***90%***	3%	0%	7%
Fear	0%	0%	2%	***92%***	0%	6%
Surprise	0%	5%	0%	0%	***92%***	3%

Table 2: Audio-visual emotion recognition rates

The *Others* column represents the percentage of results where the emotion perceived by voice did not match the one given by the face, or where the perception is not deterministic in either audio or video; the visual was more expressible than the audio. In table 3, we recall the audio recognition results of emotion done in [11], to see the amount of improvement to the recognition rates added by visual.

Identified / synthesized	*Anger*	*Joy*	*Sadness*	*Fear*	*Surprise*	*Others*
Anger	***75%***	0%	2%	7%	0%	6%
Joy	0%	***67%***	0%	2%	13%	18%
Sadness	5%	0%	***70%***	5%	0%	20%
Fear	3%	0%	5%	***80%***	0%	12%
Surprise	0%	10%	0%	2%	***73%***	15%

Table 3: Audio emotion recognition rates

Compared with the emotional TTS, we found that the recognition rates have increased up to 92% while it was no more than 80%. The minimum recognition rate has improved from 67% to 87%.

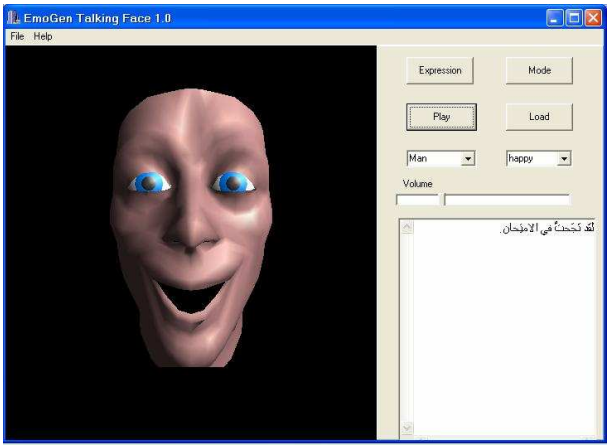

Figure 3 a screen of the execution of our system

Figure 3 shows a screen of the execution of our system. The spoken sentence is "I have passed the exam", with the *Joy* emotion.
The system runs under Windows environment 2.3 GHz CPU 64 MB Graphics card memory, the frame animation duration average was 30 ms.

6. CONCLUSION

In this paper, we presented the first version of an emotional audio-visual speech synthesizer for Arabic texts.
The encouraging results shown above gave us the motivation to go further to reach a perfect coherence between the sound and the image, and perhaps to develop an audio-visual synthesis with various speakers.

REFERENCES

[1] K.Waters, "A muscle model for animating three-dimensional facial expression," in *Proc. SIGGRAPH'87*, vol. 21, no. 4, pp. 17-24.
[2] B. Le Goff, C. Benoît, "A text-to-audiovisual-speech synthesizer for French", in *Proc. ICSLP96*.
[3] O. Al dakkak, N. Ghneim, "Towards Man-Machine Communication in Arabic" in *Proc. Syrian-Lebanese Conference*, Damascus SYRIA, October 12-13, 1999.
[5] V. Aubergé, "La Synthèse de La parole: des Règles aux Lexiques", Thèse de l'université Pierre Mendès France, Grenoble2, 1991.
[4] J.Q. Wang, K.H. Wong, P.A. Heng, H. M. Meng and T. T. Wong, "A real-time Cantonese text-to-audiovisual speech synthesizer," International Conference on Acoustics, Speech and Signal Processing (ICASSP), Montreal, Quebec, Canada, 17-21 May 2004
[6] Masatsune Tamura, Shigekazu Kondo, Takashi Masuko, Takao Kobayashi, "Text-to-Audio-Visual Speech Synthesis Based on Parameter Generation from HMM," in *Proc.* EUROSPEECH' 99, Budapest, Hungary, pp.959-962.
[7] Emmanuel TANGUY, 3D Facial Animation, http://membres.lycos.fr/maybeweb/projet/dissertation.doc.
[8] N. M. Thalmann, http://www.informatik.uni-trier.de/~ley/db/indices/a-tree/m/Magnenat=Thalmann:Nadia.html
[9] J. Edge, "*Expressive visual speech using geometric muscle functions*," in Proc. of *Eurographics UK*, pp. 11–18, April 2001
[10] E. Moulines and J. Laroche, "Non-parametric techniques for pitch-scale and time-scale modification of-speech," *Speech Communication,* vol. 16, pp. 175-205, 1995.
[11] O. Al-Dakkak, N. Ghneim, M. Abou Zleikha, and S. Al-Moubayed, "Emotion inclusion in an Arabic text-to-speech," in *Proc EUSIPCO 2005, 4-8 September,* Turkey.