LA TRANSCRIPTION ORTHOGRAPHIQUE-PHONETIQUE DE LA LANGUE ARABE

Tahar SAIDANE (1), 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 présente les composants nécessaires à la synthèse de la parole arabe. Nous nous attarderons sur la transcription graphème phonème, étape primordiale pour l'élaboration d'un système de synthèse d'une qualité acceptable.

Nous présenterons ensuite quelques-unes des règles utilisées pour la réalisation de notre système de traitement phonétique. Ces règles sont, pour notre système, stockées dans une base de données et sont parcourues plusieurs fois lors de la transcription.

Our paper presents the components which are necessary for the arabic speech synthesis. We will dwell on the transcription graphème phoneme, a primordial stage for the development of a synthesis system with an acceptable quality.

Then, we will present some of the rules used for the realization of our phonetic treatment system. These rules are, for our system, stocked in a data base and are browsed several times during the transcription.

Keywords – Mots Clés

Transcription graphème-phonème, langue arabe, règles de transcription.
Grapheme-phoneme transcription, Arabic language, transcription rules.

1 Introduction

Notre travail s'inscrit dans le cadre de la réalisation d'un système embarqué de synthèse de la parole arabe. La conception de ce type de système nécessite un double effort sur deux fronts

différents : la synthèse proprement dite mais aussi les contraintes d'embarquabilité du système à proposer.

Dans cet article, nous allons nous limiter à la présentation de l'étape de transcription. Nous décrirons les différentes étapes et composantes nécessaires à une telle réalisation.

La transcription orthographique-phonétique est indispensable pour la mise en oeuvre des applications de synthèse à partir du texte telles que les applications dans les machines à lire pour les aveugles ou de lecture du courrier électronique. Lorsque nous lisons à haute voix un texte, nous effectuons naturellement une transcription orthographique-phonétique avant de commander notre appareil vocal. Il est en fait difficile de connaître le détail de notre démarche et même de connaître comment sont structurées les principales phases du traitement. Plusieurs niveaux doivent être pris en compte : niveaux phonétique et phonologique, niveaux lexical, syntaxique et même sémantique. Il est vraisemblable que certaines de ces analyses sont effectuées en parallèle.

2 Les constituants d'un système de synthèse de la parole

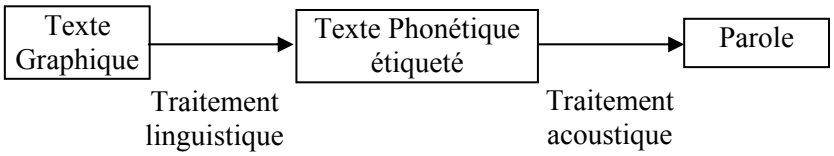

Figure 1 : Schéma de principe de notre système de synthèse de la parole.

Notre système de synthèse de la parole est divisé en deux grandes parties : une partie linguistique et une partie acoustique. La partie linguistique ou symbolique permet, à partir d'un texte écrit, de générer un texte phonétique étiqueté. Ce passage comprend quatre étapes:

1. Le prétraitement du texte (espaces, paragraphes, ponctuation, chiffres, etc.).
2. La consultation du lexique des exceptions pour l'élimination des mots spéciaux.
3. L'application des règles de transcription établies pour la langue arabe.
4. La conversion graphème-phonème.

A l'issue du traitement linguistique, nous aboutissons à un texte phonétique étiqueté qui sera archivé dans une base de données en vue de son utilisation lors des phases de test et d'évaluation.

3 La langue arabe

L'alphabet arabe comporte 28 consonnes et 6 voyelles de l'arabe standard (3 longues et 3 courtes) et quelques autres réalisations vocaliques. Nous pouvons classer les consonnes selon plusieurs critères : des consonnes articulées avec une vibration des cordes vocales et des consonnes qui n'engendrent pas une vibration des cordes vocales, le franchissement de l'air à travers le conduit vocal donne naissance à d'autres variétés de sons. Mais pour les besoins de la transcription les 28 consonnes arabes ont été divisées en deux groupes :

- 14 consonnes solaires qui assimilent le « ـلـ » de l'article.
- 14 consonnes lunaires qui n'assimilent pas le « ـلـ » de l'article.

Solaires	Lunaires
ت ث د ذ ر ز س ش ص ض ط ظ ل ن	أ ب ج ح خ ع غ ف ق ك ه م و ي

Figure 2 : Classification des consonnes tenant compte des contraintes de la transcription.

En ce qui concerne les voyelles de l'arabe, on distingue trois voyelles courtes et trois voyelles longues. La durée d'une voyelle longue est environ double de celle d'une voyelle courte. Ces voyelles sont caractérisées par la vibration des cordes vocales. Elles sont représentées dans le tableau suivant :

Courtes	Longues
ُ / ِ / َ /	ا ِي ُو

Figure 3 : Classification des voyelles de la langue arabe.

3.1 La transcription de la langue arabe et ses problèmes

Certaines langues sont plus faciles que d'autres pour effectuer cette transcription. Un compromis a du être trouvé entre les tailles respectives du dictionnaire de règles de transcription à vocation générale et du dictionnaire de mots comportant les exceptions les plus courantes et facilitant l'analyse syntaxique souvent nécessaire de la phrase.

Le tableau suivant représente l'ensemble des caractères de l'arabe avec pour chaque graphème, sa transcription phonétique.

Gr.	Ph.	Gr.	Ph.	Gr.	Ph.	Gr.	Ph.	Gr.	Ph.	Gr.	Ph.
أ	E	خ	x	ش	ʃ	غ	ɣ	ن	n	ُ	u
ب	b	د	d	ص	ʂ	ف	f	ه	h	ِ	i
ت	t	ذ	ð	ض	ɖ	ق	q	و	w	ً	an
ث	θ	ر	r	ط	ʈ	ك	k	ي	j	ٌ	un
ج	ʒ	ز	z	ظ	ðˤ	ل	l	ء	a	ٍ	in
ح	ħ	س	s	ع	ʕ	م	m	َ	a	ْ	silence

Figure 4 : Correspondance graphème phonème de la langue arabe suivant l'alphabet phonétique internationale IPA 96.

La complexité du problème de transcription peut être appréciée après avoir examiné le tableau des correspondances graphèmes-phonèmes. Ces problèmes sont liés à la langue traitée elle-même et sont de différentes sortes :

- Des graphèmes qui ont plusieurs réalisations phonétiques. Le w و de « بُومٌ » et « مَوْزٌ » correspond à deux sons différents ayant la même graphie.
- Des phonèmes qui ont plusieurs réalisations graphémiques. Le noun dans «يُقَيمُونَ» et dans «أنْزَلَ» n'a pas la même représentation graphémique.
- Des graphèmes qui ne sont pas pris en compte. Le alif dans «قَامُوا» ne correspond pas à un son (silence).
- Une absence totale de correspondance graphème-phonème. Le mot «هَذَا» devrait être écrit «هَاذَا».

3.2 La connaissance de la langue arabe

Une des premières recherches qu'on a effectuée avant toute autre démarche consiste à formaliser au mieux les problèmes posés par la langue arabe. Il s'agit tout d'abord de disposer de données grammaticale et linguistique de la langue arabe pour pouvoir établir des règles, les listes de mots d'exception et en générale de formaliser le problème posé.

4 La transcription proprement dite

Plusieurs phases sont à distinguer :

4.1 Le repérage des mots

Les séparateurs (blancs, tirets...) permettent le repérage des mots. Les séparateurs (virgules, points, deux points, points d'interrogation, d'exclamation...) sont aussi traités pour l'analyse des pauses et des arrêts.

4.2 L'utilisation d'un lexique

Le lexique traite les exceptions qui ne peuvent être prises en compte par les règles. En effet on ne peut pas se permettre de dépenser un nombre important de règles pour le traitement spéciales de quelques mots. Un compromis est alors nécessaire.

4.3 L'utilisation de règles

Avant d'écrire les règles de transcription, on aura soin de définir un certain nombre de catégories qui faciliteront l'écriture de ces règles. Par exemple : les consonnes solaire ou lunaire, les voyelles, etc. Ces règles traitent principalement les problèmes suivants :

- l'épellation,
- la liaison ou non,
- la prononciation des géminées,
- l'élision en syllabe initiale du mot, en syllabe intérieure du mot, l'élision en finale,
- la prononciation des graphèmes à l'intérieur d'un mot en fonction du contexte dans lequel il est représenté par un son particulier ou bien une catégorie de sons avec d'éventuelles exceptions, etc.

Notre analyse linguistique nous a permis d'établir pour notre système un ensemble de 133 règles. Il est à noter que l'ordre d'application de ces règles est trés important et influe énormément sur le résultat final. L'ensemble de ces règles est stocké dans une table de notre base de données. En voici quelques exemples :

- Au début d'un mot, le alif devient un hamaza (Exemple : الوَلَدُ → Elwaladu).
- Le alif est supprimé quand il est suivi d'un caractère lunaire (Exp: طَلَعَ القَمَرُ → طَلَعَ لقَمَرُ).

4.4 La conversion graphème phonème :

Cette phase concerne le passage du texte graphique au texte phonétique (après le traitement des exceptions et des règles de prononciation) suivant la table de correspondance (figure 2). Cette table est constituée de 49 enregistrements qui correspondent aux différents caractères de l'alphabet arabe et des quelques symboles utilisés.

5 Les outils de transcription

Plusieurs techniques peuvent être exploitées pour faciliter les opérations décrites ci-dessus. Il est plus simple d'utiliser une grammaire contextuelle. Elle peut faciliter l'écriture, puis la lecture et la modification, des règles avec un contrôle automatique de la syntaxe des règles utilisées. Une règle, se lit de droite à gauche et doit être formalisée de la façon suivante :

[Phonème] = {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 et CL est une consonne lunaire.

Généralement, les règles peuvent indiquer des classes de phonèmes ou de graphèmes (voyelles, consonnes, fricatives, chiffres...). De même, les règles sont ordonnées, la suite de leur application est importante.

5.1 Présentation d'un échantillon de règles

En ce qui suit la description de quelques règles parmi les 133 élaborées et incorporées dans la base des règles :

- [uu]={CS}+{ُ}+{و} [uu]={CL}+{ُ}+{و} [uu]= §+{ُ}+{و} [uu]= $+{ُ}+{و}

Lorsque leو est précédé par la voyelleُ et qu'il est suivi par une consonne, on obtient le phonème de la voyelle longue [uu]. Lorsque leو est précédé par la voyelleُ et qu'il est en fin de mot, on obtient le phonème de la voyelle longue [uu].

Exemple : دُونَ لَبَسُوا

- [aa]={َ}+{ـا}

Lorsque leـا est précédé par la voyelle َ, on obtient le phonème de la voyelle longue [aa] quelque soit ce qui suit.

Exemple : لِمَا

- [ii]={CL}+{ِ}+{ي} [ii]={CS}+{ِ}+{ي} [ii]=§+{ِ}+{ي} [ii]=$+{ِ}+{ي}

Lorsque leي est précédé par la voyelle ِ et qu'il est suivi par une consonne, on obtient le phonème de la voyelle longue [ii]. Lorsque leي est précédé par la voyelleِ et qu'il est en fin de mot, on obtient le phonème de la voyelle longue [ii].

Exemple : قَلِيلاً لَمَسَنِي

- [CC]={ّ}+{C}

Lorsqu'une consonne est suivie par la ّ, elle est doublée, on obtient alors le phonème [CC].

Exemple : وَدَّ

- [C]={ْ}+{C}

Lorsqu'une consonne est suivie par ْ, elle reste inchangée, on obtient alors le phonème [C].

Exemple : يَرْغَبُ

- {CS}+{أ}={CS}+{الـ}+# {CL}+{لـ}+{CL}={CL}+{الـ }+{CL}

Lorsque le الـ est en début de phrase et qu'il est suivi par une consonne solaire, il est équivalent à la non présence du لـ . Lorsque le الـ est entre deux consonnes lunaires, il est équivalent à la non présence du أ .

Exemple : السَميعُ مُنِعَ الأكْلُ

5.2 Vérification

Les procédures de vérification sont certainement aussi importantes que le système lui-même car elles permettent de l'améliorer en précisant le rôle et l'importance de chacune des règles et en mettant en évidence les disfonctionnements. Les phrases traitées et archivées dans la base de données permettent de faire cette évaluation.

6 CONCLUSION

Nous avons présenté dans ce document les outils nécessaires ainsi que les méthodes utilisées lors de l'élaboration des modules de prétraitement et de traitement phonologique, tous deux opérationnels. La base de données des règles de transcription ne cesse de s'enrichir en vue d'un meilleur résultat.

Néanmoins, l'aboutissement au système escompté reste long et devra passer par un nombre important de tests, par l'approfondissement des outils de programmation sur puces ainsi que par la résolution des problèmes de taille de mémoire généralement de mise pour de tels systèmes.

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.