Synthèse des variations de la fréquence fondamentale de la parole arabe à partir du texte

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

[(1)]Equipe Signal et Image, ENSEIRB UMR 5131, B.P 99, F-33 402 TALENCE Cedex, France

[(2)]LEESA, Faculté des Sciences, BP 1014 Rabat, Maroc.

Tél.: +33 556 84 61 85 - Fax: +33 556 84 84 06

{najim,zaki}@tsi.u-bordeaux.fr, arajouani@yahoo.fr

Résumé – *Cette communication s'articule autour de la génération automatique des variations de la fréquence fondamentale (F0) pour la langue arabe standard. Cette étude pour la modélisation de l'intonation de l'arabe est menée dans le contexte de la synthèse par règle. Le modèle proposé est fondé sur l'hypothèse qui stipule que l'information linguistique est contenue dans les points cibles du contour intonatif. La perception de l'accent lexical en arabe est corrélée avec les variations de F0. Les règles employées pour déterminer les points cibles sont fondées sur l'algorithme d'accentuation. La validation de ce modèle est effectuée par l'utilisation de deux systèmes de synthèse. Une évaluation perceptive des résultats est proposée. Le traitement proposé de l'intonation permet une amélioration considérable du naturel de la parole de synthèse.*

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 validation of this model is carried out by TTS system. 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

La prosodie et plus particulièrement l'intonation, joue un rôle déterminant dans la perception de l'aspect naturel de la parole de synthèse. Dans la plupart des systèmes de synthèse la génération de la prosodie est traitée selon deux étapes :

– La première correspond à une description abstraite de la prosodie de l'expression écrite qui est issue du niveau linguistique.

– La deuxième étape consiste à prédire, à partir des informations issues de la première étape, l'ensemble des paramètres physiques qui sont associés à la prosodie.

Du point de vue acoustique, la prosodie est définie par les variations des trois paramètres principaux suivants : la fréquence fondamentale, la durée segmentale et l'intensité. Cet ensemble de paramètres représente l'information nécessaire pour qu'un synthétiseur puisse reproduire une parole synthétique comme on peut l'illustrer sur la figure 1.

L'étude présentée dans cette communication représente une étape primordiale pour le développement d'un modèle prosodique pour un système de synthèse de la parole arabe à partir du texte.

Plusieurs approches ont été proposées pour la génération automatique à partir du texte des variations de la fréquence fondamentale qui caractérisent le contour intonatif. Elles peuvent être classées en trois approches : synthèse par règles [1], synthèse par concaténation des contours intonatifs pré-stockés [2][3] et l'approche stochastique telle que les modèles HMM [4] et les réseaux de neurones [3][5]. Les travaux sur la modélisation des variations de la fréquence fondamentale dans le contexte de la synthèse de la parole arabe à partir du texte sont assez rares.

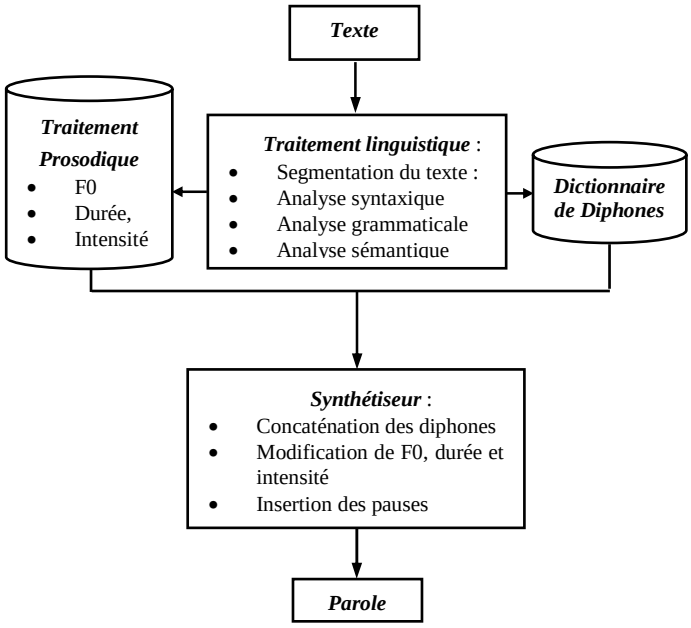

FIG. 1 : organigramme d'un TTS basé sur les diphones

On propose dans cette communication l'application de l'approche fondée sur les règles pour la modélisation de l'intonation de la langue arabe standard. Cette approche a été utilisée pour la langue arabe mais avec une méthodologie de modélisation différente de celle présentée dans cette communication [6]. L'hypothèse de l'approche par règles développée ici stipule que l'information linguistique est contenue dans les points cibles du contour intonatif. Comme

pour d'autres langues, la langue arabe comprend aussi le phénomène de déclinaison. L'absence de ce phénomène macro-prosodique affecte considérablement le naturel de la parole de synthèse.

Cette communication est organisée comme suite : dans la section 2 on présente une introduction sur l'aspect linguistique et prosodique de la langue arabe. La section 3 sera consacrée à la description du modèle intonatif. Dans la section 4, on présente les tests et résultats du modèle intonatif. La section 5 constitue une conclusion générale.

2. Fondement linguistique

2.1 Introduction

La langue arabe standard contemporaine est la version moderne de la langue arabe classique utilisée dans tous les pays arabes. C'est la langue de l'enseignement et des media (journaux, TV etc.) par opposition aux dialectes.

Le système vocalique de la langue arabe est composé de 12 voyelles [7]. Elles peuvent être classifiées selon la longueur (6 longues et 6 courtes) ou selon la catégorie (6 emphatiques et 6 non emphatiques). Les voyelles se réalisent graphiquement sur ou sous les consonnes.

Le système consonantique est composé de 28 consonnes. Comme d'autres langues, l'arabe comprend la syllabe qui joue un rôle important sur le plan linguistique. Le nombre de syllabes est limité à six types de syllabes : CV, CVV, CVC, CVVC, CVCC, CVVCC. Les quatre premières syllabes figurent au début, milieu et à la fin du mot. La CV est la syllabe la plus fréquente contrairement à la CVVCC qui ne sera pas prise en compte dans notre étude. Les deux dernières syllabes figurent généralement dans un contexte isolé ou à la fin d'un mot. Notre modèle intonatif est fondé sur l'unité syllabique pour la génération automatique des variations de F0.

2.2 Prosodie et accent lexical

Dans le contexte de la synthèse à partir du texte, les études acoustiques [8] effectuées sur la perception de l'accent lexical ont montré l'existence d'une relation étroite entre les variations de F0 est la réalisation de l'accent lexical. Ainsi, la formalisation de cette relation facilitera la génération automatique de F0 à partir du texte. La réalisation de l'accent lexical est gérée par un ensemble de règles dites règles d'accentuation [9]. Les règles d'accentuation adoptées dans cette étude sont :

1) quand il s'agit d'un mot composé d'une chaîne syllabique de type CV, c'est la première syllabe qui reçoit l'accent primaire et le reste des syllabes reçoit l'accent faible dit de 3ème degré ;
2) quand il s'agit d'un mot qui contient uniquement une syllabe longue, celle-ci reçoit l'accent primaire et le reste des syllabes reçoit un accent de 3ème degré. Il faut noter dans ce cas qu'une syllabe longue à la fin du mot n'est pas considérée ;
3) quand il s'agit d'un mot qui contient deux syllabes longues ou plus de deux syllabes, la syllabe longue la plus proche de la fin du mot (la dernière syllabe n'est pas considérée) reçoit l'accent primaire et dans la majorité des cas la syllabe longue la plus proche du début (et non pas celle de début) reçoit l'accent secondaire.

Ces règles qui concernent le mot dans un contexte isolé peuvent être étendues avec une légère modification aux mots qui forment une phrase [10]. Dans le contexte de la synthèse de la parole à partir du texte, nous avons dégagé à partir de l'analyse par synthèse d'un corpus de parole les résultats suivants :

- lors d'une liaison phonologique entre deux mots, la syllabe résultante de cette liaison phonologique (syllabe située à la fin du premier mot et qui peut être soit une CVC ou CVVC) reçoit un accent de deuxième degré.
- les propositions monosyllabiques /fii/, /min/, />an/, /maa/ etc. reçoivent l'accent du 3ème degré. Elles sont dites mots clitiques.
- le contour intonatif dont les variations de la fréquence fondamentale sont corrélées avec l'accent suit une tendance de déclinaison du début à la fin de la phrase.

3. Description du modèle

Dans notre démarche de modélisation, nous nous sommes inspirés de l'approche dite « *Tone Sequence Model* » [11]. Cette approche a été appliquée à l'origine pour l'anglais [1]. Une adaptation a été proposée pour le français [12].

Cette approche opère en deux étapes principales :

- l'étape de description intonative qui dépend de la métrique et de la phonologie de la langue étudiée.
- l'étape de quantification intonative qui correspond au modèle phonétique approprié pour la génération des variations de F0.

3.1 Etape de description intonative

Cette étape consiste à représenter le contour intonatif par un ensemble de points cibles, associés aux syllabes du texte. A ces points cibles, on affecte des tons de niveaux différents : bas, intermédiaire ou haut. On distingue deux types de tons les tons absolus et les tons relatifs. Les tons absolus sont prédits en se basant sur l'algorithme d'accentuation dont les règles sont introduites dans la section 2.2. Par contre les tons relatifs sont déterminés à partir des règles phonétiques contextuelles. Cette description prosodique est codée par des symboles inspirés de l'alphabet prosodique du système INTSINT[1] [13]. Les tons absolus sont codés par les symboles suivants: **T**(op), **M**(id), **B**(ottom). Les tons relatifs sont codés par les symboles suivants: **L**(ower), **H**(igher), **S**(ame), **U**(pstepped), **D**(ownstepped).

3.2 Etape de quantification intonative

L'étape de génération automatique des variations de F0 se fonde sur les résultats issus de l'étape précédente. Le modèle de quantification de F0 est fondé sur la génération du registre du locuteur. Ce dernier est délimité par deux droites qui sont appelées: « *top-line* » et « *base-line* ». Ces droites donnent accès aux variations d'amplitude de la fréquence fondamentale. Les deux droites sont calculées par la méthode

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

des moindres carrés à partir des minima et des maxima du contour intonatif.

Nous avons appliqué l'approche développée à la génération de l'intonation de phrases avec deux types de modalités : déclarative et interrogative. Dans le cas des phrases interrogatives, on distingue deux types de phrases interrogatives :

– les phrases qui admettent comme réponse « oui » ou « non ». Ce type de phrase est repéré soit par la présence de ces deux pronoms qui marquent l'interrogation Hal « هَلْ » et ?A « أ » ou bien l'absence totale de marqueur d'interrogation au début de la phrase ;

– les phrases qui commencent par les autres marqueurs d'interrogation de la langue arabe standard : man « مَنْ », kam « كَمْ », ayna « أَيْنَ », maaada « مَاذَا », etc.

Cette distinction est faite également au niveau du contour intonatif global et par conséquent sur la morphologie du registre du locuteur dynamique.

Le calcul du registre du locuteur nous conduit à trois formes globales selon la modalité de la phrase comme on peut l'illustrer sur les figures 2-a, 2-b et 2-c.

Les équations de génération automatique des frontières du registre du locuteur sont celles d'une droite :

$$Y = S(N)x + I(N),$$

où X est la position de la syllabe dans le texte et les coefficients *S* et *I* sont respectivement la pente et l'intercepte de la droite. Ces deux coefficients sont calculés à partir des équations optimisées avec une régression non linéaire dans le cas de la pente (FIG. 3) et une régression linéaire dans le cas de l'intercepte en fonction du nombre de syllabes du texte.

TAB. 1 : paramètres de calcule du registre du locuteur pour les phrases déclaratives

	Pente	***Y-Intercepte***
Base-line	$S = \frac{-16}{2+N}$	$I = 0.94 * N + 124.7$
Top-line	$S = \frac{-33}{1+N}$	$I = 0.23 * N + 103.11$

3.3 Règles de modélisation des variations de F0

Comme nous l'avons souligné auparavant, les règles d'attribution des tons absolus aux syllabes sont gérées par les règles d'accentuation. Les syllabes qui reçoivent un accent de 3ème degré[2] reçoivent des tons qui ne sont pas toujours absolus. Ainsi des règles ont été élaborées pour attribuer des tons convenables à ces syllabes, dits tons relatifs. La première étape du traitement tonique[3] consiste à attribuer à toutes les syllabes de la phrase des tons absolus : **T** pour les syllabes qui portent un accent de première degré, **M** pour les syllabes qui portent un accent de deuxième degré et **B** pour les syllabes qui portent un accent de troisième degré. La deuxième étape du traitement tonique consiste à transformer certains tons **B** en tons relatifs en tenant compte du contexte gauche et droit pour assurer une transition adéquate entre tons. Les règles de transitions ont été élaborées en se basant sur l'analyse par synthèse.

Par exemple, dans le cas où la première étape du traitement tonique produit une séquence tonique **T-B-T**, le deuxième traitement remplace le **B** de cette séquence par **H**. Les valeurs en termes de fréquences des tons sont attribuées par alignement sur le registre du locuteur. Les tons absolus **T**, **M** et **B** sont respectivement alignés sur le « *top-line* », la ligne intermédiaire du registre du locuteur et le « *base-line* ». Les valeurs fréquentielles des tons relatifs sont calculées en tenant compte des valeurs gauche/droit des tons des syllabes adjacentes.

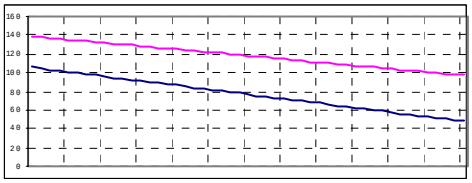

FIG. 2-a : registre des phrases déclaratives

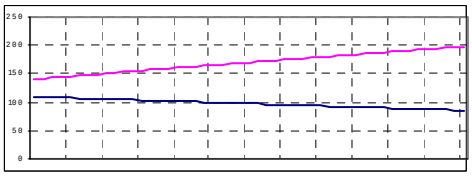

FIG. 2-b : registre des phrases interrogatives oui/non question

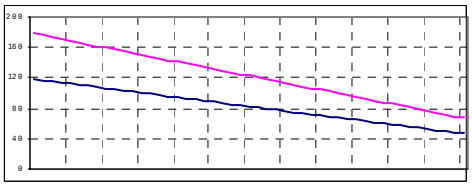

Fig. 2-c : registre des autres phrases interrogatives

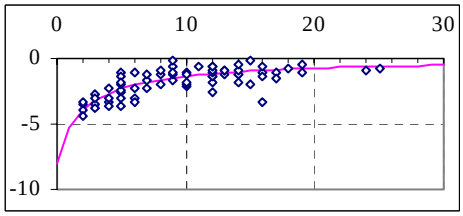

FIG. 3 : fonction optimisée pour le calcule de la pente du *Base-line*

4. Implémentation, tests et résultats

Le modèle intonatif présenté a été implanté sur deux synthétiseurs différents. Les deux systèmes de synthèse sont fondés sur la méthode de concaténation des diphones par deux techniques différentes MBROLA[4] [14] et TD-PSOLA[5] [15]. Ces techniques permettent l'adaptation de la prosodie des unités chargées du dictionnaires avec les paramètres estimés du texte comme l'illustre la figure 1.

Pour tester le modèle développé, une procédure de test perceptif a été élaborée. Le test consiste à comparer trois signaux différents avec différentes combinaisons. Pour chaque phrase de test, trois signaux de synthèse sont générés : signal original re-synthétisé « OS », signal de synthèse à partir du texte avec le modèle proposé « SM » et le troisième

[2] L'accent de 3ème degré signifie qu'il s'agit d'une syllabe qui reçoit un accent faible ou syllabe non accentuée.

[3] On appellera l'étape d'attribution des tons aux syllabes par le traitement tonique.

[4] Disponible sur l'URL : http//tcts.fpms.ac.be/synthesis/mbrola.html

[5] Système de synthèse d'Elan-Informatique Toulouse.

est un signal de synthèse à partir du texte avec des courbes intonatives intrinsèques des diphones concaténés « SB ». On note que pour les trois signaux, nous avons utilisé les mêmes durées segmentales estimées du signal original afin de maintenir le même rythme et focaliser l'attention des sujets sur les variations de F0. Quatre paires de signaux correspondants à la même phrase sont présentées à quatre sujets. Le corpus de test comprend 20 phrases ce qui fait 80 paires de signaux à comparer pour chaque sujet. Les sujets devaient comparer l'aspect intonatif de deux signaux de parole présentés et préciser si le premier est meilleur (1>2), si les deux sont équivalents (⇔) ou enfin si le deuxième est meilleur que le premier. Les résultats du test perceptif sont présentés dans le tableau 2 sous forme de pourcentage des différentes réponses selon chaque type de test.

Tab. 2 : résultats du test perceptif

	OS(1)/OS(2)			OS(1)/SM(2)			OS(1)/SB(2)			SB(1)/SM(2)		
	1>2	⇔	1<2	1>2	⇔	1<2	1>2	⇔	1<2	1>2	⇔	1<2
1	5	95	0	40	55	5	95	5	0	0	5	95
2	15	85	5	35	55	10	95	5	0	5	5	90
3	0	100	0	50	45	5	100	0	0	0	0	100
4	5	95	0	45	55	0	100	0	0	0	5	95

Une lecture du tableau 2 nous permet de conclure qu'en moyenne, 10/20 des signaux de synthèse avec le modèle intonatif développé « SM » sont jugés équivalents aux signaux originaux re-synthétisées « OS » et au moins un signal « SM » a été jugé meilleur que celui du « OS ». La comparaison des signaux de synthèse « SM » et « SB » favorise largement le modèle intonatif développé avec 95% en faveur des signaux «SM ». Les autres combinaisons de test ont permis de d'évaluer la capacité d'évaluation des sujets afin d'éviter les résultats contradictoires. Le pourcentage de mauvais jugements est très bas : 5% en moyenne.

Des résultats portant sur les phrases de parole synthétique de test sont disponibles sur le site : http://www.ts.u-bordeaux.fr/zaki/synthese/demo.html.

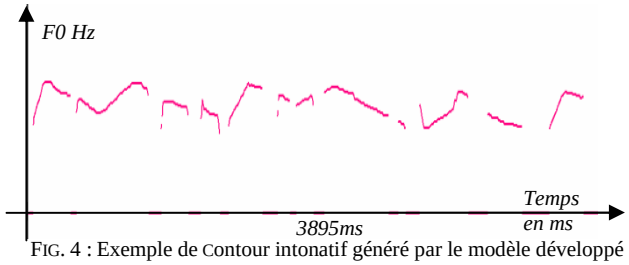

FIG. 4 : Exemple de Contour intonatif généré par le modèle développé

5. Conclusion

Le test perceptif a permis de valider la capacité du modèle intonatif proposé à générer des variations de F0 qui améliorent relativement l'aspect naturel de la parole de synthèse.

En comparaison avec l'approche neuronale [16], l'approche présentée dans cette communication, permet une implantation flexible des connaissances linguistiques caractérisant la langue traitée. L'approche développée ne produit pas de résultats aléatoires comme c'est le cas parfois avec l'approche neuronale.

Références

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

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

[3] Y. Morlec, G. Bailly and V. Aubergé. "Generating prosodic attitudes in French: Data, model and evaluation". Speech Communication 33. pp. 357-371. 2001

[4] A. Ljolje, [. Fallside, ˇSynthesis of Natural Sounding Pitch Contours in Isolated Utterances Using Hidden Markov Modelsˇ. *IEEE Transactions ASSP*, 34, pp. 1047-1080. 1986.

[5] C. Traber, "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. Pp. 287-304. 1992.

[6] S. Nasser Eldin, H. Abdel Nour, A. Rajouani "Automatic Modeling and Implementation of Intonation for the Arabic Language in TTS Systems", Proc. ICSLP vol. I pp. 597-600, Pékin, 2000.

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

[8] A. Rajouani, D. Chidami, M. Najim, « Synthèse et Perception de l'Accent Lexical en Arabe ». *Actes des XVI*ème *JEP, Hammamet*, pp. 302-305. 1987.

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

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

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

[12] P. Mertens, J-P. Goldman, E. Wehrli, A. Gaudinat, « La synthèse de l'intonation à partir de structures syntaxiques riches ». Traitement Automatique des Langues 42 (1), pp. 142-195, 2001.

[13] D. Hirst, J. Véronis, "Analysis of fundamental frequency patterns for multi-lingual synthesis using INTSINT". *Proceedings of 2nd ESCA/IEEE Workshop on Speech Synthesis*. 1994.

[14] Dutoit, H. Leich, "MBR-PSOLA: Text-To-Speech Synthesis based on an MBE Re-Synthesis of the segments DataBase". Speech Communication, vol. 13 pp. 435-440. 1993.

[15] C. Hamon, E. Moulines, F. Charpentier, "A Diphone System, Based on Time-Domain Prosodic Modifications of Speech". In *Proceedings of ICASSP-IEEE*, 238-241. Glascow: Scotland, 1989.

[16] A. ZAKI, A. Rajouani, M. Najim "Synthesizing Intonation of Standard Arabic Language Using Neural Network". Proceedings of Eurospeech'2001 Conference, pp:541-544, Aalborg, September 2001.