SAPHA : Un système pour l'analyse acoustique et le décodage phonétique de l'Arabe standard moderne

Mahieddine DJOUDI

Laboratoire d'Informatique de l'Université de Poitiers
40, avenue du Recteur Pineau,
86022 Poitiers cedex

Résumé :

Nous présentons dans cet article le système SAPHA (Système Acoustico-PHonétique de l'Arabe) que nous avons développé pour l'analyse acoustique et le décodage phonétique de l'Arabe standard moderne. Le système reçoit en entrée le signal de parole préalablement digitalisé et renvoie en résultat un treillis phonétique.
Autour des modules de reconnaissance, nous avons développé des procédures pour le calcul et l'affichage graphique des paramètres acoustico-phonétiques du signal ainsi que des modules d'évaluation des performances du système de reconnaissance. L'évaluation nécessite un corpus de phrases étiquetées manuellement.

1 Introduction

Le système SAPHA permet de faire la reconnaissance analytique des phonèmes de l'Arabe standard en parole continue et dans un contexte multilocuteur. Ce système peut être considéré comme un étage d'un système de reconnaissance de l'Arabe qui intègre des informations linguistiques ou bien comme un module d'une machine à dicter vocale. Le système de décodage proprement dit est un système expert à base de règles de production. L'acquisition des connaissances a nécessité le développement d'outils d'analyse et d'affichage graphique. L'évaluation des performances du système a été faite en utilisant le corpus DJOUMA (DJOUmal MAqrou'a, qui veut dire en Arabe phrases lues) constituées de 50 phrases phonétiquement équilibrées prononcées par 11 locuteurs (7 hommes et 4 femmes) [1].

2 Architecture du système

Le système SAPHA est structuré en modules (voir figure **??**). Il reçoit en entrée le signal de parole d'une phrase. Le résultat est un treillis phonétique. Le système possède les fonctionalités classiques pour acquérir et restituer le signal de parole, le visualiser et calculer le spectrogramme (par l'algorithme FFT) ainsi que des fonctions pour le traitement du corpus.
Nous pouvons ainsi découpé le système en 4 grandes parties, chacune comportant un certain nombre de modules et dédiée à un travail particulier :

1. acquisition et traitement acoustique du signal.
2. reconnaissance automatique des phonèmes.
3. affichage graphique et analyse phonétique.
4. traitement du corpus et évaluation des performances.

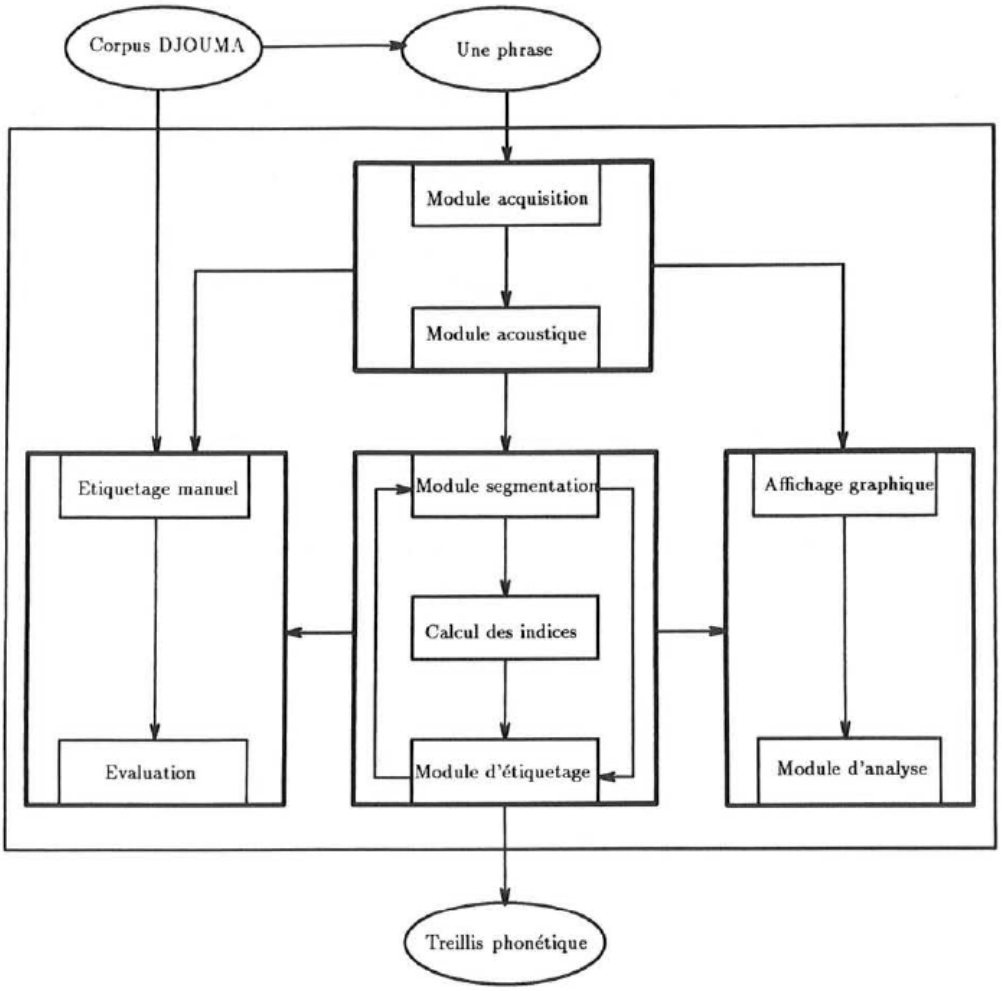

FIGURE 1 – Architecture de SAPHA

3 Acquisition et représentation paramétrique

3.1 Le module d'acquisition

Ce module est composé d'un ensemble de fonctions :

- Acquérir de la parole, le système demande le temps d'acquisition qui est limité à 4 secondes. Le signal est stocké sur fichier disque contenant les échantillons de taille 16 bits.
- Ecouter une partie ou l'ensemble du signal, le système demande le début et la fin de la zone à restituer dans le signal temporel ; l'utilisateur choisit le nombre de restitutions.
- Choisir la fréquence d'échantillonnage lors de l'acquisition et la restitution. Cette fréquence est utilisée par plusieurs modules et elle est fixée par défaut à 16kHz, mais il est possible de l'augmenter jusqu'à 33 kHz.
- Lire un fichier de parole dans l'un des corpus existants.
- Sauvegarder une partie ou l'ensemble du signal temporel. Il suffit de donner le nom du fichier

et ensuite le début et la fin du signal pour délimiter la partie de la phrase à sauvegarder.
- Afficher le signal temporel sur une fenêtre de la console graphique.
- Faire un zoom sur le signal temporel.

3.2 Le module acoustique

Ce module se charge d'extraire les paramètres acoustiques à partir du signal temporel. Il permet en particulier de :

- Calculer et afficher un spectrogramme en 13 niveaux de couleurs. Le spectrogramme normal est un spectrogramme à bande large avec une fenêtre de Hamming de 4 ms et un déplacement de 2 ms mais il est possible de calculer les spectrogrammes suivants :
  - Spectrogramme calculé sur les cœfficients LPC qui renforce les fréquences formantiques.
  - Spectrogramme à bande étroite pour séparer les harmoniques de la fréquence fondamentale.
  - Spectrogramme lissé cepstralement.

  Les algorithmes utilisent le processeur vectoriel, ce qui permet d'obtenir un temps de calcul d'environ 1 seconde pour un spectrogramme de 4 secondes de parole.
- Réafficher un spectrogramme déjà calculé en modifiant sa présentation (lissage et nombre de canaux). L'utilisateur possède plusieurs palettes de couleurs pour l'affichage (gris, rouge, bleu, vert, orange).
- Calculer l'amplitude du signal comme étant :
  $E_a = \frac{1}{N} \sum_{n=0}^{N-1} |x(n)|$
  Ce paramètre peut servir pour distinguer entre parole et non parole et entre sons voisés et sons non voisés, de même, il fournit une information sur l'intonation.
- Calculer le nombre de passages par zéro du signal temporel par seconde. Ce paramètre est utilisé pour distinguer entre parole et non parole et permet de différencier les sons voisés des sons non voisés.
- Calculer la fréquence fondamentale ou pitch, qui correspond aux vibrations des cordes vocales. La méthode utilisée est celle de l'autocorrélation [4].

4 Décodage phonétique

4.1 Le module de segmentation

Il consiste à segmenter le signal de parole en grandes classes phonétiques en utilisant des algorithmes non contextuels et reposant sur des critères simples. Nous avons retenu trois grandes classes : les voyelles, les plosives et les fricatives. Le reste des phonèmes de l'Arabe standard sont mis dans la classe des sonnantes.
Les fonctions prévues à ce niveau sont : le calcul, l'affichage et la sauvegarde de la segmentation.

4.2 Le calcul d'indices

L'extraction des indices phonétiques pertinents est une étape très importante dans le processus de décodage phonétique. Nous avons développé une procédure pour chacun des indices phonétiques suivants :

- la durée d'un segment,
- le degré de voisement,
- la barre d'explosion et ses paramètres,
- les valeurs des formants,
- les transitions formantiques,

— le centre de gravité énergétique,
— la limite inférieure du bruit de friction.

4.3 Le module d'étiquetage

C'est à ce niveau que se fait le décodage proprement dit. En partant des segments fournis par le module de segmentation, le module tente de trouver les bons phonèmes prononcés en utilisant les indices extraits lors de l'étape précédente. Deux méthodes d'étiquetage ont été utilisées, l'une procédurale et l'autre utilisant un système expert à règles de production.

5 Les outils d'analyse

5.1 L'affichage graphique

Afin de pouvoir faire une analyse détaillée sur les phonèmes, nous avons prévu des fonctions qui permettent d'afficher :

— Les deux formants d'un segment de parole le plan F1-F2. L'utilisateur choisit les limites avec la souris. Ce sont les deux premières racines du calcul de la LPC.
— Les pics des cepstres les plus visibles jusqu'à 5000 Hz. Les cepstres sont calculés sur une fenêtre de 256 échantillons avec un recouvrement de 128 échantillons.
— Les racines de LPC dont la largeur de bande est inférieur à 700 Hz.
— Les pics de cepstres qui ne correspondent pas à une racine de la LPC.
— La FFT, la FFT lissée, la LPC et le cepstre correspondant au signal que l'utilisateur clique sur le spectrogramme. Une marque rouge apparaît sur le spectrogramme et permet de savoir à quel instant correspond les courbes calculées.
— Le suivi des pics du cepstre sur le spectrogramme.
— La courbe d'énergie dans une bande de fréquences.

5.2 L'analyse phonétique

Il s'agit d'analyser phonétiquement une phrase à partir de sa représentation spectrographique. Cette analyse consiste à calculer les valeurs des paramètres phonétiques des segments. Le processus peut être activé de deux manières :

— L'une manuelle, qui consiste à préciser la nature (voyelle, plosive, fricative ou autre) et les limites (début et fin) du segment à analyser en cliquant sur le spectrogramme. Le résultat est ensuite affiché sur la console.
— L'autre automatique qui consiste à calculer les valeurs des indices de tous les segments. La nature et les limites de chaque segment sont obtenues soit par l'étiquetage manuel soit par le module de segmentation.

Les indices à extraire sont ceux utilisés lors du décodage phonétique et qui sont calculés par le module d'extraction des indices.

6 Traitement du corpus et évaluation

6.1 L'étiquetage manuel

L'étiquetage manuel permet d'affecter des étiquettes phonétiques à des segments de parole à partir de la représentation spectrographique de la phrase. L'étiquetage n'est possible que sur console graphique

et il permet de :

- Insérer une étiquette phonétique qui apparaît toujours avant la marque que l'on pose avec la souris. Le système demande ensuite l'étiquette à mettre grâce à des menus portant sur les phonèmes de l'Arabe standard répartis en grandes classes.
- Effacer une étiquette et la marque correspondante. Il suffit de cliquer le segment à détruire.
- Changer l'étiquette d'un segment. Pour cela, il suffit de cliquer sur l'étiquette à modifier et de choisir ensuite la bonne étiquette.
- Déplacer la limite d'un segment. L'utilisateur clique une première fois sur cette limite et une seconde fois pour indiquer ou' la déplacer.
- Calculer et afficher un spectrogramme haute définition.
- Ecouter un morceau de signal autant de fois que l'on veut dans le but de faciliter l'étiquetage.

Le résultat de l'étiquetage manuel est sauvegardé dans un fichier qu'il est ensuite possible de lire pour le consulter ou le modifier. Il servira en particulier à évaluer les performances du système tant au niveau segmentation qu'au niveau reconnaissance.

6.2 Evaluation des performances

Le module d'évaluation a pour tâche de calculer les performances du système par rapport à l'étiquetage manuel. Deux évaluations sont prévues, l'une concerne la segmentation et l'autre l'étiquetage [2].

Evaluation de la segmentation : elle consiste à rapprocher les résultats de la segmentation automatique avec la segmentation manuelle effectuée sur les phrases du corpus en utilisant un algorithme de programmation dynamique. Pour chaque phonème, le système rend le nombre d'occurrences dans chaque classe phonétique ainsi que les taux de bonne segmentation, d'insertion et d'omission.

Evaluation du décodage : elle consiste à calculer le taux de reconnaissance phonétique du système par comparaison du résultat du décodage avec la transcription correcte des phrases obtenue lors de l'étiquetage manuelle. Le résultat est une matrice de confusion. Pour chaque phonème est donné le nombre de fois ou' il a été confondu avec un autre. La mise en correspondance entre le résultat du système et la transcription de la phrase utilise une matrice initiale de confusion qui fixe au préalable les confusions possibles entre les phonèmes de la langue. Le remplissage de la matrice de confusion tient compte du mode d'articulation et de la proximité entre les lieux d'articulation des phonèmes.

7 Conclusion

Nous avons présenté dans cet article un système d'analyse acoustique et de décodage phonétique de l'Arabe standard moderne. Sa réalisation a nécessité de développement des algorithmes de traitement du signal de parole ainsi que des outils d'affichage de visualisation des courbes. Le module de reconnaissance utilise un système expert à base de règles de production. L'évaluation du système de décodage faite sur un corpus de 50 phrases prononcées par trois locuteurs masculins donne un taux global de reconnaissance de 65% [2]. Les perspectives à donner au présent travail est d'utiliser en plus de l'approche basée sur la connaissance, de nouvelles méthodes de décodage acoustico-phonétique (méthodes connexionistes, statistiques, etc.) afin d'améliorer le pourcentage de reconnaissance et d'intégrer le décodeur phonétique SAPHA dans un système de reconnaissance et/ou de compréhension du langage arabe parlé [3].

References

[1] M. Djoudi. Contribution à l'étude et à la reconnaissance automatique de la parole en arabe standard. Thèse de Doctorat de l'Université de Nancy 1, 1991.

[2] M. Djoudi. Assessment of Acoustic Phonetic Decoder for Standard Arabic. In *13th National Computer Conference*, volume 2, pages 761–770, Riyadh Saudi Arabia, 28 November - 2 December, 1992.

[3] M. Djoudi, D. Fohr, and J. P. Haton. MARS : Un système de reconnaissance de l'Arabe moderne. In *Actes des* $18^{ème}$ *Journées d'Etudes sur la Parole*, pages 217–221, Montréal, Mai, 1990.

[4] L. R. Rabiner, M. J. Cheng, A. E. Rosenberg, and C. A. McGonegal. A comparative study of several pitch detection algorithms. *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP-24 :399–418, Oct. 1976.