Réalisation d'un système hybride de synthèse de la parole arabe utilisant un dictionnaire de polyphones

Tahar SAIDANE (1), Ahmed HADDAD (2), Mounir ZRIGUI (2) et Mohamed BEN AHMED (3)

(1) Société Tunisienne d'Electricité et du Gaz,
Centre de production de Sousse, Tunisie
saidane.tahar@planet.tn
(2) Laboratoire RIADI, Unité Monastir
Faculté des Sciences de Monastir, Tunisie
mounir.zrigui@fsm.rnu.tn
(3) Laboratoire RIADI, Ecole Nationale
des Sciences de l'informatique, Tunis, Tunisie
Mohamed.BenAhmed@riadi.rnu.tn

Résumé – Abstract

Notre article s'intègre dans le cadre du projet intitulé "Oréodule" : un système embarqué temps réel de reconnaissance, de traduction et de synthèse de la parole. L'objet de notre intérêt dans cet article est la présentation d'un système hybride de synthèse de la parole arabe et plus précisément du volet du traitement acoustique. En effet, nous présenterons le module de syllabation en unités acoustiques de tailles variables (phonème, diphone et triphone), ainsi que le dictionnaire de polyphones correspondant. Nous détaillerons les étapes de constitution de ce dictionnaire et les difficultés rencontrées lors de son élaboration. Nous intègrerons également les différents résultats pratiques obtenus lors de chaque phase (nombre de polyphones, tailles des enregistrements, volume total du dictionnaire, etc.).

Our paper is integrated in the scope of the project titled "Oréodule" : a real time embedded system of speech recognition, translation and synthesis. The object of our interest in this work is the presentation of our hybrid synthesis system of the arabic speech and more precisely of the acoustic treatment shutter. Indeed, we will present the module of syllabication in acoustic units of variable sizes (phoneme, diphone and triphone), as well as the corresponding polyphones dictionary. We will list the stages of constitution of this dictionary and the difficulties met during its development. We will also integrate the different convenient results gotten during each phase (number of polyphones, sizes of the registrations, total volume of the dictionary, etc.).

Keywords – Mots Clés

Synthèse de la parole arabe, Phonèmes, Diphones, Triphones, Unités acoustiques, Dictionnaire de polyphones.

Arabic speech synthesis, Phonemes, Diphones, Triphones, Acoustic units, Dictionary of polyphones.

1 Introduction

Notre étude porte sur la conception et la réalisation d'un système de synthèse de la parole arabe qui donne la voix la plus naturelle possible tout en tenant compte des particularités de la langue. Le résultat de ces études nous a guidé vers un système hybride de synthèse utilisant la concaténation d'unités acoustiques de tailles variables tout en utilisant des règles établies. Cet article présentera les modules de ce système de synthèse à savoir le module de transcription, le module de syllabation et de concaténation et le dictionnaire d'unités acoustiques (le dictionnaire des polyphones). Les étapes de conception et de réalisation de ces modules seront présentées.

2 Le système hybride de synthèse de la parole

2.1 Choix d'une grammaire de formalisation

Pour homogénéiser la formalisation de notre travail de conception nous avons opté pour l'utilisation d'une grammaire adaptée à la transcription et à la concaténation. Une règle grammaticale, se lit de droite à gauche et doit s'écrire de la façon suivante :

[RésultatPhon.] = {CG(contexte gauche)} + {C(caractère)} + {CD(contexte droit)}

 # est un signe de début de phrase,
$ est un signe de fin de phrase,
§ est une extrémité de mot,
C est une consonne,
V est une voyelle,
CS est une consonne solaire,
CL est une consonne lunaire.

Une telle règle signifie qu'un caractère *C*, précédé d'un caractère ***CD*** et suivie par un caractère ***CG***, aura pour transcription ***RésultatPhon*** [Zrigui 91].

2.2 La conception du module de transcription

L'une des premières recherches que nous avons effectuée a consisté en la formalisation adéquate des problèmes posés par la langue arabe. Ces problèmes sont de différentes sortes : des graphèmes qui ont plusieurs réalisations phonétiques, des phonèmes qui ont plusieurs réalisations graphémiques, des graphèmes qui ne sont pas pris en compte [Ghazali 90]; nous avons alors constaté une absence totale de correspondance Graphème-phonème. La transcription proprement dite se compose de plusieurs phases : le repérage des mots, la syllabation, l'utilisation de règles, l'utilisation d'un lexique, la conversion graphème phonème et enfin la vérification. Le schéma suivant illustre le processus de transcription :

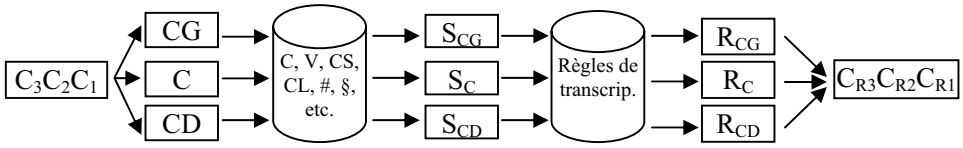

Figure 1 : Utilisation des règles de transcription : le système réalise la transcription en principalement deux étapes, un prétraitement et une application des règles.

2.3 Le prototype du module de transcription

L'alphabet arabe compte 28 consonnes, 6 voyelles (3 longues et 3 courtes) et d'autres réalisations vocaliques [Guerti 83]. L'écriture de ces réalisations se fait de droite à gauche et les consonnes se lient généralement entre elles. Nous avons mis en évidence un ensemble de règles de transcription tout en tenant compte des exceptions. Selon des analyses d'échantillons de texte, ***133 règles*** ont été établies et utilisées à travers une base de données comprenant aussi tous les graphèmes et leurs phonèmes correspondants (une table de 49 caractères). Ci-après un extrait de cette table :

Gr	Ph	Rem.	Gr	Ph	Rem.
ب	b	même valeur qu'en français	غ	ɣ	le "r" parisien grasseyé
ت	t	même valeur qu'en français	ف	f	même valeur qu'en français
ج	ʒ	même valeur qu'en français : je	ك	k	même valeur qu'en français

Figure 2 : Un extrait de la table de correspondance graphème phonème qu'on utilise pour la transcription basé sur l'alphabet IPA96.

En ce qui suit la description de quelques règles élaborées et incorporées dans la base des règles:

- ***Règle 1 :*** [uu]={CS}+{ُ}+{و} ***Règle 2 :*** [uu]={CL}+{ُ}+{و}

Lorsque le و est précédé par la voyelle ُ et qu'il est suivi par une consonne (lunaire ou solaire), on obtient le phonème de la voyelle longue [uu] (حُوتٌ , دُونَ).

- ***Règle 86 :*** {CS}+{أ}={CS}+{ـال}+#
- ***Règle 90 :*** {CS}+{أ}+{V}+{C}={CS}+{ـال}+{V}+{C}

Lorsque le ـال est en début de phrase ou qu'il est précédé par une consonne voyellée et qu'il est suivi par une consonne solaire, il est équivalent à la non présence du لـ (ذَهَبَ الرَجُلُ , السَمِيعُ).

La langue arabe présente par ailleurs quelques exceptions qui ne peuvent pas être décrites par des règles générales. Ces mots ont été rassemblés dans une table des exceptions qui sera parcourue lors de l'étape de transcription.

3 Le système de syllabation et choix des unités acoustiques

Après étude, nous avons adopté un système de synthèse par concaténation dont les unités acoustiques sont de trois types : les triphones, les diphones et les phonèmes. Ceci nous a permis d'apporter plus de souplesse et surtout une meilleure qualité à notre module acoustique [Moulines 96]. La figure 3 présente un exemple type de syllabation.

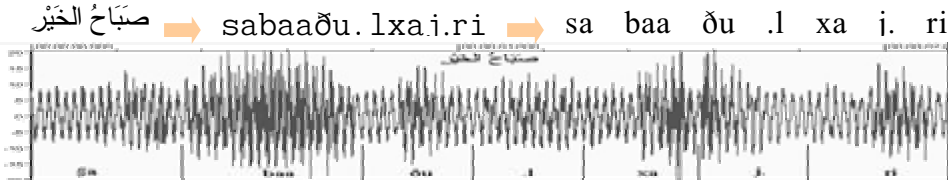

Figure 3 : Exemple de syllabation de l'expression صبَاحُ الخَيْر (bonjour) en polyphones.

L'entrée du module de sélection est une séquence de phonèmes, l'algorithme converge alors vers une suite optimale d'unités acoustiques à concaténer. La sélection dynamique des unités se traduit alors par la recherche de la séquence optimale de représentants, visant à minimiser

les discontinuités au point de concaténation. La problématique de la sélection des unités a été formalisée en utilisant ***6 règles***, illustrées dans le tableau suivant :

[CVV] ={V}+{V}+{C}	Lorsqu'une consonne est suivie de deux voyelles les trois graphèmes constituent une unité acoustique de notre système.
[CV]={C}+{V}+{C}	Lorsqu'une consonne est suivie d'une voyelle puis d'une consonne les deux premiers graphèmes constituent une unité acoustique.
[CC]={C}+{C}+{C}	Lorsque nous avons une succession de trois consonnes les deux premiers graphèmes constituent une unité acoustique.
[C]={V}+{C}+{C}	Lorsque nous avons deux consonnes suivies par une voyelle seul le premier graphème constitue une unité acoustique de notre système.
[VV]={V}+{V}	Lorsque nous avons une succession de deux voyelles les deux constituent une unité acoustique de notre système.
[V]={V}	Lorsque nous avons une voyelle isolée elle constitue une unité acoustique de notre système.

Figure 4 : Les six règles de syllabation qu'utilise notre système pour la langue arabe.

Il est à noter que l'ordre d'application de ces règles ainsi établies est très important pour une bonne syllabation et donc une meilleure concaténation sonore. C'est à partir de ces résultats que nous avons recueilli les échantillons sonores utiles à la constitution de la base d'enregistrements nécessaire à la synthèse.

4 Les étapes de constitution du dictionnaire de polyphones

Pour constituer un dictionnaire d'unités acoustiques il faut disposer de toutes les combinaisons réalisables. Le module de concaténation a besoin de la totalité des unités acoustiques sous la forme d'enregistrements sonores. Ces enregistrements constituent le dictionnaire de notre système. Le dictionnaire ainsi établi contient 196 unités acoustiques (***28 phonèmes de type C, 84 diphones de type CV et 84 triphones de type CVV***), suffisantes pour la réalisation des différentes occurrences possibles. Les étapes de réalisation peuvent se résumer en ce qui suit :

- La saisie du corpus de mots et d'expressions.
- L'enregistrement sonore des expressions.
- La segmentation des enregistrements sonores obtenus en phonèmes, diphones et triphones.
- Le test du dictionnaire obtenu.

La qualité du résultat final de la synthèse dépend directement de la qualité des enregistrements effectués lors de l'élaboration du dictionnaire d'unités acoustiques ; quelques précautions ont alors été prévues tel que l'utilisation d'un seul locuteur par dictionnaire et la limitation des séances d'enregistrement.

4.1 Le choix du corpus initial

Nous avons choisi, pour faciliter la prononciation et en conséquence la segmentation, d'enregistrer un corpus de mots significatifs appartenant au vocabulaire arabe tout en plaçant les unités voulues au milieu des mots afin d'éviter la troncature de la réalisation phonémique. Pour l'extraction de la totalité des polyphones nous avons utilisé les enregistrements de près de ***137 phrases*** et expressions utilisant le vocabulaire arabe usuel. Nous avons par la suite relevé la fréquence d'utilisation des différents polyphones dans ces expressions afin de se donner le maximum de possibilités pour une bonne extraction des unités. Le tableau suivant illustre un extrait de ces mesures :

Polyphone	Fréquence	Polyphone	Fréquence	Polyphone	Fréquence
ma	30	saa	13	m	13
la	28	maa	13	hi	12
sa	14	tuu	2	dii	2

Figure 5 : La fréquence d'utilisation des polyphones pour le corpus des phrases enregistrées pour la réalisation du dictionnaire de polyphones.

4.2 Le choix du locuteur

Les locuteurs, doivent avoir une bonne élocution, représentative de la langue (pas d'accent marqué), et doivent prononcer les pages de logatomes avec une prosodie la plus neutre possible (des professionnels peuvent y être prédisposés). Pour notre système, des phrases contenant toutes les unités acoustiques de l'arabe standard, nécessaires à notre système, ont été enregistrées par deux locuteurs différents (une femme et un homme), tous deux maîtrisant la langue et l'accent local. Ce choix a été considéré afin de pouvoir évaluer le type de locuteur qui se prête le mieux à ce genre de système.

4.3 Les opérations d'enregistrement

Pour faciliter l'opération de segmentation, lors de la procédure générale d'enregistrement du corpus nous avons tenu compte de quelques recommandations :

- Les enregistrements ont été effectués en une seule séance.
- Les mots ont été enregistrés par petites séries pour éviter l'effet de liste et la tendance du timbre de la voix à devenir de plus en plus grave au cours du temps.
- Deux mots consécutifs ont été prononcés sans liaison.

Pour notre système nous avons utilisé des fichiers WAV en format PCM échantillonné à ***44.1 kHz en mode 16 bits et en stéréo soit à 172 kbits/s*** (afin de conserver au mieux la qualité des enregistrements). Nous avons utilisé un matériel standard pour pouvoir juger de la dépendance matériel – qualité, mais aussi dans l'optique d'un système peu contraignant visant un maximum d'utilisateurs.

4.4 La segmentation

La méthode de segmentation que nous avons utilisée dans notre travail est manuelle, le processus de segmentation complètement automatique de corpus est jusqu'à présent peu concevable et peu fiable pour son utilisation dans les systèmes de synthèse par concaténation. Au cours de cette étape, l'identification des différentes unités s'est fait à travers l'utilisation de plusieurs outils parmi lesquels : la forme temporelle de l'onde acoustique correspondant à l'enregistrement, le spectrogramme de l'enregistrement et l'audition, qui reste le critère de choix majeur pour la segmentation.

Le dictionnaire d'unités acoustiques ainsi établi a une taille de ***9 MØ*** (en moyenne un ***phonème prend 20 kØ, un diphone 40 kØ et un triphone 60 kØ***). Ci-après un exemple de segmentation :

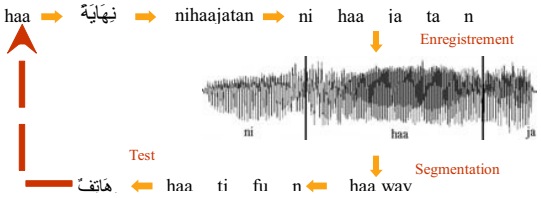

Figure 6 :Un exemple de traitement pour l'obtention du triphone « haa » de l'identification au test en passant par l'enregistrement et la segmentation

5 Conclusion

Lors de l'élaboration du dictionnaire d'unités acoustiques, nous avons rencontré plusieurs difficultés de nature à ralentir le travail et surtout à influer énormément sur la qualité de la voix synthétisée en aval. La majorité de ces contraintes survient lors de l'étape de segmentation. Nous citons en exemple les points suivants :

- L'unité acoustique à extraire doit être au milieu d'un mot, afin d'éviter les variations incontrôlées d'intonation du début et de la fin du mot.
- Des lettres comme (ف ه ح خ ض ظ ذ) sont prononcées au moyen d'une forte expiration, la qualité du microphone peut influer sur le résultat.
- Des lettres comme (غ) et (ر) posent des problèmes de naturel lors des essais de synthèse à cause de leur nature de prononciation.
- La voix féminine est plus nette que celle du locuteur masculin, ce qui influe sur la qualité de la parole produite.
- La qualité de synthèse ne dépend pas que de la nature de la voix d'origine mais principalement de la qualité de la segmentation.

6 Perspectives

Les unités acoustiques, quelles que soient les précautions prises lors de la sélection et de l'enregistrement des unités, ne possèdent pas exactement à leur frontière les mêmes caractéristiques acoustiques. Il est alors nécessaire de procéder à un « lissage » des extrémités des unités acoustiques. Ce volet sera notre principale tâche, ce qui nous permettra d'aboutir à la finalisation d'un produit complet de synthèse de la parole arabe.

Références

Ghazali S., Habaili H., Zrigui M. (1990). Correspondance graphème-phonème pour la synthèse de la parole arabe à partir du texte, *IRSIT*. Congrès dialogue homme machine Tunis.

Guerti M. (1983). Contribution à la synthèse de la parole par diphones en arabe standard, *Institut de Linguistique et de Phonétique. Alger*.

Lemmety S. (2000). Review of speech synthesis technology, *Helsinki University of Technology*. Thèse.

Moudenc T., Emerard F. (2003). Synthèse vocale et handicap, *Annales de télécommunications*. pp 928-934.

Moulines E., Cappe O. (1996). Synthèse de la parole à partir du texte, *Techniques de l'ingénieur*.H1960 pp 7.

Zrigui M., Ghazali S., Ben miled Z., Jemni M. (1990). Synthèse de l'Arabe standard à partir du texte par TD PSOLA, *18ème journée d'étude sur la parole. Belgique*.

Zrigui M., Mili A, Jemni M. (1991). Vers un système automatique de synthèse de la parole arabe, *Maghrebin symposium on programming and system, Alger*. pp 180-197.