Rules Based Model for Automatic Synthesis of F0 Variation for Declarative Arabic Sentences

A. Zaki[1,2], A. Rajouani[2], Z. Luxey[1], M. Najim[1]

[(1)]Equipe Signal et Image-LAP UMR 5131, ENSEIRB.
[(2)]LEESA, Faculté des Sciences, BP 1014 Rabat, Maroc.
{zaki,najim}@tsi.u-bordeaux.fr, arajouani@yahoo.fr

Abstract

This paper deals with automatic generation of fundamental frequency (F0) contours for standard Arabic language. We use synthesis by rule for modelling Arabic intonation. The proposed model is based on the assumption that linguistic information is contained in target points of intonative contour. The perception of Arabic lexical accent is correlated with variations of F0. Rules used to determine target points are based particularly on accentuation algorithm. The existence of a declination phenomenon is discussed and introduced in our model. The validation of this model is carried out by TTS system based on TD-PSOLA technique. We provide an evaluation of our results based on perceptive test. The proposed processing of intonation allows a noticeable improvement of speech synthesis naturalness.

1. Introduction

Prosody and more particularly intonation play a key role in the perceived naturalness of synthetic speech. In most TTS systems, prosody generation is carried out in two steps: first an abstract description of the sentence prosody, which is derived from linguistic level. Secondly, given this information, the physical parameters (i.e. F0 contour, segmental durations, pauses etc.) are predicted to produce the acoustic description of the intended prosody.

The work reported in the present paper is an important step of the development of prosodic model for an Arabic text-to-speech system.

Many approaches have been proposed for the automatic generation of intonative contours starting from text for various languages. They can be classified in three classes: synthesis by rule [1], synthesis by a concatenation of stored F0 contours [2][4] and stochastic approach: HMM [3], Neural networks [4].

Our purpose consists of using rules to synthesise F0 variations. The assumption used for this approach is that the linguistic information is contained in target points of intonative contour. As other languages, a declination phenomenon is noticeable in Arabic. The absence of these macro-melodic phenomena considerably affects the naturalness of speech synthesis.

This paper is organised as follows: in section 2 we present a linguistic and prosodic background of Arabic. In section 3 we describe intonative model. In section 4 we present the results of our experiments.

2. Background

2.1. Introduction

In this study we focus our attention on standard Arabic language. The contemporary Standard Arabic language, a modernised version of classical Arabic, is the language commonly in use in all Arab-speaking countries today. It is the language of science, learning and media (news paper, TV news…) in opposition with the particular dialects.

The vocalic system of Arabic is composed of 12 vowels. These can be classified according to their length (6 longs and 6 shorts) or their category (6 emphatic and 6 non emphatic). The vowels are realised graphically under or on the consonants.

The consonant system consists of 28 consonants. As other natural languages Arabic includes the syllable unites. The number of syllables is limited to six unites (CV, CVC, CVV, CVVC, CVCC, CVVCC). The four first syllables can appear in the beginning, middle and in the end of the word. The CV frequently appears but the last syllable CVVCC rarely appears, so it will be ignored in our study. The last two syllables appear isolated or only in the final position of the word. Syllabic system can be classified in two categories: the short and long ones. We can also distinguish closed syllables (CVC, CVVC, CVCC) and open syllables (CV, CVV). The prosodic model we present here is based on the syllabic unit for automatic generation of F0 variations. The prosody is defined [5] in the linguistic level as a description (phonetic aspect) and a formal representation (phonologic aspect) of the oral elements of utterances such as stress, tons, intonation and quantity, whose, a physical manifestation in production of speech is associated with F0 variations, the segmental duration and the intensity.

2.2. Prosody and lexical accent

In the Arabic language, the grammarians have completely ignored the prosodic dimension of language and notably accentual proprieties. Some contemporary authors believed themselves in right to draw argument from this fact to minimise, to see denying the linguistic or metric role of accent. According to [6]. "*In Arabic, a word stress never played any distinctive role*". From the linguistic point of view, some *generativists phonologists* and *Arabists* has shown that an accent can play a distinctive role with some minimal pairs. The examples in that case are rather rare. In the acoustic level, accent clearly affects the prosodic parameters (F0, duration and intensity). The studies made in the context of speech synthesis, does not cease confirming the need of lexical accent for an adequate management of the prosody.

Thus, the generation of intonatives contours of the Arabic is based on the lexical accent determination.
Even if the linguists have not devoted exhaustive studies to the lexical accent in Arabic, we can find in the literature more than one accentuation algorithm. This can be explained by dialect influence. In the study reported by [7] and [8], the accentuation is fixed by the word boundary. The word can be defined as accentual phrase accompanied or not by its *clitic*. The accentual area is limited by the last phoneme of syllable string.

In the domain thus defined, the accent can take place at any syllable, whatever its length or its weight. The rules proposed are:

(a) if the last syllable of the word is over heavy (CVVC, CVCC), this one receives the accent;
(b) if (a) does not apply and if the penultimate is heavy (CVV, CVC), this one receives the accent;
(c) if (a) and (b) do not apply, the antepenultimate receives the accent.

The rules formulated in [9] assign the accent on the first syllable in the words whose neither the penultimate nor the antepenultimate are long. The secondary and the tertiary accents are involved in this algorithm.

(a) when a word is made up of a string of the CV type syllables, the first syllable receives the primary stress and the remaining syllables receive weak stresses, e.g. /ka(1)ta(3)ba(3)/ CV(1)CV(3)CV(3) "he wrote";
(b) when a word contains only one long syllable, this syllable receives the primary stress while the rest of the syllables go unmarked, receiving weak stresses. The final long syllable never receives a primary stress. /kaa(1)tib(3)/ CVV(1)CVC(3) "writer";
(c) when dealing with polysyllabic word, a stress is placed on the first long syllable beginning from the penultimate. The nearest long syllable to the beginning of the word receives the secondary stress. /mus(3)taw(2)da(3)>aa(1)tu(3)hum(3)/ CVC(3)CVC(2)CV(3)CVV(1)CV(3)CVC(3) "their deposits".

In [10] and [11] the fundamental frequency is presented as a pertinent parameter for perception of lexical stress in Arabic. The place of principal accent is conserved on the sentence level. The maximum of F0 curve corresponds to the accentuated syllable. This correlation of the realisation of accent and fundamental frequency variations is useful for automatic processing of prosody.

The rules reported in [9] do not make the difference between the heavy and over heavy syllables. In addition, the accentuation of the last syllable is neglected. Indeed, these rules are most convenient to follow the variations of the fundamental frequency in the word level. It is not the case with the rules presented in the first part of this section. This result is ratified by using a speech corpus; recorded by two Arabic native speakers from Maghreb and Middle East. 90% of intonative contours of the first corpus (Moroccan speaker) and 85% for the second corpus (Palestinian speaker) present a clear correlation of the lexical accent realisation and F0 variations. These permit a rejection of dialectal influence hypothesis of rules. These rules are used in this work for automatic generation of F0 variations in the context of speech synthesis.

In the following, we use a corpus for one speaker, consisting of 100 utterances of declarative sentences with lengths from 1 to 10 words offering a large variety of lexical, syntactic and semantic forms. The corpus is recorded in a soundproof room and digitised at 16 KHz with 16 bits/sample. The intonative contour of each sentence is estimated from natural speech by using interactive software graciously provided by Elan-Informatique Company[1]. This ensures an automatic stylisation of F0 curves. The errors occurred when computing F0 values implies a manual tuning to correct data of F0 without changing intonation of sentences at perceptive level. The gap of unvoiced phonemes is automatically filled.
This corpus is used to study a declination phenomenon and analyses the intonative contours, to find a linguistic phenomenon, which contributes to F0 variations on sentences level.

3. Model description

Synthesising intonation has been performed for Arabic language for the first time with development of Arabic TTS system based on synthesis by rules [11][12]. The intonative model established is based on the concatenated intonative contours of the word; is called here a classical model. Each word contour of the sentences is generated using accentuation rules, as it was reported previously in §2 of section 2. The word contour consists of raising and falling movement. The peak of contour is correlated with the principal accent. The other accent levels are neglected. On the perceptive level, the synthesis speech resulting from the classical model, presents undesirable monotony. This latter affects the global trend of the macro-intonation. This can be explained by the fact that the model kept all of the peaks of F0 contour on the same intonative level. In addition, the brutal fall of the F0 in certain word boundaries is disruptive from perceptive point of view.

These problems directly affect the naturalness of speech synthesis. The solutions we propose here for automatic processing of intonation are:

- introduce declination phenomenon;
- taking into account the second and third levels of accent;
- involving a set of phonologic and phonotactic rules for smoothing F0 contour.

The modelling process consists in description of the melodic curve by a set of target points succession. These latter's are located at the picks and valleys of the F0 contour. The connection between themes is made by use an appropriate transition function. The prediction of target points is based on the syllable unites. The syllable can be associated with one or two target points. Each syllable is defined by its pitch and position in the sentences. The macro-prosody is decomposed into target points associated to each syllable. The micro-prosodic phenomenon is not sufficiently known to make it possible to build a quantitative model. Consequently it is not treated in this approach. The strategy of modelling presented here is based on the computing of speaker register as it was reported in [13].

3.1. Declination phenomena in Arabic

The declination can be defined from the acoustic point of view as a trend, which has the fundamental frequency decreasing from the beginning to the end of the sentence. It is a macro-melodic phenomenon. Its realisation tack place at least on the sentence level. Here, we will not discuss the cognitive aspect of the declination, nor the fact that it is programmed or not by

[1] Prosel software: Alignment of speech signal with phonetic string of text.

the speaker, expected or not by the listener. First we aim at showing the existence of declination phenomena from acoustic realisation of intonation in Arabic. Secondly we propose to model a speaker register, which is defined by, the bottom and the top line: they are the envelope for the F0 contour [14]. The speaker register gives access to amplitude variations of the fundamental frequency. In the first step, for each sentence, we compute by using linear regression the bottom and the top lines closest to F0 values. These lines correspond respectively to the minimum and the maximum of the intonative contour. In the second step, we compute the slope and the *Y*-Intercept[2] of both bottom and top lines according to the number of syllables in the sentences. The Figure 1 represents the evolution of the slope according to the number of syllables in each sentence for both lines. Sentences with a small number of syllables have a great slope but when this number reaches 25 the slope tends to zero. The high values of slope for sentences with syllables number is under to 20 prove the existence of declination phenomenon. For long sentences we must take into account the re-initialisation effect. This study deals with sentences containing less than 25 syllables.

3.2. Modelling of declination

According to the Figure 1, we can observe that the slope evolution can be approached by logarithmic function. From *Y*-Intercepts evolution, we notice that the Intercept follow a linear evolution according to the number of syllables. In the aim of modelling a declination phenomenon, and computing speaker register, we model a slope evolution for both lines by two approximate functions using a non-linear regression. We approximate *Y*-Intercept evolution by linear function of the straight-line using a linear regression.

Table 1: Equations modelling the slope (S) and Y-Intercept (I) of each declination line defines the speaker register according to the number of syllable (N).

	Slope	***Y-Intercept***
Bottom line	$S=\frac{-16}{2+N}$	$I=0.94*N+124.7$
Top line	$S=\frac{-33}{1+N}$	$I=0.23*N+103.11$

The slope is inversely proportional to the syllable number but the *Y*-Intercept has a linear dependence. Given the number of syllables, the implementation procedure for computing a declination phenomenon, and eventually a speaker register is based on computing the bottom and top lines using a linear function $Y=S(N)*x+I(N)$, x represents a syllable position. The pitch of each target point will be quantified according to the declination lines as we can see in the Figure 2. In the following we present a set of rules, which permits the prediction of target points necessary to generate F0 contour.

3.3. Rules used for modelling F0 contour

Indeed, if we analyse the phonologic liaison between two words in the sentence we notice the following phenomena:

- the last syllable of the first word, which has a weak stress when pronounced alone, receives a secondary accent;
- the liaison makes it possible to create a so called sense group, by bringing together both words.

The monosyllabic prepositions /fii/, /min/, />an/, /maa/ etc. receive a weak stress instead of a principal one. The realisation of the accent for different levels, correlated with fundamental frequency variations, takes place with respect to declination from the beginning to the end of sentence (see Figure 2). The method proposed here operates in 2 steps:
First, the localisation of the target points and determination of their pitch by using a set of rules. Secondly, we assign the fundamental frequency to each target point by means of speaker register.

We use the pitch symbols (**T**, **B**, **M**, **H,** U, **D**, **L)** of INTSINT system[3] to define a pitch of target point [15]. **T**, **B**, **M** are known as absolute pitch. They depend on the accentuation rules. **H**, **L,** U and **D** are relative pitch, which are resulting from phonotactic rules. **T**, **M** and **B** are respectively associated to principal, secondary and third accents. In the following we propose a set of phonotactic rules:

- when the realisation of this pitch targets sequence **T-B-T**, the pitch target **B** is replaced by **H**;
- when the realisation of this pitch targets sequence **M-B-T**, the pitch target **B** is replaced by **U**;
- when the realisation of this pitch targets sequence **T-B-M**, the pitch target **B** is replaced by **D**;
- The pitch targets of more than one **B** pitch targets succession are not all aligned with the bottom line. It is the case only for the last **B** target, but for the others, the **L** pitch target will replace them.

The phonotactic rules are results from the forward and the backward effect of accentuated syllables. They are formulated according to the analysis by synthesis study of the declarative corpus. The proposed rules could be illustrated with the example presented in Table 2.

3.4. Implementation stage

The implementation stage is based on the automatic localisation of target points. We compute the speaker register (Bottom and Top line). This step permits to associate to each pitch target an appropriate F0 value. The **B**, **M** and **T** are aligned respectively to the bottom, middle and top lines of speaker register. For the relative symbols (**H**, **U**, **D**, **L**), the F0 values are computed by taking into account the contextual information, the pitch of the right and left targets. We use exponential transition function for interpolating F0 of target points.

4. Test and results

We test our model with 20 sentences. Its evaluation is made with a perceptive test. For this step we have prepared 20 pairs of waveform files of speech synthesis. Each pair consists of: sentence synthesised with our model and other synthesised with a classical model. Four native Arabic subjects familiar with speech synthesis and prosody are involved in our test. Each subject is asked to choose from each waveform pairs the sentence with a preferable intonation. As we can see in Table 3, the average of answers rate, which gives preferential treatment to the sentences synthesised by using the proposed model is 96%. This result shows the improvement brought by the new treatment. We can see in Figure 2 an example of

[2] The most useful form of the straight-line equation is the "slope-intercept" form: *Y*=m*X*+b. This called the slop–intercept form because "m" is the **slope** and "b" gives the ***Y*-intercept**.

[3] **IN**ternational **T**ranscription **S**ystem for **INT**onation

preliminary result of our model. The graphic shows F0 variations with respect to the declination phenomenon. Some samples of synthesised speech using a new model are available in: http://tsi.u-bordeaux.fr/zaki/arabic-synthesis.html

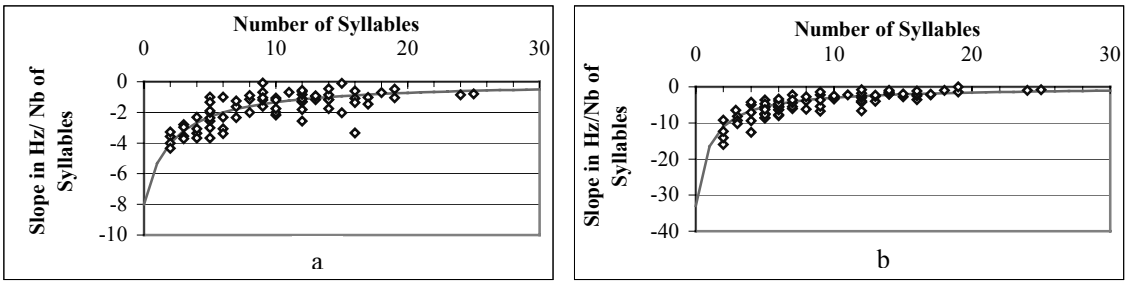

Figure 1: the graphics (a) and (b) represent respectively the evolution Slope of the bottom and Top line.

Table 2: Example of phonologic representation of target points.

Sentences	MUSTAW DA>A:TUHUMU LLATI: FI LMAONA>I												
Meaning in English	Their deposit which are in its factory												
Syllabification	MUS	TAW	DA	>A:	TU	HU	MUL	LA	TI:	FIL	MAO	NA	>I
Syllable type	CVC	CVC	CV	CVV	CV	CV	CVC	CV	CVV	CVC	CVC	CV	CV
Accent level	3	2	3	1	3	3	2	1	3	3	1	3	3
Pitch Target	**B**	**M**	**U**	**T**	**L**	**B**	**M**	**H**	**L**	**B**	**T**	**L**	**B**
Syllable position	1	2	3	4	5	6	7	8	9	10	11	12	13

Table 3: Scores of subjective test.

Subject number	Score of the new model	Score of the classical model
1	100%	0%
2	95%	5%
3	100%	0%
4	90%	10%

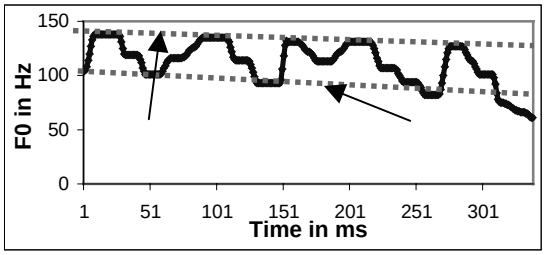

Figure 2: Example of synthetic F0 contour generated by our model of sentences "qaabala >ucmaanu fii wazni rriichati muhammadan".

5. Conclusion

The treatment proposed in this paper provides a noticeable improvement of speech synthesis naturalness. In fact, the declination phenomenon has an important effect on the global trend of the melody. The rules we have proposed are useful to automatic processing of intonation. They can ensure F0 variations relatively comparable to the natural curve. A coming work will deal with application of this approach for other modalities of Arabic language such as interrogation, exclamation, call and imperative sentences.

6. References

[1] Anderson, M., D., Pierrehumbert, J. B., Liberman, M. Y., 1984. Synthesis by Rule of English Intonation Patterns. In *Proceedings of ICASSP-IEEE*, *San Diego*, 281-284.

[2] Malfrère, F., Dutoit, T., Mertens, P., 1998. Fully Automatic Prosody Generator for Text-To-Speech Synthesis. In *Proceedings of ICSLP*, *Sidney*, 1395-1398.

[3] Ljolje, A., Fallside, F., 1986. Synthesis of Natural Sounding Pitch Contours in Isolated Utterances Using Hidden Markov Models. *IEEE Transactions ASSP*, 34, 1047-1080.

[4] Traber, C., 1992. F_0 Generation with a Database of Natural F_0 Patterns and with a Neural Network. In *Talking Machines: Theories, Models, and Designs*. Bailly, G., Benoît, C., Sawallis, T., R., (eds). North-Holland: Elsevier Sciences. 287-304.

[5] Di Cristo, A., 2000. Interpréter la Prosodie. In *Actes des XXIII^{ème} JEP*, *Aussois*, 13-29.

[6] Cantineau, J., 1960. *Etudes de Linguistique Arabe*. Library C. Klincksiech. Paris.

[7] Kouloughli, D., 1976. Contribution à l'Etude de l'Accent en Arabe Littéraire. In *Annales de l'université d'Abidjan.* Série H, vol. IX. 124-125.

[8] Mokhtar, O., 1991. *Phonetic Study*. Arabic version. Aalam Al-Kutub Publisher. Cairo.

[9] Al Ani, S., 1970. *Arabic Phonology: An Acoustical and Physiological Investigation*. The Hague, Netherlands: Mouton.

[10] Rajouani, A., Chidami, D., Najim, M., 1987. Synthèse et Perception de l'Accent Lexical en Arabe. In *Actes des XVI^{ème} JEP*, *Hammamet*, 302-305.

[11] Es-Skali, L., Rajouani, A., Najim, M., Chidami, D., 1987. Eléments d'un Modèle Intonatif pour la Phrase Affirmative en Arabe. In *Actes des XVI^{ème} JEP*, *Hammamet*, 282-285.

[12] Rajouani, A., 1989. *Contribution à la Synthèse de la Parole Arabe par Règles*. Thèse de doctorat d'état, Université Mohamed V, Faculté des Sciences Rabat.

[13] Beaugendre., F., 1994. *Une Etude Perceptive de l'Intonation du Français*. Thèse de doctorat,. Université Paris XI-Orsay.

[14] Pierrehumbert, J., 1981. Synthesizing Intonation. *J. Acoust. Soc. Am*. 70(4), 985-995.

[15] Hirst, D., Dicristo, A., 1998. *Intonation Systems A Survey of Twenty Languages*. Cambridge University Press, 1-44.