Un système de lissage linéaire pour la synthèse de la parole arabe : Discussion des résultas obtenus

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

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

Mots clés – Keywords

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

Arabic speech sythesis, Phoneme, Diphones, Triphones, Acoustic units, Polyphones dictionary.

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 de notre système de synthèse hybride de la parole arabe. Nous présenterons, dans ce papier, les différents modules et les différents choix techniques de notre système de synthèse hybride par concaténation de polyphèmes. Nous détaillerons également les règles de transcription et leurs effets sur le traitement linguistique, les règles de syllabation et leurs impacts sur le coût (temps et difficulté) de réalisation du module acoustique et nous poursuivrons par l'exposé de nos choix au niveau du module de concaténation. Nous décrirons le module de lissage, un traitement acoustique, post concaténation, nécessaire à l'amélioration de la qualité de la voix synthétisée. Enfin, nous présenterons les résultats de l'étude statistique de compréhension, réalisée sur un corpus.

This research paper is within the project entitled "Oreillodule" : a real time embedded system of speech recognition, translation and synthesis. The core of our interest in this work is the presentation of the hybrid system of the Arabic speech synthesis and more precisely of the linguistic and the acoustic treatment. Indeed, we will focus on the grapheme-phoneme

transcription, an integral stage for the development of this speech synthesis system with an acceptable quality. Then, we will present some of the rules used for the realization of the phonetic treatment system. These rules are stocked in a data base and browsed several times during the transcription. We will also 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 the establishment of this dictionary and the difficulties faced during its development. Finally, we will present the results of the statistical survey of understanding, achieved on a corpus.

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. Cet objectif a nécessité l'étude de toutes les étapes de la synthèse de la parole et le choix des solutions les plus adaptées à chaque tâche. Le résultat de ces études nous a guidé vers un système de synthèse hybride 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 transcripteur, le module de syllabation, le dictionnaire d'unités acoustiques et le module de concaténation muni de son système de lissage (Dutoit, 1993).

2 LA TRANSCRIPTION

L'analyse linguistique nous a permis d'établir un ensemble de 133 règles. Il est à noter que l'ordre d'application de ces règles est très important et influe sur le résultat final. En ce qui suit la description de quelques règles élaborées (Saidane, 2004) :

1. [CC]={ّ}+{C}

Lorsqu'une consonne est suivie par laّ, elle est doublée, on obtient alors le phonème [CC]. Exemple : وَدَّ , رَوَّجَ.

2. {CL}+{ـل}+{V}+{CL}={CL}+{الـ }+{CL}
3. {CL}+{ـل}+{V}+{CS}={CL}+{الـ }+{CS}

Lorsque le الـ est entre suivi par une consonne lunaire, il est équivalent à la non présence du أ. Exemple : مُنعَ الأكل, أكِلَ الأكلُ (Zrigui 1991).

3 LA SYLLABATION

Les unités acoustiques de notre système de synthèse sont de trois types : les triphones, les diphones et les phonèmes. On a établi un ensemble de règles de concaténation à partir desquelles les différentes occurrences de trois phonèmes pouvaient se transformer en : un triphone, un diphone suivi d'un phonème, un phonème suivi d'un diphone, ou éventuellement trois phonèmes. 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 (Boula, 2001). Le schéma suivant présente un exemple de syllabation pour l'expression «أيْنَ المُسَافِرُونَ » (eaj.na.lmusaafiruuna[1]: Où sont les voyageurs) (Saidane, 2004):

[1] Suivant l'alphabet phonétique internationale IPA 96

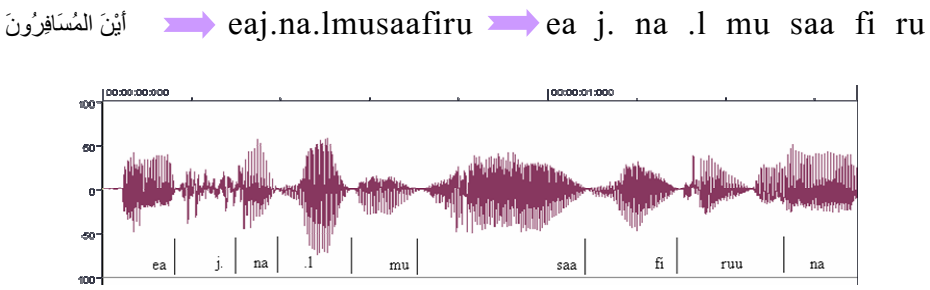

Fig. 1. Exemple de syllabation

La problématique de la sélection des unités a été formalisée en utilisant des règles. Ces règles de syllabation peuvent se résumer en ce qui suit :

1. [CVV] ={V}+{V}+{C} : lorsqu'une consonne est suivie de deux voyelles les trois graphèmes constituent une unité acoustique de notre système.
2. [CV]={C}+{V}+{C} : lorsqu'une consonne est suivie d'une voyelle puis d'une consonne les deux premiers graphèmes constituent une unité acoustique.
3. [CC]={C}+{C}+{C} : lorsque nous avons une succession de trois consonnes les deux premiers graphèmes constituent une unité acoustique.
4. [C]={V}+{C}+{C} : lorsque nous avons deux consonnes suivies par une voyelle seul le premier graphème constitue une unité acoustique.
5. [VV]={V}+{V} : lorsque nous avons une succession de deux voyelles, les deux constituent une unité acoustique.
6. [V]={V} : lorsque nous avons une voyelle isolée elle constitue une unité acoustique.

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 (Emerard, 1977). Ces six règles de syllabation élaborées vont imposer les types d'unités acoustiques à utiliser pour la synthèse de la parole. Le dictionnaire ainsi établi contient 196 unités acoustiques suffisantes pour la réalisation des différentes occurrences possibles. Le nombre de phonèmes est de 28, le nombre de diphones est de 84 et le nombre de triphones est de 84.Néanmoins, la pratique et l'étude de la langue arabe ont permis de dégager une dizaine d'autres unités dues principalement aux contraintes de la langue.

Le module de concaténation a besoin de la totalité des unités acoustiques sous la forme d'enregistrements sonores (Lemmety, 2000). Ces enregistrements constituent le dictionnaire de notre système. 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).

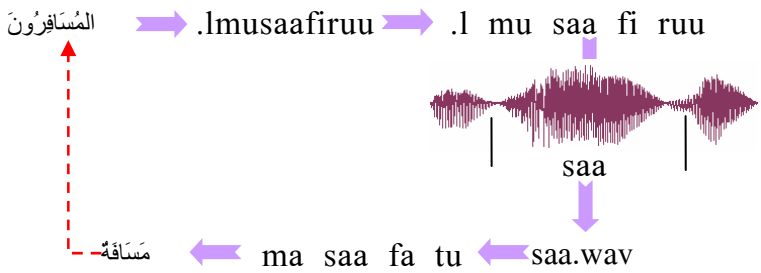

Fig. 2. Un exemple de traitement pour l'obtention du triphone « saa »

4 LA CONCATENATION

Pour notre système nous avons voulu commencer par un traitement de lissage temporel pour mesurer l'effet d'un post traitement sur la qualité de la parole obtenue. Après l'analyse des différentes unités acoustiques de l'arabe, il s'avère que celles-ci présentent une atténuation aux niveaux de leurs extrémités. L'idée retenue consiste alors à procéder, lors de la concaténation, à une accentuation aux niveaux d'un certain nombre de valeurs d'extrémités avant le collage en bout à bout. Ce traitement touchera évidement la fin de la première unité et le début de la suivante.Un signal numérique de la parole étant :

$$s(t) = \sum_{1}^{N} s_n \, \delta(t - nT) \qquad (1)$$

s(t) : signal numérisé de la parole (échantillonné), $s_n = s(nT)$: la valeur du signal à l'instant nT et δ(t) : impulsion de Dirac. La concaténation de deux unités sera :

$$s(t) = s_1(t) + s_2(t) = \sum_{1}^{N} s_{1n} \, \delta(t - nT) + \sum_{1}^{M} s_{2n} \, \delta(t - nT) \qquad (2)$$

L'idée consiste alors à isoler X valeurs du premier signal et Y valeurs du second. Ces valeurs subiront alors une atténuation proportionnelle définie par :

$$s_i^{atténué} = s_i \frac{K - i}{K} \quad i = 1 \,..\, K \qquad (3)$$

Le résultat se présentera sous la forme :

$$s(t) = \sum_{1}^{N-X} s_{1n} \, \delta(t-nT) + \sum_{N-X+1}^{N} s_{1n} \frac{N-n}{N} \delta(t-nT) + \sum_{1}^{Y} s_{2n} \frac{Y-n}{Y} \delta(t-nT) + \sum_{Y+1}^{M-Y} s_{2n} \, \delta(t-nT) \qquad (4)$$

La fonction d'atténuation ainsi définie a été appliquée pour un nombre de points représentants 10 % de la durée du signal de l'unité acoustique. Les résultats obtenus sont montrés en ce qui suit :

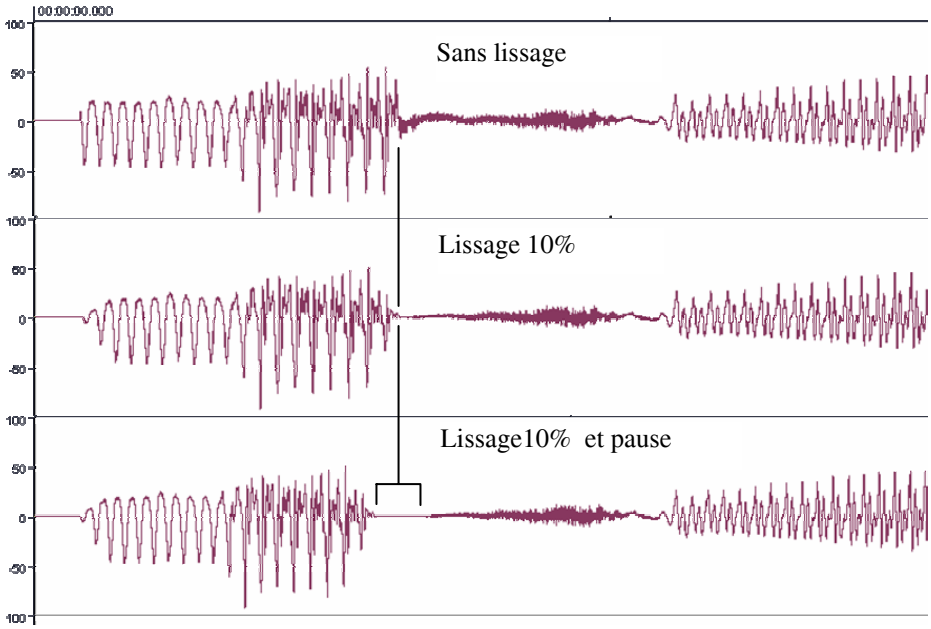

Fig. 3. Effet du lissage temporel sur la forme d'onde au niveau des points de discontinuités.

Les courbes précédentes montrent l'effet de ce lissage temporel sur un exemple de synthèse du mot « مَسَافَةٌ » (masaafatun : distance). En effet, la première courbe montre une concaténation bout à bout nous y constatons une discontinuité flagrante aux niveaux des points de jointures. La courbe du bas introduit, quant à elle, le résultat d'une concaténation lissée et la fluidité aux niveaux des points de concaténation. Le résultat obtenu a sensiblement amélioré la qualité de la voix synthétisée. Néanmoins, nous constatons un chevauchement entre les unités. Pour éviter un tel problème nous avons introduit un temps de silence de 10 millièmes de seconde. L'insertion d'une pause entre les unités avec nous a alors permis d'obtenir une meilleure intelligibilité.

5 RESULTATS DES TESTS

Afin d'évaluer notre système, nous avons établi une procédure de test basée sur l'écoute et l'identification de phrases synthétisées. Nous avons utilisé 20 phrases, soit 53 mots, 211 unités acoustiques dont 73 différentes ce qui constitue 37.2 % de la totalité des unités acoustiques qu'utilise notre système. Nous les avons fait écouter à 8 personnes (4 femmes et 4 hommes) ce qui a permis une évaluation statistique réaliste du résultat. Chaque phrase est écoutée trois fois, à chaque passage le sujet doit orthographier ce qu'il entend. En ce qui suit les résultats obtenu :

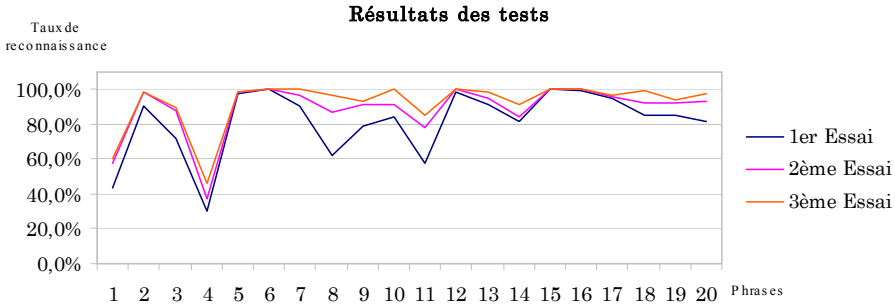

Fig. 4. Les résultats de la phase de test

Nous avons alors pu conclure à un pourcentage d'identification de plus de 81 % dès la première écoute, ce taux passe à plus de 92% pour la troisième phase. Par ailleurs nous avons remarqué qu'une phase d'adaptation de 2 à 3 phrases a été nécessaire pour avoir une stabilisation des taux de reconnaissance. De ces relevés nous avons aussi constaté que les mots non courants sont difficilement identifiables (exp : لذعْتُهُ phrase n° 4), et que quelques caractères sont plus difficiles que d'autres pour l'identification (exp :ذ phrase n° 3, 4 et 11).

6 CONCLUSION

Nous avons présenté dans cet article notre système de synthèse de la parole, ces différents constituants, les différentes phases de son élaboration et les choix techniques retenus pour chaque module. Le module de syllabation constitue à notre sens le point de départ pour une autre vision de la langue arabe, vue la rupture totale avec les méthodes jusque là utilisées en

langue arabe. Nous avons aussi exposé l'opération de concaténation ainsi que le poste traitement que nous avons choisi pour remédier aux problèmes de discontinuités.

La comparaison des résultats obtenus par rapport à l'existant demeure difficile. Les travaux sur les systèmes de synthèse de la parole arabe sont peu nombreux et les résultats d'évaluation ne font pas l'objet d'articles publiés. Néanmoins nous avons relevé que notre système a permis de se restreindre à trois types de syllabes seulement (CVV, CV et C) contrairement aux autres travaux préconisant cinq voir six types de syllabes différents (Ben Sassi, 2001). Nous n'utilisons que 196 unités acoustiques pour synthétiser n'importe quelle occurrence de l'arabe standard alors que le minimum jusque là était de 310 unités (Elshafei, 2002).

Références

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

2 Saidane Tahar, Zrigui Mounir, Pr Ben Ahmed Mohamed. 2004. La Transcription Orthographique-Phonétique de la Langue Arabe. RÉCITAL 2004, Fès, Maroc.

3 Emerard Frençoise. 1977. Les diphones et le traitement de la prosodie dans la synthèse de la parole. Bulletin de l'institut de phonétique de grenoble.

4 Dutoit Thierry. 1993. High quality text to speech synthesis of the french language. Thèse. Faculté polytechnique de Mons.

5 Elshafei M., Al-Muhtaseb, H., Al-Gamdi M. 2002. Techniques for high quality Arabic speech synthesis, Information sciences, Vol.140, 255-267.

6 Ben Sassi S., Braham R., Belgith A. 2001. Neural speech synthesis system for Arabic language using celp algorithm, Proc. Conference on Computer Systems and Applications.

7 Saidane Tahar, Haddad Ahmed, Zrigui Mounir, Pr Ben Ahmed Mohamed. 2004. Réalisation d'un système hybride de synthèse de la parole arabe utilisant un dictionnaire de polyphones. JEP-TALN 2004, Traitement Automatique de l'Arabe, Fès, Maroc.

8 Boula de Mareuil Philippe, Célérier Philippe, Cesses Thierry, Fabre Serge, Jobin Carine, Le Meur Pierre-Yves, Obadia David, Soulage Benoît, Toen Jacques. 2001. Elan text to speech : un système multilingue de synthèse de la parole à partir du texte. Elan TTS Toulouse.

9 Lemmety Sami. 2000. Review of speech synthesis technology. Thèse. Helsinki University of Technology.