Calculatrice vocale basée sur les SVM

Zaïz Fouzi *, Djeffal Abdelhamid *, Babahenini MohamedChaouki*, Taleb Ahmed Abdelmalik,**

** Laboratoire LESIA, Département d'Informatique, Université Mohamed Kheider Biskra, Algérie*
*** IUT GE2I, Laboratoire LAMIH UMR CNRSUVHC 8530, Université de Valenciennes, France*

threezed@yahoo.com; Abdelhamid_Djeffal@yahoo.fr ; chaouki.babahenini@gmail.com; Abdelmalik.Taleb-Ahmed@univ-valenciennes.fr

RÉSUMÉ. Avec le développement très important des moyens de calcul et de stockage, les interfaces homme machines sont devenues de plus en plus proches de l'interaction humaine naturelle. Dans ce papier, nous présentons une calculatrice vocale basée sur l'apprentissage et la classification par la méthode de la machine à vecteurs supports. Les techniques d'apprentissage utilisées dans la littérature tel que les réseaux de neurones soufrent de la faiblesse du taux de classification et la vitesse d'apprentissage. La classification par SVM est une méthode très puissante qui a démontré de très bons résultats dans plusieurs domaines notamment dans la reconnaissance des visages et des caractères manuscrits. L'application de cette méthode dans la reconnaissance vocale appliquée à une calculatrice a permis un taux très intéressant de classification

MOTS-CLÉS : Reconnaissance vocale, Apprentissage, machine à vecteurs supports.

1. Introduction

Depuis longtemps, la reconnaissance vocale n'a cessé d'attirer l'attention des chercheurs et de consommer des budgets, vu le développement très important qu'elle puisse apporter aux systèmes de sécurité, aux interfaces homme-machine et aux systèmes d'aide des non voyants. De tels systèmes se composent généralement de plusieurs étapes. On commence par l'acquisition de la voix et sa numérisation, puis son codage. Les données vocales codées passent ensuite par une phase d'analyse en but d'extraction des caractéristiques essentielles de la voix. Le système de reconnaissance nécessite une étape d'apprentissage où l'utilisateur doit introduire un ensemble de données types qui vont être utilisés à travers un classifieur pour créer un modèle pour prendre des décisions à propos des voix en entrée lors de la phase d'utilisation.

Dans la première phase, l'isolation des mots des slots de silence est très importante. Les mots sont ensuite soumis à une méthode dite LPC (Linear predictif coding) qui

permette de les convertir en vecteurs contenants leurs caractéristiques statistiques les plus importantes. Les vecteurs obtenus représentent des points dans un espace à n dimensions ou n est les nombre d'attributs c'est-à-dire le nombre de composantes des vecteurs. La méthode SVM consiste à la recherche d'un hyperplan dans l'espace des vecteurs qui sépare le mieux possible ces vecteurs en deux classes, un l'hyperplan doit être trouvé pour chaque voix. Ainsi, chaque nouveau mot détecté va être exposé à chacune des classes (mots) possibles pour trouver la classe la plus adéquate. Une fois le mot (le chiffre ou l'opération) est détecté, il est passé à la calculatrice pour l'utiliser.

Dans ce papier, on va décrire les algorithmes utilisés dans chaque phase et enfin les résultats obtenus.

2. Schéma général du système

L'objectif de notre système est la réalisation d'une calculatrice vocale, pour ce faire, on utilise un ensemble de commandes vocales où chaque commande passe par une succession d'opérations : acquisition, segmentation et extraction des vecteurs acoustiques, apprentissage et classification, et finalement calcul et synthèse du résultat.

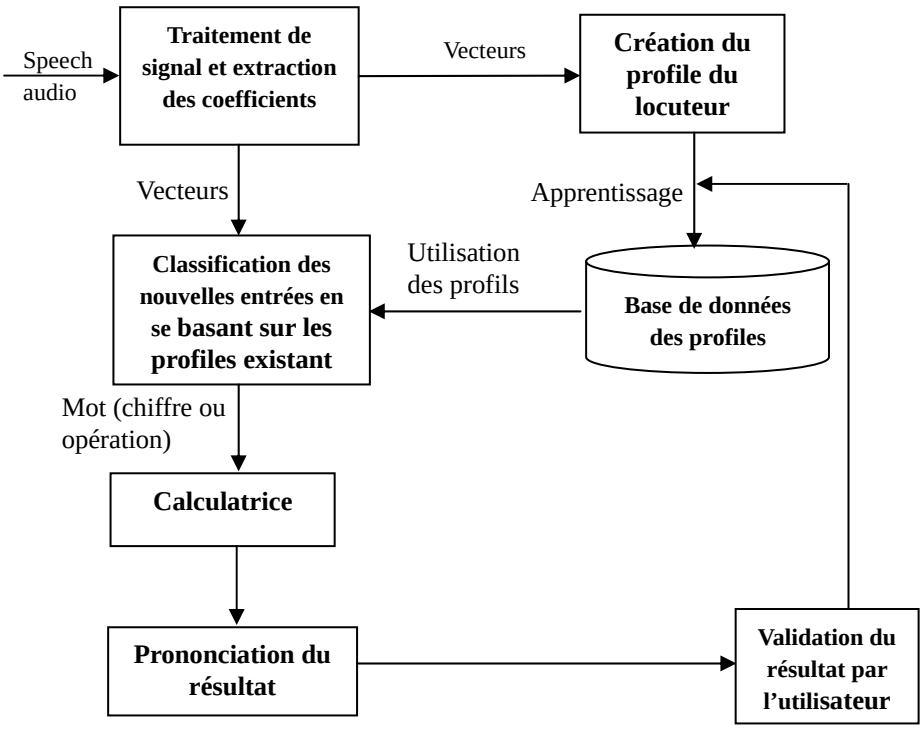

Figure 1 : *Schéma général du système*

2.1. Traitement du signal et extraction des cœfficients (Tebelskis 95 , Anibal 2004, Bellanger 2002)

Dans cette étape, le signal vocal passe par deux opérations: la segmentation parole/silence (mots isolés), puis l'extraction des caractéristiques de la voix (méthode LPC).

2.1.1. Segmentation parole/silence

Cette segmentation se base sur une analyse temporelle du signal. Une inspection minutieuse de la structure temporelle (forme d'onde), selon un certain nombre de critères, permet une segmentation primaire, fiable et précise.

Après avoir éliminé la composante continue, le signal est recodé selon les passages par zéro de sa dérivée. Nous retenons donc les instants ki, ki+1…, ainsi que les amplitudes a (i) associées aux extremums. Ce prétraitement est avantageux, car il permet de réduire la quantité d'informations à traiter et le signal se présente sous une forme plus simple à manipuler. Le senseur d'identification des segments Silence et Parole se base sur la comparaison des amplitudes du signal avec le niveau de bruit. Cependant le senseur ne fait pas intervenir un traditionnel seuil d'énergie ou d'amplitude moyenne, mais un codage particulier des extremums.

2.1.1.1. Codage des extremums du signal

Les extremums sont des amplitudes a(i) sur une portion de silence supérieures à la moitié du niveau de bruit. Le niveau de bruit Ln est calculé en prenant la moyenne des valeurs l'écart type :

$$Ln = |a(i)| + \sigma / 2 .$$

Les extremums du signal sont ensuite classés en tenant compte de leur voisinage immédiat. La classe AI ("Amplitude Inférieure") correspond au type Silence et la classe AS ("Amplitude Supérieure") au type Parole (figure 2).

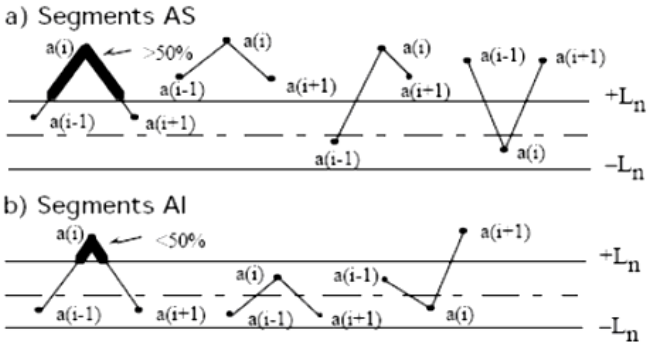

Figure 2 : *Codage des extremums selon le niveau de bruit (Ln).*

L'identification des éléments AI et AS fait ressortir des tendances de type silence (succession de segments de type AI de durée longue, entre coupés de segments AS de durée plus courte), et des tendances de type parole (présence de segments de type AS plus longs). Par une élimination judicieuse des éléments les plus courts, nous déterminons le début et la fin des segments de parole.

2.1.1.2. Détermination des segments Silence et Parole

Pour commencer, les éléments AS d'une durée plus courte qu'un quart de la période de voisement sont éliminés (figure 3).

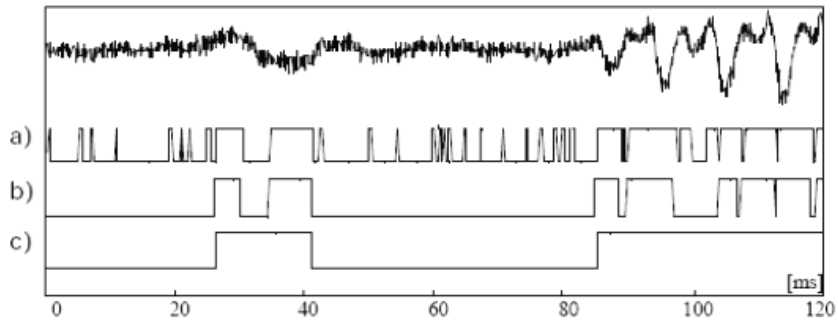

Figure 3 : *Détermination du début du segment de parole.*

Ce critère est basé sur l'observation de portions de signaux de faible amplitude (début ou barre de voisement). Ensuite, on élimine les segments de type AI dont la durée ne dépasse pas une période de voisement (figure 3.c). La procédure décrite est bien adaptée aux signaux voisés, mais peut présenter des erreurs dans le cas des fricatives ou des barres d'explosion des occlusives de faible amplitude. De ce fait un second procédé est utilisé en parallèle, basé sur le niveau *Ld* de la première dérivée du signal et qui a pour effet de mieux faire apparaître les composantes de haute fréquence :

$$Ld = \frac{1}{M} \sum_{K=N}^{N+M} \left| x(K+1) - x(K) \right|$$

Où *M* est la durée de la fenêtre d'analyse, *N* le point de départ et *x(K)* les échantillons du signal.

Les fricatives étant caractérisées par une distribution d'énergie plus importante dans les hautes fréquences, *Ld* est plus élevé (figure 4). Un seuil sur *Ld* permet d'assurer l'inclusion des frictions de faible amplitude dans les segments de type Parole.

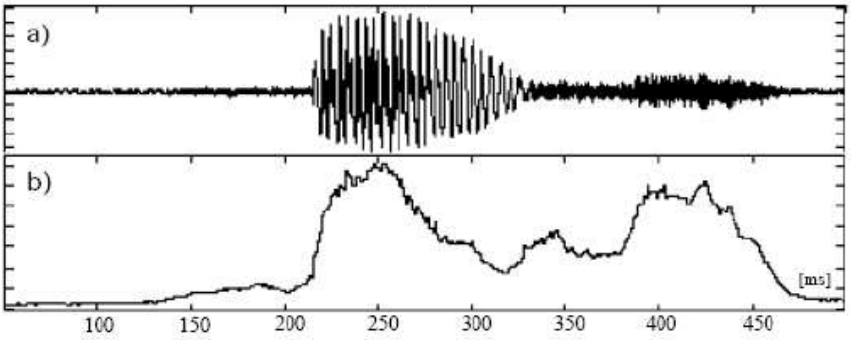

Figure 4. *Enveloppe de la première dérivée du signal.*

2.1.2. Extraction des caractéristiques de la voix

Il existe une diversité de méthodes pour extraire les caractéristiques d'un signal vocal, mais celle dont la fiabilité a été prouvée est bien le codage prédictif linéaire ou LPC (Linear Predictif Coding) car elle extrait l'information d'une petite partie de l'enveloppe spectrale de la parole.

Les coefficients de LPC représentent les caractéristiques les plus importantes. Les études montrent l'efficacité des coefficients de LPC dans la reconnaissance et l'identification.

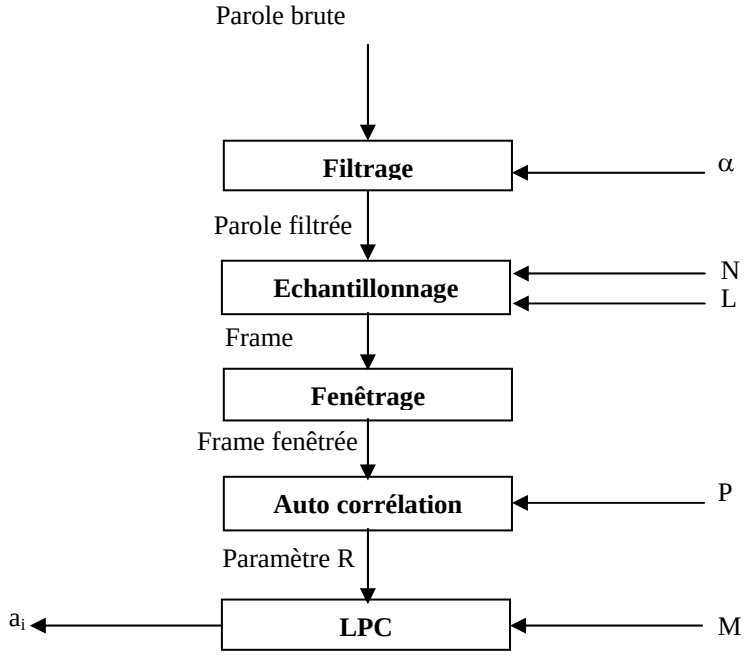

Figure 4 : *L'extraction des paramètres vocaux par LPC*

3. Support vector machines (SVM) et son (Cristianini & all 2000, Abe 2005, Wang 2005)

Après l'extraction des paramètres du signal vocal par la méthode LPC, ces paramètres sont utilisés comme une donné d'entrée pour le composant de classification (SVM), qui va rechercher un hyperplan séparateur qui sépare les exemples dans la phase d'apprentissage et prend une décision de classification dans la phase d'identification.

Dans le module SVM, il y a deux phases : une pour l'apprentissage et l'autre pour la classification. La figure 7 représente la relation entre ces deux phases.

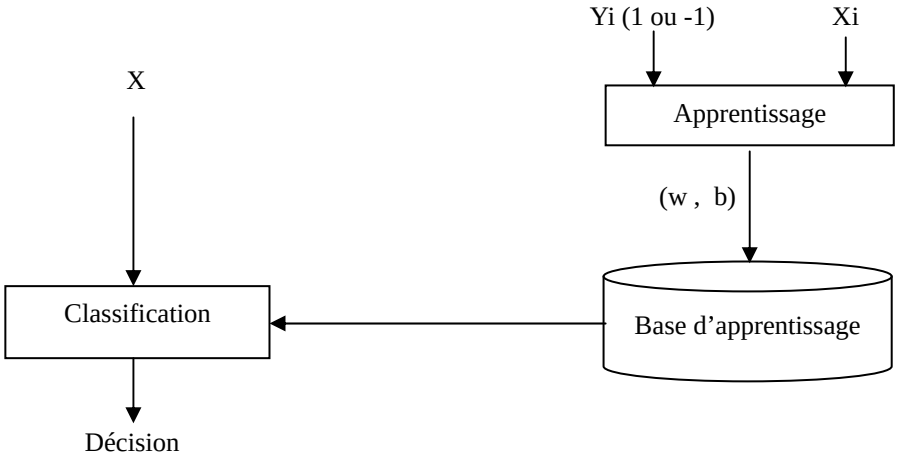

Figure 7 : *Apprentissage par SVM*

Où X, Xi représentent le vecteur caractéristique des enregistrements du son ; et Yi les étiquettes de chaque classe.

4. Résultats

La méthode proposée, commence par demander à l'utilisateur d'enregistrer des échantillons des sons utilisés (zéro, un, deux, …, onze, …. Trente, … cent, mille, … fois, plus, … virgule, …), ensuite effectue une analyse LPC de ces sons pour extraire leurs coefficients. Une fois les coefficients définis, on passe à la phase d'apprentissage où on détermine un w et un b pour chaque ensemble de sons du même symbole, en le

supposant appartenir à une classe et les autres sons à une autre. Dans la base d'apprentissage, on n'enregistre que les paramètres w et b pour chaque symbole.

Figure 8. *Interface de la calculatrice et d'apprentissage*

Dans la phase d'utilisation, on détecte le mot prononcé, on extrait les coefficients LPC, puis on applique la fonction de décision de chaque son dans la base jusqu'à trouver la classe qui convient.

Une fois le symbole identifié, il est utilisé par la calculatrice comme une opération ou un opérande selon son type. Lorsque l'utilisateur tape le caractère " = ", la calculatrice prononce le résultat en utilisant des sons déjà préparés et stockés dans une base de donnée.

Les résultats de cette calculatrice, dépendent des paramètres choisis pour la méthode SVM et du nombre d'exemple d'apprentissage utilisés et leur qualité.

Pour choisir le noyaux nous nous somme basés sur les résultats de plusieurs références (Ganapathiraju 2004, Wang 2005) qui montent que le noyau le plus adapté à l'apprentissage de la voix par SVM est le noyau gaussien avec les paramètres Gamma = 0.5 et C=10.

Les taux d'apprentissage atteints en utilisant le noyau gaussien, sur 150 exemples pour chacun des symboles 0, 1, 5, 7, +, *, = sont données dans le tableau suivant :

Classe	Nombre d'exemples	Taux de reconnaissance (%)	Taux d'erreurs (%)
0	150	76.66	23.34
1	150	90.33	9.67
5	150	95	5
7	150	100	0

+	150	89	11
*	150	98	2
=	150	92	8
Total	1200	91.57	8.43

Tableau 1 : *Résultats de reconnaissance de quelques chiffres*

En comparant ces résultats par ceux obtenus dans (Ganapathiraju & all 2004) pour les mêmes valeurs de Gamma et C où le taux de reconnaissance était 63%, le taux d'apprentissage obtenu ici est très intéressant. Cependant le temps de calcul était relativement important.

5. Conclusion

Ce papier résume un travail d'application de la méthode SVM dans le domaine de la reconnaissance vocale en but de réaliser une calculatrice vocale. Les résultats obtenus sont encourageants et démontre la puissance de cette technique dans la reconnaissance vocale. Le travail futur sera concentré sur l'amélioration des prétraitements tel que l'utilisation de la méthode MFCC (Mel Frequency Cepstrum Coefficients) au lieu le la méthode LPC, puis l'étude des différents noyaux et leut influence sur les résultats, et l'optimisation du temps pour l'algorithme SMO.

6. Références

Abe S., « *Support Vector Machines for Pattern Classification* », Springer 2005.

Anibal J., « Méthodes à vecteurs de support et indexation sonore », *IRIT, SAMoVA*, 2004.

Bellanger M., « *Traitement numérique du signal théorie et pratique* », Edition Dunod, 2002.

Cristianini N., Taylor J., « *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods* », Cambridge University Press, 2000.

Ganapathiraju A., Hamaker J., Picone J., « Application of support vector machines to speech recognition », *IEEE Transaction on signal processing*, 18 mars 2004.

Scaringella N., Mlynek D., « A mixture of support vector machines for audio classification », IEEE MIREX, London, 2005

Tebelskis J., « Speech Recognition using Neural Networks », Thèse PHD, School of Computer Science Carnegie Mellon University Pittsburgh, Pennsylvania, 1995.

Wang L., « *Support Vector Machines: Theory and Applications* », Springer 2005