EMOTION INCLUSION IN AN ARABIC TEXT-TO-SPEECH

O. Al-Dakkak, N. Ghneim*, M. Abou Zliekha** and S. Al-Moubayed***

* 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

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

ABSTRACT

Many attempts have been conducted to add emotions to synthesized speech [1]. Few are done for the Arabic language. In the present paper, we introduce a work done to incorporate emotions: anger, joy, sadness, fear and surprise, in an educational Arabic text-to-speech system. After an introduction about emotions, we give a short paragraph of our text-to-speech system, then we discuss our methodology to extract rules for emotion generation, and finally we present the results we had and try to draw conclusions.

1. INTRODUCTION

When compared with human speech, synthetic speech is in general less intelligible, and less expressive [2]. These are drawbacks for conversational computer systems or for reading machines. The role of emotions in speech is to provide the context in which speech should be interpreted and signal speaker intentions, and this is essential in synthesized speech.

Synthesis systems have to simulate emotions if they want to produce them. There are two ways to perceive emotions: (1) Generative (speaker) model, which depends on the mental and physical states of the speaker, and the syntax and semantic of the utterance, (2) acoustic (listener) model, which describes the acoustic signal parameters as perceived by the listener [2], [3] which we have adopted in our work.

In the present article, we are merely concerned with the production of emotions in Arabic, and the incorporation of these emotions in synthetic speech produced by an Arabic TTS system.

2. ARABIC TTS SYSTEM

We intend to build a complete system of standard spoken Arabic with a high speech quality. The steps to achieve this goal were (1) the definition of the phonemes' set used in standard Arabic including the open /E/ and /O/ [4], (2) the establishment of the Arabic text-to-phonemes rules by using the TOPH (Orthographic-PHonetic Transliteration) system [5] after its adaptation to Arabic Language [6], and (3) the definition of the acoustic units; the semi-syllables, and the corpus from which these units are to be extracted, and in parallel, (4) recording the corpus and extracting the acoustic units prior to analyzing them using PSOLA techniques [7], and in parallel (5) the incorporation of prosodic features in the syntactic speech.

The first three steps are already done. As we intend to use more phonemes than MBROLA systems [4], [8], we decided to choose the MBROLA system to perform preliminary text-to-speech. In fact, Arabic is rather a syllabic language, and semi-syllables are more appropriate for the synthesis [9], [10]. Our corpus is already decided and is in the recording phase.

The output of our third step is converted according to MBROLA transcription. MBROLA system allows control on pitch contour and duration for each phoneme. That enabled us to test our prosody and emotion synthesis. We recall works previously done in the field of general prosody generation for Arabic TTS, such as the ones in [11], [12].

In the present paper we focus on the incorporation of emotions in the system.

3. RULE EXTRACTION FOR VARIOUS EMOTIONS

3.1 Methodology

The most crucial acoustic parameters to consider for emotion synthesis are the prosodic parameters: pitch, duration and intensity [2], [3]. The variations of each of these parameters are described through the following other sub-parameters [13], [2], [14], [15]:

F0 Parameter:

- F0 Range (difference between F0max and F0min)

- Variability (degree of variability: high, low..)
- Average F0
- Contour slope (shape of contour slope)
- Jitter (irregularities between successive glottal pulses)
- Pitch variation according to phoneme class

Duration Parameter

- Speech rate
- Silence rate
- Duration variation according to phoneme class
- Duration variation according to pitch

Intensity Parameter

- Intensity variation according to pitch

Our methodology was to (1) record a corpus of sentences emotionless and with different emotions, (2) analyze these sentences to extract the various parameters and sub-parameters, and extract rules, (3) synthesize emotions according to these rules, and finally test the results and apply tuning on the rules when necessary.

3.2 Recording, analysis and rules extraction

Twenty sentences were chosen for each emotion. Each sentence was recorded twice, one emotionless and the other with the intended emotion. All these sentences were analyzed using PRAAT system to find the prosodic parameters. A statistical study followed to find the relevant changes between the pairs of sentences for each emotion. The following results were found (Table 1):

Emotion	Prosodic Rules
Anger	F0 mean: + 40%-75% F0 range: + 50%-100% F0 at vowels and semi-vowels: + 30% F0 slope: + Speech rate: + Silence rate: - Duration of vowels and semi-vowels: + Intensity mean: + Intensity monotonous with F0 Others: F0 variability: +, F0 jitter: +
joy	F0 mean: + 30%-50% F0 range: + 50%-100% F0 at vowels and semi-vowels: + 30% F0 slope: - Speech rate: - Duration of vowels and semi-vowels: + Intensity mean: + Intensity monotonous with F0 Others: F0 variability: +, F0 jitter: +
sadness	F0 mean: + 40%-70% F0 range: + 180%-220% F0 at vowels and semi-vowels: + Speech rate: - Silence rate: + Duration of vowels and semi-vowels: + Intensity mean: +
fear	F0 mean: + 50%-100% F0 range: +100%-150% F0 at vowels, semi-vowels, nasals and fricatives: + Speech rate: + Silence rate: - Duration of vowels and semi-vowels: + Intensity mean: + Intensity monotonous with F0 Others: F0 variability: +, F0 jitter: +
surprise	F0 mean: + 50%-80% F0 range: + 150%-200% F0 at vowels and semi-vowels: + Speech rate: + Silence rate: - Duration of vowels and semi-vowels: + Others: F0 variability: +

Table 1: Results on natural speech

3.3 Emotion synthesis

To test the above rules, we have developed a tool linked to our TTS system, to control emotional parameters over the Arabic text automatically. The inherent synthetic prosody (emotionless), built in the system is rather coarse, thus the application of the above rules did not give always the desired emotion perception. We had to tune those rules to cope with the synthesizer. The final experimental emotional rules are given below (Table 2):

Emotion	Prosodic Rules
Anger	F0 mean: + 30% F0 range: + 30% F0 at vowels and semi-vowels: + 100% Speech rate: + 75%-80% Duration of vowels and semi-vowels: + 30% Duration of fricatives: + 20%
joy	F0 mean: + 50% F0 range: + 50% F0 at vowels and semi-vowels: + 30% F0 at fricative: + 30% Speech rate: + 75%-80% Duration of vowels and semi-vowels: + 30% Duration of last vowel phonemes: + 20% Others: F0 variability: +40%
sadness	F0 range: + 130% F0 at vowels and semi-vowels: + 120% F0 at fricative: + 120% Speech rate: - 130%
fear	F0 mean: + 40% F0 range: + 40% F0 at vowels, semi-vowels, nasals and fricatives: +30% Speech rate: - 75%-80%

	Others: F0 variability: +60%, F0 jitter: +3%
surprise	F0 mean: + 220% F0 at vowels and semi-vowels: +150% Speech rate: - 110% Duration of vowels: +200% Duration of semi-vowels: +150% Others: F0 variability: +60%

Table 2: Results on synthetic speech

The following five figures show F0 contours for each emotional type sentence with its corresponding emotionless sentence.

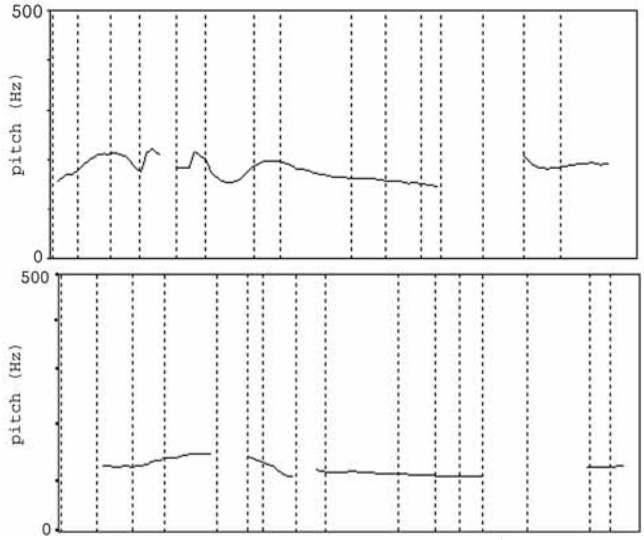


Figure 1: Anger emotion and emotionless /من تظن نفسك؟/ "who do you think you are?"

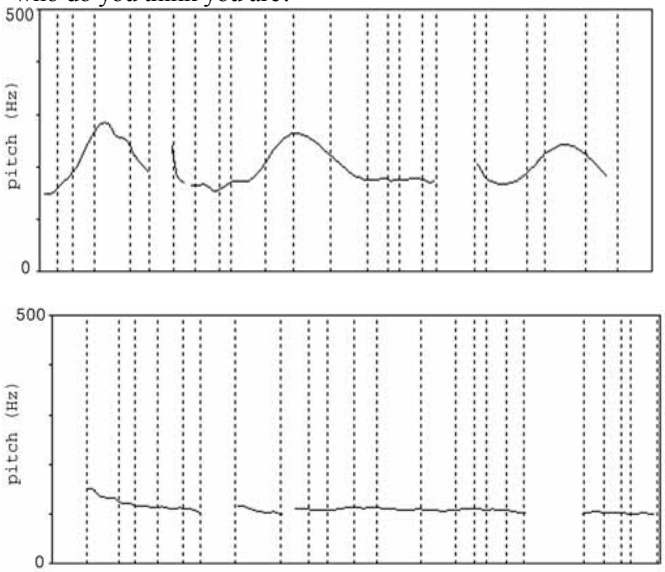
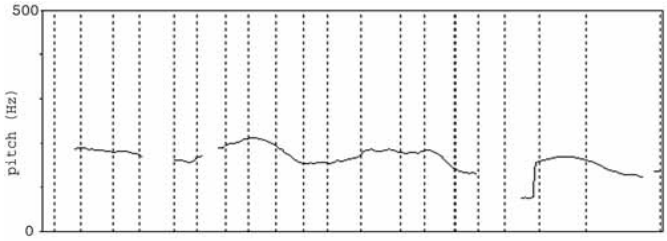


Figure 2: Joy emotion and emotionless /زالت الغيوم من السماء/ "No more clouds in the sky"



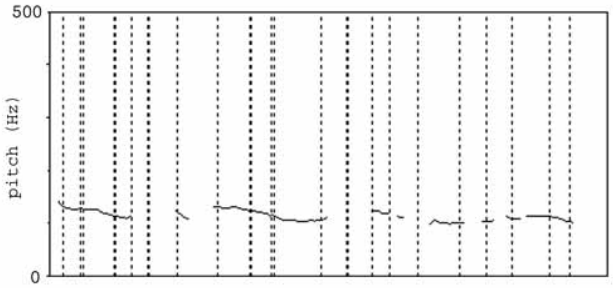


Figure 3: Sadness emotion and emotionless / أنا حزين جداً اليوم/ "I am so sad today!"

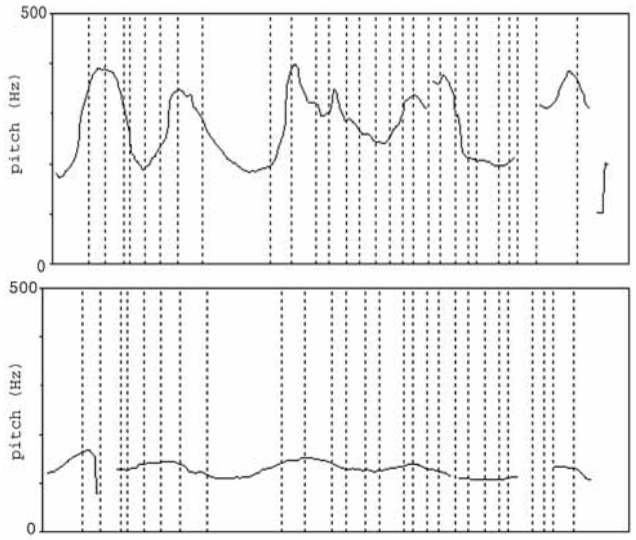
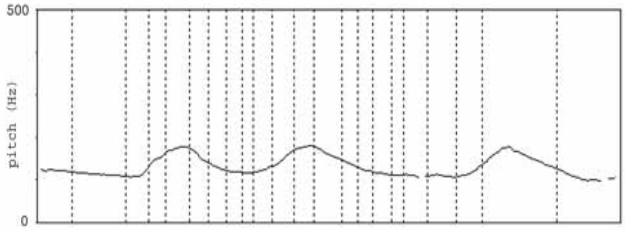


Figure 4: Fear emotion and emotionless /يا إلهي ما هذا المنظر المخيف/ "God! What a scary scene!"



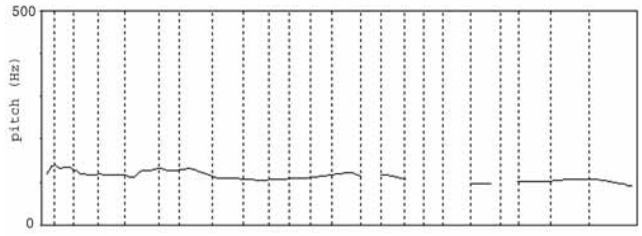


Figure 5: Surprise emotion and emotionless /يا له من منظر جميل!/ "What a beautiful scene!"

4. RESULTS

Using the experimental rules, five sentences for each emotion were synthesized and listened by 10 people. Each individual was asked to give the perceived emotion for each sentence. Table 3 shows the results of this test.

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: Emotion recognition rates

Some people believed that some tested sentences have more than one emotion.

5. CONCLUSION

An automated tool has been developed for emotional Arabic synthesis. The new prosodic model, proposed and tested in this work proved to be successful, especially when applied in conversational contexts.

A further work will follow to incorporate other emotions like disgust, and annoyance.

The quality of the TTS System with its prosody plays a crucial role in emotion synthesis. We intend to refine our prosodic model; the emotional rules have to be revalidated to cope with it.

REFERENCES

[1] M. Schroder, "Emotional speech synthesis: A review". in *Proc. of Eurospeech 2001*, Aalborg, Denmark vol. 1, pp. 561–564, 2001. URL http://www.dfki.de/~schroed/ publications.html.

[2] J. Cahn "The generation of affect in synthesized speech" *Journal of the American Voice I/O Society*, Vol. 8. pp. 1-19, Jul. 1990.

[3] O. Pierre-Yves, "The production and recognition of emotions in speech: features and algorithms" *International Journal of Human-Computer Studies*, vol.59 n.1-2, pp.157-183, Jul. 2003.

[4] 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.

[6] N. Ghneim, H. Habash, "Text-to-Phonemes in Arabic", *Damascus University Journal for the Basic Sciences*,. vol. 19, n. 1,. 2003.

[7] 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.

[8] T. Dutoit, V. Pagel, N. Pierret, F. Bataille and O. van der Vrecken,"The MBROLA project: towards a set of high quality speech synthesizers free of use for non-commercial purposes", *Proc. of ICSLP'96*, pp. 1393-1396, 1996.

[9] N. Chenfour, A. Benabbou and A. Mouradi, " Etude et Evaluation de la di-syllabe comme Unité Acoustique pour le Système de Synthèse Arabe PARADIS", *Second International Conference on language resources and evaluation*, Athenes, Greece, 31 May-2 June 2000.

[10] N. Chenfour, A. Benabbou and A. Mouradi, " Synthèse de la Parole Arabe TD-PSOLA Géneration et Codage Automatiques du Dictionnaire ", *Second International Conference on language resources and evaluation*, Athenes, Greece, 31 May-2 June 2000.

[11] S. Nasser Eldin, H. Abdel Nour and A. Rajouani "Enhancement of a TTS System for Arabic Concatenative Synthesis by Introducing a Prosodic Model", ACL/EACL 2001 workshop, Toulouse-France 2001.

[12] S. Baloul, M. Alissali, M. Baudry and P. Boula de Mareuil "Interface syntaxe-prosodie dans un système de synthèse de la parole à partir du texte en arabe", *XXIViemes Journées d'Etudes sur la Parole*, CNRS/Université Nancy2, Nancy, France, 24-27 juin 2002.

[13] F. Zotter, "Emotional speech", at URL: http://spsc.inw.tugraz.at/courses/asp/ws03/talks/zotter.pdf

[14] I. R. Murray, M. D. Edgington, D. Campion, and J. Lynn, "Rule-based emotion synthesis using concatenated speech", *ISCA Workshop on Speech & Emotion*, Northern Ireland, 2000, pp. 173-177.

[15] J.M. Montero, J. Gutiérrez-Arriola, J. Colás, E.Enríquez and J.M.Pardo, "Analysis and modelling of emotional speech in Spanich", at URL: http://lorien.die.upm.es/~juancho/conferences/0237.pdf