Etude et Evaluation de la Di-syllabe comme Unité Acoustique pour le Système de Synthèse Arabe PARADIS

N. Chenfour[1], A. Benabbou[2], A. Mouradi[3]

[1] Faculté des Sciences de Fès, chenfour@yahoo.fr
[2] Faculté des Sciences et Techniques Fès, abenabbou@yahoo.fr
[3] ENSIAS Rabat, mouradi@ensias.um5soussi.ac.ma

Résumé

L'étude que nous présentons dans cet article s'inscrit dans le cadre de la réalisation d'un système de synthèse de la parole à partir du texte pour la langue arabe. Notre système PARADIS est basé sur la concaténation des di-syllabes avec TD-PSOLA comme technique de synthèse. Nous présentons dans cet article l'intérêt du choix de la di-syllabe comme unité de concaténation pour le synthétiseur et son apport au niveau de la qualité de synthèse. En effet, la di-syllabe permet d'améliorer amplement la qualité de synthèse et de réduire les problèmes de discontinuité temporelle lors de la concaténation. Cependant, on est confronté à plusieurs problèmes causés par la taille considérable de l'ensemble des di-syllabes et leur adaptation aux modèles prosodiques qui sont d'habitude associés à la syllabe comme unité rythmique. Nous décrivons alors le principe sur lequel nous nous sommes basés pour réduire le nombre de di-syllabes. Nous présentons ensuite la démarche que nous avons mise au point pour la génération et l'étiquetage automatique du dictionnaire de di-syllabes. Ainsi, nous avons choisi des logatomes ayant des formes particulièrement appropriées à l'automatisation de la procédure de génération du corpus des logatomes et à l'opération de segmentation automatique. Par ailleurs, nous présentons une technique d'organisation du dictionnaire acoustique parfaitement adaptée à la forme de la di-syllabe arabe.

1. Introduction

Notre système de synthèse PARADIS (Psola ARAbic DI-syllable concatenation based System) est composé de plusieurs modules (Figure 1). Le premier module concerne la transcription graphèmes-phonèmes qui consiste à lire un texte arabe voyellé et le transcrire en un texte phonétique correspondant. Ce module a été généré automatiquement par notre compilateur de règles LSPERT (Langage de SPEcification des Règles de Transcription) à partir d'une spécification formelle des règles de transcription (Benabbou, 1997; Chenfour, 1997). Le texte phonétique est ensuite traité par un module prosodique permettant d'y insérer des marqueurs de variation de pitch et de durée. La phase suivante consiste à découper le texte phonétique en di-syllabes. Un module d'accès direct basé sur une technique de Hash-code permet, d'une part, l'organisation du dictionnaire de di-syllabes lors de la phase d'analyse, d'autre part, l'extraction rapide des unités acoustiques au cours de la synthèse. Après décodage des données extraites, le synthétiseur TD-PSOLA permet enfin de générer le signal vocal correspondant au texte d'entrée.

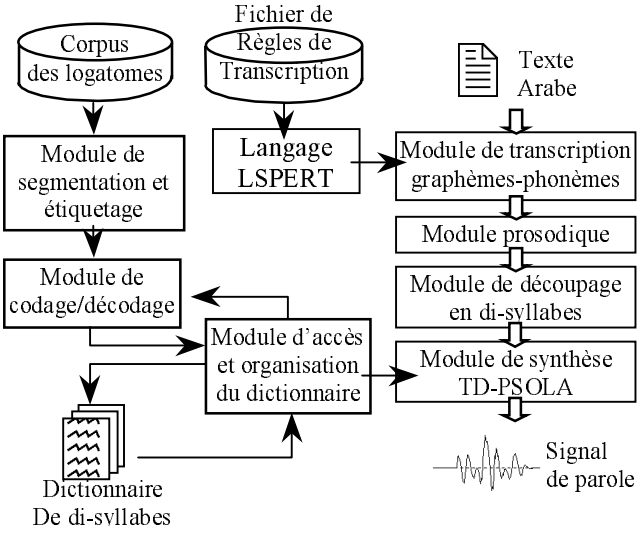

Figure 1 : Architecture générale du système PARADIS

L'élaboration du dictionnaire est une phase déterminante pour le système de synthèse. La première étape consiste à enregistrer un corpus de parole à partir duquel seront extraites toutes les unités acoustiques. Nous avons réalisé l'enregistrement du corpus à base de logatomes contenant chacun une di-syllabe. Etant donnée la grande taille du corpus obtenu, nous avons adopté une démarche complètement automatique en vue d'une génération rapide du dictionnaire. Elle comprend l'opération de segmentation du corpus en di-syllabes et l'étiquetage (marquage de pitch) des segments obtenus.

2. Catégorisation des Unités Acoustiques de Concaténation

Un système de synthèse de la parole à partir du texte par concaténation nécessite, pour obtenir une synthèse de haute qualité, un choix judicieux de la technique de synthèse et de l'unité acoustique de concaténation ainsi qu'une préparation minutieuse du dictionnaire d'unités acoustiques. La technique TD-PSOLA que nous avons implémentée présente, grâce à la notion de superposition des signaux à court-terme, des possibilités importantes de traitement sur le signal de parole. En outre, l'opération de superposition des signaux à court-terme attribue à TD-PSOLA la faculté d'interpolation entre les segments du signal de parole. Ces segments ou unités acoustiques doivent être bien choisis afin de mieux exploiter les possibilités de l'algorithme TD-PSOLA et offrir une bonne qualité de synthèse.

Une solution raisonnable serait le choix d'une unité contenant l'articulation entre deux phonèmes consécutifs. L'unité minimale assurant cette contrainte est le diphone. Cette unité permet lors de la synthèse par concaténation de reproduire une partie de la dynamique de la production de la parole. Ainsi, plusieurs systèmes ont adopté le diphone comme unité de concaténation (Mouradi, 1989; Moulines, 1990; Dutoit, 1992) . En effet, le diphone est une unité très courte acoustiquement et engendre un nombre de combinaisons acceptable (environ 1300 pour le français et 1200 pour l'arabe). Il en découle que la capacité de stockage nécessaire est fort convenable. On a retenu (d'après Moulines) pour le français un volume de stockage entre 5 et 10 Mo sans codage avec une fréquence d'échantillonnage de 16 KHz. Pour l'arabe, on peut retenir une capacité nécessaire de l'ordre de 6 à 8 Mo.

Cependant, les études ont montré que le diphone est une unité non suffisante pour transporter tout le phénomène de coarticulation (d'après Moulines). Les tests d'intelligibilité ont montré que certains groupes consonnantiques restent mal perçus par des auditeurs non expérimentés. En plus, certains diphones présentent des discontunuités perceptibles au niveau de la région de concaténation.

Les unités plus longues assurent en général une synthèse de meilleure qualité car elles intègrent le phénomène de coarticulation à plus long terme. Pour la langue arabe, nous avons choisi comme unité de base la di-syllabe. Celle-ci a donné de meilleurs résultats au prix d'un nombre d'unités beaucoup plus élevé que celui des diphones.

3. Description Fonctionnelle de la Di-Syllabe

La définition d'une di-syllabe (Chenfour, Mouradi, Benabbou, 1997) ressemble à la définition du diphone projetée sur l'échelle de la syllabe:

"Une di-syllabe est la transition du noyau vocalique d'une syllabe vers le noyau vocalique de la syllabe suivante".

Les mots arabes ne contiennent jamais plus que deux consonnes consécutives, nous obtenons alors six formes possibles d'une di-syllabe : CV, VC, VCC, V#, VCV, VCCV. C étant un représentant de la classe des consonnes (au nombre de 28) et V représente la classe des voyelles (courtes et brèves, au nombre de 6). La forme CV se trouve au début des mots. Les formes VCV et VCCV occupent des positions médianes. Enfin les formes V#, VC et VCC figurent en position finale.

Ces unités comportent d'une part la coarticulation de voyelle à voyelle à travers les consonnes. D'autre part elles améliorent la procédure de concaténation qui juxtapose uniquement les parties stables d'une même voyelle. Ainsi, on évite le problème d'interpolation NV/NV (Non-Voisé/Non-Voisé).

Par ailleurs, la concaténation doit s'accompagner d'une interpolation efficace pour éliminer les distorsions et discontinuités spectrales qui résultent du fait que les 2 di-syllabes sont extraites de deux contextes différents. Le modèle TD-PSOLA se trouve alors bien adapté à cette opération. L'opération de recouvrement-addition (OLA) entre le dernier signal à court-terme de la pré-di-syllabe et le premier signal à court terme de la post-di-syllabe, qui sont tous deux des signaux voisés, sera considérée comme une interpolation linéaire entre les deux segments.

4. Réduction du Nombre de Di-syllabes

Le choix de la di-syllabe donne une synthèse de meilleure qualité. Néanmoins, on remarque que la taille du dictionnaire acoustique correspondant est très grande. Avec 28 consonnes et 6 voyelles, nous obtenons le tableau des combinaisons suivant :

V	C	VC	CV	VCV	VCC	VCCV	Total
6	28	168	168	1008	4704	28224	34278

Tableau 1 : Bilan des di-syllabes arabes.

Le tableau indique un total de 34278 unités. Nous sommes alors amenés à réduire ce nombre.

- Première réduction :

Une première réduction consiste à considérer toutes les unités uniquement avec des voyelles courtes. En ce qui concerne les voyelles longues, elles seront générées automatiquement par le synthétiseur en utilisant un mécanisme de duplication de périodes de la partie stable de la voyelle courte correspondante. Ceci réduit énormément la combinatoire (Tableau 2).

V	C	VC	CV	VCV	VCC	VCCV	Total
3	28	84	84	252	2352	7056	9744

Tableau 2 : Bilan des di-syllabes arabes, 1ère réduction

- Deuxième réduction :

Dans la langue arabe, certaines consonnes ne se suivent jamais pour constituer l'une des formes VCCV ou VCC. Le nombre total de cas que nous avons recensés est 76. Combiné avec les formes VCC et VCCV il donne un total de 912, ce qui ramène la taille du dictionnaire à 8832 unités.

5. Projection des Facteurs Prosodiques sur l'Axe des Di-Syllabes

Dans le module de traitement prosodique, nous avons constaté un problème d'incohérence entre l'unité rythmique «la syllabe », à laquelle sont affectés généralement les facteurs des variations prosodiques, et notre unité acoustique «la di-syllabe ». Nous avons résolu ce problème à l'aide d'une projection/interpolation.

Un mot de N syllabes S_1, S_2, S_3, ..., S_N (mot = $S_1.S_2.S_3...S_N$) contient N+1 di-syllabes (DS_1, DS_2, DS_3, ..., DS_{N+1}). Avec DS_i est la transition de S_{i-1} à S_i pour tout i>1, DS_1 est la transition Silence-S_1, et DS_{N+1} est la transition S_N-Silence. On note F_i un facteur prosodique associé à S_i. La correspondance des facteurs est effectuée en associant à chaque DS_i le facteur F_i de S_i, et à DS_{N+1} le facteur F_N. Les di-syllabes du mot seront alors décrites par (F_1, F_2, ... F_{N-1}, F_N, F_N). Mais l'application d'un facteur prosodique F_i à la di-syllabe DS_i sera effectuée après une interpolation demi-cosinus avec le facteur F_{i-1}. Le principe est expliqué à l'aide de la figure 2 sur un exemple :

$$\underset{Mot}{} = \overset{Syllabes}{X.Y.Z.T} = \overset{Di-syllabes}{\#x.xy.yz.zt.t\#}$$

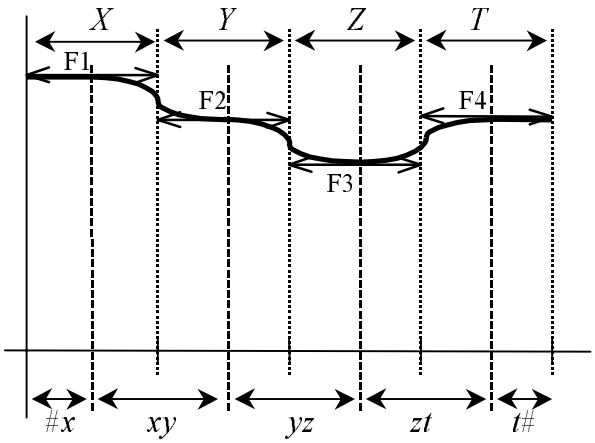

Figure 2 : Schéma de la projection sur les di-syllabes des facteurs prosodiques associés aux syllabes d'un mot.

6. Génération Automatique du Dictionnaire des Di-syllabes

Un système de synthèse de la parole à partir du texte par concaténation nécessite une préparation minutieuse et exhaustive du dictionnaire d'unités acoustiques pour obtenir une synthèse de haute qualité. Cependant, une génération manuelle est une opération fastidieuse et difficile à maîtriser. Pour cela nous avons procédé de manière automatique. La démarche consiste en une procédure à trois phases principales : phase de préparation et enregistrement du corpus, phase de segmentation et marquage de pitch et phase de construction et organisation du dictionnaire.

6.1 Préparation et Enregistrement du Corpus

6.1.1 Préparation des Logatomes

Nous avons élaboré un corpus de logatomes contenant chacun une unité di-syllabe. Le choix d'un logatome comme mot porteur d'une di-syllabe est justifié par les critères suivants :

- Utilisation d'un contexte neutre.
- Automatisation de la procédure de génération du corpus des logatomes.
- Facilité de segmentation automatique, car le contexte d'une di-syllabe est connu et uniforme. Ceci permet l'implémentation d'une procédure de segmentation générale et indépendante du logatome.

Pour chaque famille de di-syllabes nous avons choisi un modèle de logatomes correspondant (Tableau 3).

Di-syllabe	*Modèle du Logatome*	*Exemples de logatomes*
CV	CV+/sasa/	easasa, basasa, tusasa
VC	/sat/+VC	satab, satub, satib
VCV	/t/+VCV+/sa/	tabasa, tajisa, tuxasa
VCCV	/t/+VCCV+/sa/	tabrasa, turrusa

Tableau 3 : Modèles de logatomes associés aux différentes di-syllabes

Les unités V# sont tirées facilement du même corpus. Tandis que les unités VCC n'ont pas été considérées provisoirement car nous avons supposé en entrée un texte complètement voyellé. Dans ce cas, on ne rencontre jamais des unités VCC.

6.1.2 Enregistrement du Corpus

L'enregistrement du corpus à été réalisé par un locuteur bien entraîné, avec un Micro-Shure en se servant d'un DSP. Notre locuteur a essayé au mieux de garder un même rythme, même intensité et sans introduction d'effets mélodiques. Nous avons réalisé l'enregistrement par groupe de 28 logatomes ayant un même modèle.

6.2 Segmentation Automatique en Di-syllabes

L'opération de segmentation en di-syllabes consiste à extraire à partir du corpus toutes les di-syllabes qui sont à la base de l'élaboration du dictionnaire de synthèse. Nous avons opéré à l'aide d'une procédure complètement automatique. La procédure est basée sur la détection des frontières des logatomes, du voisement pour identifier les voyelles et du non voisement pour l'identification des phonèmes /t/ ou /s/ qui délimitent la di-syllabe dans le logatome.

6.2.1 Extraction des Unités VCV et VCCV

Les unités VCV et VCCV en nombre de 7308, constituant la majorité des di-syllabes (74.34% de la totalité des di-syllabes et 97.71% si on néglige la forme VCC), créent une véritable nécessité de segmentation automatique. Ces deux types de di-syllabes peuvent être considérés de la même manière car toutes les deux sont délimitées par des voyelles et sont enveloppées par le même contexte. Nous les avons alors traités par un même processus de segmentation automatique.

Les logatomes analysés par cette procédure sont de la forme : /t/+V_1CV_2+/sa/ ou /t/+V_1CCV_2+/sa/. La segmentation est réalisée à l'aide d'un pointeur de détection de voisement comme indiqué sur la figure 3.

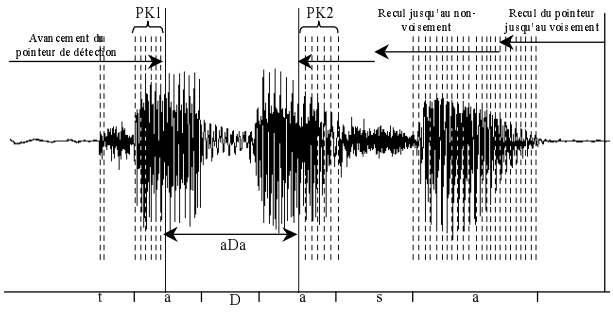

Figure 3 : Extraction d'une unité VCV (/aDa/) à partir du logatome /taDasa/. PK_1 et PK_2 indiquent les nombres de marques de pitch qui permettent d'atteindre les frontières de la di-syllabe.

6.2.2 Extraction des Unités VC et CV

Les logatomes analysés par cette procédure sont de la forme : /sat/+VC et CV+/sasa/. La segmentation est réalisée en respectant le même principe d'avancement du pointeur de détection à la recherche de signal voisé ou non voisé.

6.3 Marquage de Pitch

Il est primordial de respecter les exigences de l'analyse pitch synchrone. Le signal doit être muni d'une suite de marques de pitch distribuées de façon synchrone

avec la fréquence fondamentale sur les portions voisées du signal. Sur les portions non voisées, les marques de pitch sont distribuées arbitrairement ou à une cadence uniforme. Dans notre cas, nous avons utilisé un marquage toutes les 10 ms.

Respectant ces critères, nous avons alors développé une procédure automatique à l'aide de laquelle le marquage de pitch a été entièrement réalisé. La procédure de marquage est basée sur la méthode temporelle AMDF (Average Magnitude Difference Function) (Boris, 1994). Le principe de la méthode est basé sur la grande ressemblance entre les périodes de pitch successives d'un signal voisé. Cette pseudo-périodicité est traduite par une valeur d'énergie minimale de la différence entre les signaux de deux périodes successives. L'énergie ou la fonction de périodicité est donnée par l'expression suivante :

$$f(p) = \frac{\sum_{i=1}^{T} |S(i+p) - S(i)|}{\sum_{i=1}^{T} |S(i)|}$$

Le travail à faire est de déterminer la valeur de p pour laquelle la fonction atteint son minimum sur une période T. Nous avons pris la valeur 128 ms comme valeur de T. Des corrections de décision sont parfois nécessaires (si une période voisée est trouvée dans une portion non voisée ou inversement). Enfin, un calcul du maximum est réalisé au niveau de la portion considérée voisée. La marque de pitch est alors obtenue.

Signalons enfin que nous avons développé une autre procédure moins complexe, basée seulement sur le calcul du nombre de passage par zéro et du maximum local avec une intensité significative. Le résultat de cette procédure est semblable à celui de la méthode AMDF.

7. Dictionnaire Acoustique des Di-syllabes

Pour améliorer le temps d'accès au dictionnaire des di-syllabes, nous l'avons organisé sous forme d'un fichier à accès direct. Cependant la taille du bloc de données représentant chaque di-syllabe est variable. Nous avons donc utilisé une organisation en un fichier indexe et un fichier de blocs de données. Les indexes des différents blocs de données sont des clés d'entrée dont l'accès est calculé par une technique de Hash-Code à partir de la forme de la di-syllabe.

7.1 Hash-Code Utilisé

Le système phonétique de la langue arabe est constitué de 6 voyelles et 28 consonnes. Nous avons introduit une voyelle supplémentaire neutre notée ***v*** et une consonne supplémentaire neutre notée **ç** pour avoir un code homogène pour toutes les di-syllabes. Ainsi, toutes les di-syllabes auront la forme générale unifiée obtenue de la manière suivante :

V ⇔ V**ç**_**ç**_***v***
VC ⇔ VC**ç**_***v***
CV ⇔ ***v*****ç**CV
VCV ⇔ VC**ç**V
VCC ⇔ VCC***v***
VCCV ⇔ VCCV

En outre, 3 bits sont suffisants pour coder toutes les voyelles et 5 bits pour coder toutes les consonnes. Une di-syllabe quelconque, ramenée à la forme VCCV, peut alors être représentée par un code sur 16 bits qui la détermine. Ce Hash-code H(VCCV) est une simple juxtaposition des codes des différents éléments de la di-syllabes en base {3bits, 5 bits, 5 bits, 3 bits}.

V	C	C	V
3 bits	5 bits	5 bits	3 bits

7.2 Module d'Accès et Organisation du Dictionnaire

Etant donnée une di-syllabe VCCV, on calcule tout d'abord son code H(VCCV). Le code représente une clé d'accès direct dans le fichier indexe. On récupère la clé d'entrée dans le dictionnaire de paramètres, ainsi que la taille du bloc à extraire. Enfin, avec un accès direct au fichier de données, on extrait les informations recherchées (Figure 4).

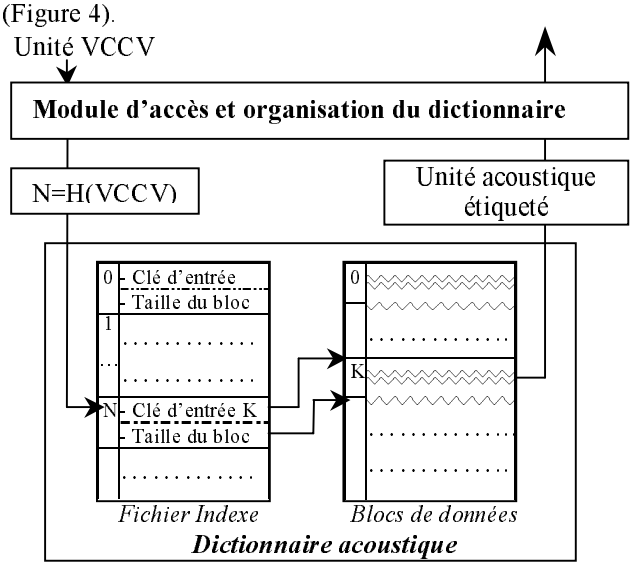

Figure 4 : Accès et organisation du dictionnaire de di-syllabes.

8. Conclusion

Dans cette contribution, nous avons exposé les avantages et l'apport de la di-syllabe comme unité acoustique de concaténation dans notre système de synthèse PARADIS. Elle a ainsi permis une meilleure exploitation de l'opération d'interpolation de TD-PSOLA. La parole synthétique obtenue est de très bonne qualité, l'intelligibilité et le naturel sont atteints.

En outre, le procédé automatique de génération du dictionnaire acoustique que nous avons réalisé, nous offre la possibilité d'obtenir de manière systématique et rapide des dictionnaires multi-locuteurs. Nous avons évalué le taux global de succès de la procédure de segmentation à 98%. Pour les unités mal segmentées, nous avons procédé seulement par réajustement des paramètres de prédiction de l'algorithme.

La recherche des segments de parole dans le dictionnaire lors de la synthèse est effectuée avec une grande rapidité grâce à l'organisation particulière du dictionnaire de di-syllabes à l'aide du code di-syllabique.

Références

Benabbou A. (1997): implémentation sur PC d'un système à formant pour la synthèse par règles. Thèse de 3ème cycle. Faculté des sciences Rabat. Université Mohammed V.

Boris D. (1994) : Estimation de la Fréquence Fondamentale des signaux sonores. Thèse pour l'obtention du titre de docteur de l'université Paris VI. pp. 28-33.

Chenfour N., Mouradi A., Benabbou A. (1997): Synthèse de la Parole Arabe par Concaténation de Di-syllabes. Journées Sientifiques et techniques du réseau Francophone de l'Ingénierie de la langue de l'AUPELF-UREF. Avignon France. pp. 459-462.

Chenfour N. (1997) : Réalisation d'un Système de Synthèse de la Parole Arabe à partir du texte par Concaténation de Di-syllabes. Thèse pour l'obtention du titre de docteur de troisième cycle. Faculté des sciences Rabat. Université Mohammed V.

Dutoit T., Leich H. (1992): Improving the TD-PSOLA text-to-speech synthesizer with a specially designed MBE re-synthesis of the segments database. Proc. European Signal Processing Conf., Bruxelles, pp. 343-346.

Moulines E. (1990) : Algorithmes de codage et de modification des paramètres prosodiques pour la synthèse de la parole à partir du texte. Thèse de Doctorat. ENST Paris.

Mouradi, A. (1989): Validité et limites du diphone en tant qu'unité de synthèse pour la langue arabe standard. Journal Acoustique 2 (pp. 307-309).