Optimisation de la paramétrisation pour la reconnaissance floue des signaux de parole

Optimization of features parameters for fuzzy speech recognition

I. Ben Fredj K. Ouni
Unité de Recherche Systèmes Mécatroniques et Signaux
Ecole Supérieure de Technologie et d'Informatique (ESTI), Université de Carthage
Tunis, Tunisie
ines_benfredj@yahoo.fr
kais.ouni@esti.rnu.tn

Résumé :
Nous présentons une méthode de reconnaissance phonémique de la base Timit en utilisant la logique floue. Elle consiste en l'extraction de vecteurs de référence flous : le vecteur maximal, le vecteur moyen et le vecteur minimal. Pour classer un phonème donné, nous calculons son degré d'appartenance à toutes les classes des phonèmes, et nous choisissons par la suite le plus haut degré d'appartenance. Nous évaluons cette méthode avec différentes techniques de paramétrisation du signal de parole telles que les coefficients Mel Ceptre (MFCC), les coefficients de prédiction linéaire cepstraux (LPCC) et la méthode de prédiction linéaire perceptuelle (PLP). Nous avons introduit les dérivées premières et secondes ainsi que l'énergie de signal afin de modéliser ses caractéristiques transitoires pouvant contribuer à l'amélioration de la tâche de reconnaissance. Les taux de reconnaissance retenus sont 83.47%, 82.09% et 99.48% respectivement pour 36 MFCC, 11 LPC and 11 PLP.

Mots-clés :
Logique floue, LPC, MFCC, PLP, Timit.

Abstract :
We provide a phoneme recognition technique of Timit corpus based on the concept of fuzzy logic. The fuzzy method consists on the extraction of a three fuzzy reference vectors : maximal, mean and minimal vectors. To classify a phoneme request, we calculate his degree of membership to all classes. The class of a phoneme request is then the one which maximizes one degree of membership calculated according to reference vectors. We also compare the performance of our recognizer using different speech parameterization techniques such as MFCC, LPC and PLP. We extracted temporal dynamic of the signal to improve recognition task. So, we introduced first and second derivatives and energy. Our experimental results indicate that the classifier behaves differently depending on the number of samples. Fuzzy logic gave a considerable matter since retained recognition rates were 83.47%, 82.09% and 99.48% respectively for 36 MFCC, 11 LPC and 11 PLP.

Keywords :
Fuzzy logic, LPC, MFCC, PLP, Timit.

1 Introduction

La reconnaissance automatique de la parole comporte deux processus [1] [2] : le premier est la paramétrisation de signal qui consiste à représenter le signal de parole par un jeu de paramètres réduit, pertinent et robuste pour réduire le flux d'informations à traiter par le moteur de reconnaissance. La paramétrisation est généralement suivie par le processus de reconnaissance [3]. Un tel processus nécessite l'utilisation extensive des techniques d'intelligence artificielle. Ces techniques, telles que la logique floue, visent à résoudre les problèmes de la modélisation de données et de classification.
Dans ce travail, nous présentons une nouvelle technique de reconnaissance phonémique de la base Timit basée sur la logique floue [4]. Elle consiste en l'extraction de vecteurs de référence flous utilisés ensuite pour calculer le degré d'appartenance d'un phonème donné à toutes les classes définies des phonèmes. Le classifieur a été évalué avec différentes techniques de paramétrisation de signal tel que MFCC [5], LPC [6] et PLP [7]. Le but principal est de déterminer le nombre optimal des paramètres de signal générant les taux de reconnaissance les plus significatifs.

Dans la section suivante, nous proposons un aperçu général des techniques de paramétrisation de signal utilisées. Par la suite, nous définissons le concept de la logique floue qui sera suivi par l'approche de reconnaissance. Ensuite, nous discutons les résultats expérimentaux et nous finissons avec les conclusions et quelques perspectives.

2 Techniques de paramétrisation

2.1 MFCC

Le codage MFCC (Mel Frequency Cepstral Coding) est une technique très utilisée en traitement de la parole.
Il est basé sur la variation des bandes critiques de l'oreille humaine avec la fréquence, les filtres espacés linéairement aux basses fréquences et logarithmiquement à hautes fréquences [5]. Ces filtres sont modélisés par une échelle non-linéaire issue de connaissances sur la perception humaine : l'échelle Mel.
Pour les MFCCs, on utilise la fenêtre de Hamming durant la transformation du domaine temporel au domaine fréquentiel. Cette transformation est faite en utilisant la transformée de Fourier.
Un filtrage, est appliqué ensuite, par banc de filtres triangulaires espacés selon l'échelle de Mel. Cette échelle reproduit la sélectivité de l'oreille qui diminue avec l'accroissement de la fréquence.
Après le calcul de log, une transformée en cosinus discrète est appliquée pour assurer un retour au domaine temporel.
Figure 1 illustre l'algorithme de calcul des coefficients MFCC.

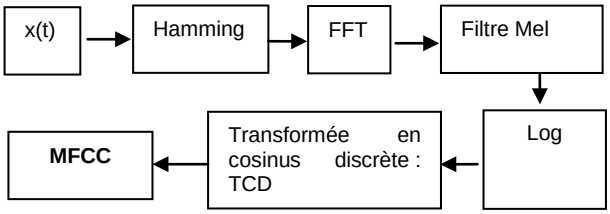

Figure 1 – Algorithme de calcul des coefficients MFCC

2.2 LPCC

Les coefficients LPC cepstraux peuvent être calculés à partir de l'analyse LPC du signal [6] par une procédure récursive. Autrement dit, ce sont les coefficients LPC convertis en cepstres.

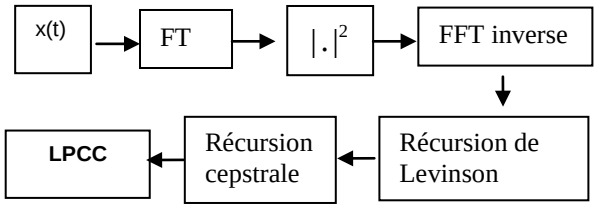

Figure 2 – Algorithme de calcul des coefficients LPC

L'analyse LPC est illustrée dans [6].

2.3 PLP

La méthode de prédiction linéaire perceptive (PLP), étudiée par Hermansky en 1991 [7], essaye de simuler le système auditif humain en introduisant des mécanismes psycho acoustiques de l'oreille humaine. Elle repose sur l'analyse LPC;
En effet, la PLP modélise un spectre auditif par un modèle tout pôle d'ordre réduit en utilisant la technique d'auto-corrélation de la prédiction linéaire.
La PLP consiste en un filtrage en bandes critiques du spectre de signal à court terme suivi d'une correction de l'intensité. L'amplitude du signal est alors compressée et enfin l'analyse par prédiction linéaire intervient. Cette dernière étape est en réalité une technique de compression spectrale qui modifie le spectre (ensemble de fréquences constituant un signal) de puissance à court terme avant son approximation par un modèle autorégressif.
L'algorithme de PLP est détaillé ci-contre :

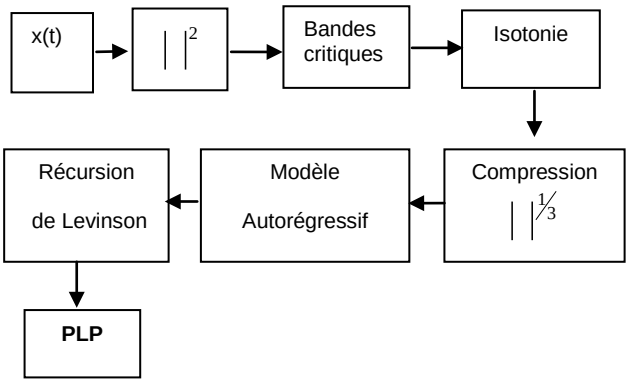

Figure 3 – Algorithme de calcul des coefficients PLP

3 HMM : méthode classique de reconnaissance de la parole

Les systèmes de reconnaissance de la parole sont généralement basés les modèles de Markov cachés : HMM.

Les HMM sont des modèles statistiques riches en structure mathématiques dont la sortie est une séquence de symboles ou de quantités [16]. Ils sont largement utilisés dans la reconnaissance de la parole étant donné leur puissante adaptation à la variabilité des observations. De plus, leur apprentissage automatique des données vocales a amélioré leurs performances de reconnaissance. D'autres parts, la formulation probabiliste offerte par les HMM fournit un cadre unifié pour les hypothèses de notation et les combinaisons des sources de connaissances.

Cependant, les HMM présentent deux limitations majeures: Tout d'abord, l'hypothèse de Markov d'indépendance conditionnelle (étant donné qu'un état ne dépend que de l'état précédent). Deuxièmement, les HMM sont bien définis uniquement pour les processus qui sont une fonction d'une seule variable indépendante, comme le temps ou la position unidimensionnelle.

Concernant la première limitation, théoriquement il est possible de définir un HMM dans lequel la dépendance peut s'étendre aux états et aux sorties précédentes. Néanmoins, de telles extensions compliquent les modèles définis et peuvent impliquer rapidement à une énorme complexité de calcul. Au même temps, beaucoup de travail reste à faire sur l'amélioration et l'utilisation des modèles déjà existants.

En outre, La deuxième limitation est fondamentale. Il semble impossible de définir un HMM dépendant de plus qu'une seule variable indépendante. Par conséquent, pour des problèmes tels que la reconnaissance vocale où le signal est naturellement est une fonction de deux variables, les chercheurs ont dû élaborer des méthodes de conversion des problèmes de deux dimensions en un problème d'une seule dimension avant d'utiliser les HMM comme solution.

Finalement et en dépit de cette contrainte, les HMM ont enrichi l'état de l'art dans le domaine de la reconnaissance vocale et se sont répandus progressivement à d'autres domaines comme la reconnaissance des formes électro-encéphalographique EEG, la reconnaissance des caractères, le domaine musical et d'autres [17].

Dans nos travaux de recherche, pour remédier aux problèmes présentés par les HMM, nous avons essayé d'implémenter une nouvelle approche de reconnaissance phonémique fondée sur la logique floue. L'algorithme flou utilisé a été testé dans la classification des objets 3D, et a donné de résultats remarquables [11].

4 Approche de reconnaissance proposée

4.1 Base de données

La base TIMIT [10] est utilisée pour la classification et la reconnaissance. Elle est composée par 630 locuteurs de 8 dialectes différents des États-Unis. Chaque locuteur prononçant 10 phrases ce qui donne 6300 phrases.

Le tableau 1 décrit la structure du corpus Timit.

Tableau 1 –Corpus Timit

Dialecte	Désignation	Locuteurs	
		H	F
DR1	New England	31	18
DR2	Northern	71	31
DR3	North Midland	79	23
DR4	South Midland	69	31
DR5	Southern	62	36
DR6	New York City	30	16
DR7	Western	74	26
DR8	Army Brat	22	11

H : Hommes
F : Femmes

Nous avons organisé le corpus Timit en sept classes homogènes de phonèmes représentant les affriquées, les fricatives, les nasales, les semi-voyelles, les occlusives, les voyelles et les autres comme l'illustre tableau 2.

Tableau 2 –Distribution des classes de phonèmes de Timit

Classe (ou macro-classes)	Etiquette
Affriquées	/jh/ /ch/
Fricatives	/s/ /sh/ /z/ /zh/ /f/ /th/ /v/ /dh/
Nasales	/m/ /n/ /ng/ /em/ /en/ /eng/ /nx/
Semi-voyelles	/l/ /r/ /w/ /y/ /hh/ /hv/ /el/
Occlusives	/b/ /d/ /g/ /p/ /t/ /k/ /dx/ /q/ /bcl/ /dcl/ /gcl/ /pcl/ /tcl/ /kcl/
Voyelles	/iy/ /ih/ /eh/ /ey/ /ae/ /aa/ /aw/ /ay/ /ah/ /ao/ /oy/ /ow/ /uh/ /uw/ /ux/ /er/ /ax/ /ix/ /axr/ /ax-h/
Autres	/pau/ /epi/ /h#/ /1/ /2/

De plus, la base Timit est divisée en données d'apprentissage et de reconnaissance. Le nombre des échantillons de ces données sera présenté par la suite avec les résultats.

Nous appliquons MFCC, LPC et PLP. Chaque phonème est représenté par un vecteur de 12 coefficients qui caractérisent les trois fenêtres du milieu. La fenêtre d'analyse est de 256 échantillons extraits chaque 10 ms en utilisant le fenêtrage de Hamming de 25 ms. La fréquence d'échantillonnage est égale à 16000 KHz.

4.2 Algorithme flou

L'algorithme de reconnaissance floue adopté repose sur l'extraction des vecteurs : minimal, moyen et maximal relatifs à chaque classe de phonèmes [11].
Le même algorithme est utilisé pour la classification et la reconnaissance.
Après l'extraction des paramètres MFCC, LPC et PLP, nous obtenons pour chaque phonème une matrice de coefficients. Nous répartissons tous les phonèmes selon la distribution décrite dans tableau 2 et nous commençons par la classification comme suit :

Soit v_{max}, v_{moy} et v_{min} le vecteur maximal, le vecteur minimal et le vecteur moyen d'une classe «c» ;

Le degré d'appartenance $D_{r,c}$ d'un vecteur « V_r » à la classe «c» est donné par :

$$D_{r,c} = \frac{v_{max} - v_{moy}}{V_r - v_{moy}} \quad \text{si} \quad v_{moy} \le V_r \le v_{max}$$

$$D_{r,c} = \frac{V_r - v_{min}}{v_{moy} - V_r} \quad \text{si} \quad v_{min} \le V_r \le v_{moy} \qquad (1)$$

$$D_{r,c} = 0 \quad \text{sinon}$$

En effet, la comparaison entre le vecteur « V_r » et les vecteurs de références v_{max}, v_{moy} et v_{min} est une comparaison des normes de chacun de ces vecteurs. Donc, l'expression « $v_{moy} \le V_r \le v_{max}$ » désigne « $\text{norme}(v_{moy}) \le \text{norme}(V_r) \le \text{norme}(v_{max})$ ».

Ainsi, nous obtenons pour chaque échantillon un degré d'appartenance relatif à chaque classe. Nous choisissons la classe relative au plus haut degré d'appartenance.

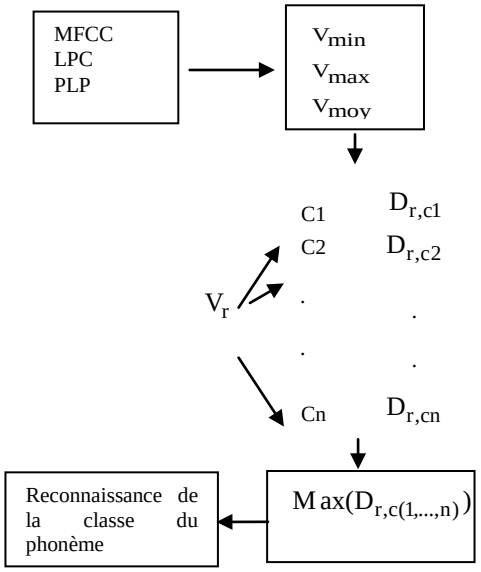

Figure 4 – Algorithme de classification floue

Notons que la classification d'un phonème en entrée signifie son identification par rapport aux macros classes ; Par exemple, pour classer le phonème /aa/, il suffit d'affirmer que c'est une voyelle. Ceci est établi si son degré d'appartenance à la classe des voyelles et supérieur aux degrés d'appartenance des autres classes. Ce phonème sera attribué à la classe du plus haut degré d'appartenance. Le même raisonnement est suivi dans l'approche de reconnaissance.

Pour développer les différentes techniques de paramétrisation utilisées (MFCC, LPC et PLP), nous avons utilisé l'implémentation de Dan Elis décrite dans [12].

De plus, la simulation du classieur a été effectuée sous l'environnement Matlab.

5 Résultats expérimentaux

Dans nos expériences, l'objectif principal était de déterminer le nombre optimal des coefficients qui offrent les meilleurs taux de reconnaissance.

Pour cette raison, un certain nombre d'expériences ont été effectuées dans lesquelles les coefficients dynamiques (dérivées premières et secondes) ainsi que l'énergie du signal ont été introduits avec les coefficients statiques des MFCC, LPC et PLP.

Les résultats préliminaires ont montré que les dérivées premières et secondes n'ont été avantageuses que pour MFCC. Bien que cette amélioration a confirmé que l'augmentation du nombre de caractéristiques améliore significativement les performances du système de reconnaissance (en notant que les meilleurs taux ont été obtenus avec 36 coefficients) ; Cette conclusion a été confirmée strictement pour MFCC. Comme ce n'était pas le cas pour les LPC et PLP, nous avons cherché à déterminer le nombre optimal de coefficients pour ces paramètres pour obtenir de meilleurs résultats. Certains récents travaux confirment qu'un nombre réduit de LPC et PLP peut entraîner meilleure représentation de signal de parole [13] [15]. De ce fait, plusieurs expériences ont été établies dans ce sens. Nous avons varié le nombre des coefficients à partir de 5 coefficients cepstraux + énergie jusqu'à 13 coefficients cepstraux + énergie. Les résultats retenus ont été obtenus par 11 coefficients cepstraux + énergie comme l'illustrent les tableaux 3 et 4.

Tableau 3 –Taux de classification

Dialecte	Nombre des échantillons	MFCC (%)	LPC (%)	PLP (%)
DR1	40084	96.87	86.62	99.14
DR2	81343	97.01	77.02	99.38
DR3	80686	97.12	84.05	99.45
DR4	73092	96.16	83.23	99.60
DR5	75978	96.85	83.94	99.56
DR6	37263	98.93	82.30	99.19
DR7	82419	98.81	80.73	99.38
DR8	23105	97.08	86.48	99.68
Taux moyen	493970	**97.32**	**82.39**	**99.43**

Tableau 4 –Taux de reconnaissance

Dialecte	Nombre des échantillons	MFCC (%)	LPC (%)	PLP (%)
DR1	11598	73.09	84.37	99.57
DR2	27884	86.66	71.43	99.44
DR3	27519	85.61	86.11	99.30
DR4	34119	81.90	95.88	99.42
DR5	29719	82.96	74.91	99.64
DR6	11680	89.73	72.32	99.66
DR7	25254	81.53	82.59	99.52
DR8	11766	84.93	82.49	99.39
Taux moyen	179539	**83.47**	**82.09**	**99.48**

Nous remarquons que les taux d'erreur les plus faible pour la classification et la reconnaissance ont été obtenus pour 11 PLP suivis par les 36 MFCC et ensuite par les 11 LPC. Une observation intéressante est que les taux de classification et de reconnaissance sont comparables ce qui indique une souplesse certaine du système de reconnaissance.
En outre, les meilleurs résultats ont été obtenus pour PLP et MFCC indépendamment du nombre de coefficients. De plus, nous pouvons déduire le bénéfice sûr et évident de l'intégration du coefficient d'énergie.
Nous pouvons noté également que les résultats varient d'un dialecte à un autre ce qui ouvre un important sujet de recherche pour étudier la cause de cette différence et d'évaluer chaque dialecte séparément.

6 Conclusions

Un nouveau système de reconnaissance de phonèmes basé sur la logique floue a été examiné dans ce travail. Des techniques classiques de paramétrisation de signal de parole ont été utilisées telles que MFCC, LPC et PLP.
Une étude comparative a été opérée pour déterminer le nombre adéquat des paramètres pour chaque technique utilisée.
Les résultats illustrent que les coefficients MFCC et PLP ont été les plus adéquats pour la classification et la reconnaissance floue présentées.
Cette étude préliminaire montre que notre méthode est une approche prometteuse pour la construction d'un moteur de reconnaissance phonémique. D'autres travaux se concentreront sur l'évaluation de cette reconnaissance floue dans un environnement bruité comme la base Aurora.

Références

[1] M.A. Anusuya, S. Katti. Front end analysis of speech recognition : a review. *International Journal of Speech Technology*, 99–145, 2011.

[2] B.H. Juang, L.R. Rabiner. Automatic speech recognition - A brief history of the technology development. *Elsevier Encyclopedia of Language and Linguistics*, 2005.

[3] I. Mporas, T. Ganchev, Mihalis, M. Siafarikas, N. Fakotakis. Comparison of Speech Features on the Speech Recognition Task. *Journal of Computer Science*, 608-616, 2007.

[4] I.B. Fredj, K. Ouni. Study of speech analysis techniques for the phonemes classification using fuzzy logic. *8th International Multi-Conference on Systems, Signals & Devices (SSD11)*, 1-5, 2011.

[5] B.T. Meyer, B. Kollmeier. Complementarity of MFCC, PLP and Gabor features in the presence of speech-intrinsic variabilities. *Interspeech*, 2009.

[6] Thiang. S.Wijoyo. Speech Recognition Using Linear Predictive Coding and Artificial Neural Network for Controlling Movement of Mobile Robot. *International Conference on Information and Electronics Engineering*, 179-183, 2011.

[7] H. Hermansky. Perceptual linear predictive (PLP) analysis of speech. *Journal of the Acoustical Society of America*, 1738-1752 (1990).

[8] L. A. Zadeh. Fuzzy logic, neural networks, and soft computing. *ACM'94*, 77-84, 1994.

[9] A.I. Pérez-Neira, M.A. Lagunas, A. Morell, J. Bas. Neuro-fuzzy Logic in Signal Processing for Communications : From Bits to Protocols. *NOLISP*, 10-36, 2005.

[10] Linguistic Data Consortium, University of Pennsylvania, http://www.ldc.upenn.edu/Catalog/readme_files/timit.readme.html

[11] A. Sadiq, R.O.H. Thami, M. Daoudi, J.P. Vandeborre. Classification des Objets 3D Basée sur la Logique Floue. *Compression et Représentation des Signaux Audiovisuels (CORESA'2004)*, 2004.

[12] D.P.W. Ellis. PLP and RASTA (and MFCC, and inversion) in Matlab using melfcc.m and invmelfcc.m, http://www.ee.columbia.edu/ ~dpwe/ resources/ matlab/ rastamat/

[13] M.A. Anusuya, S.K. Katti. Comparison of Different Speech Feature Extraction Techniques with and without Wavelet Transform to Kannada Speech Recognition. *International Journal of Computer Applications*, 19-24, 2011.

[14] L. Zadeh. Fuzzy sets. *Information and Control*. 1965.

[15] Prasad, K. Ravi, A.N. Mishra, S.N. Sharan. Text Dependant Speaker Identification Using VQ and DTW. *VSRD-IJEECE*, 453-459, 2011.

[16] I.B. Fredj, K. Ouni. Optimization of Features Parameters for HMM Phoneme Recognition of TIMIT Corpus. *Engineering and Technology*, 90-94, 2013.

[17] B. Askazad. Hidden Markov Model. University of Engineering and Technology Lahore Pakistan, 2001.