Combinaison de codeurs par algorithme génétique : Application à la vérification du locuteur

C. CHARBUILLET, B. GAS, M. CHETOUANI, J.L. ZARADER

Université Pierre et Marie Curie-Paris6, FRE2507 Institut des Systèmes Intelligents et Robotique (ISIR), Ivry sur Seine, F-94200 France

Christophe.Charbuillet@lis.jussieu.fr, Gas@ccr.jussieu.fr, Mohamed.chetouani@lis.jussieu.fr, Zarader@ccr.jussieu.fr

Résumé – Le domaine de la vérification du locuteur regroupe les applications pour lesquelles on désire identifier l'identité d'une personne à partir de sa voix. Le champ d'application couvre de nombreux secteurs tels que l'accès sécurisé, les transactions téléphoniques, la surveillance, l'indexation audio ou encore l'expertise judiciaire. Notre étude porte sur l'étape d'extraction de caractéristiques du système de reconnaissance du locuteur. Ce module a pour fonction d'extraire du signal de parole les informations pertinentes du point de vue de la discrimination inter-locuteur. Nous proposons dans cet article d'utiliser un algorithme génétique pour optimiser un système d'extraction de caractéristiques adapté à la reconnaissance du locuteur. La méthode proposée permet d'obtenir une amélioration significative du taux de reconnaissance sur la base Nist SRE 2005.

Abstract *– Speaker vérification aim at recognize the identity of a speaker based on his voice. Applications include security access control, banking phone transaction, surveillance, audio-indexing and forensic speaker recognition. Our study deals with speech feature extraction witch aim at extract speaker discriminative information from the voice signal. In this paper, we propose to use a genetic algorithm to design a feature extraction system adapted to the speaker recognition task. Result show that the proposed method improves significantly the system's performances on the 2005 Nist SRE database.*

1. Introduction

L'étape d'extraction de caractéristiques occupe une place fondamentale dans un système de vérification du locuteur. Les méthodes d'analyse du signal de la parole couramment utilisées aujourd'hui se divisent en deux groupes: les méthodes basées sur une modélisation de la production de la parole (LPC, LPCC) ainsi que les méthodes modélisant le système auditif humain (PLP, LFCC, MFCC). Ces méthodes générales sont utilisées aussi bien en reconnaissance de la parole, de la langue et du locuteur. Depuis quelques années, un certain nombre de travaux ont porté sur des méthodes permettant d'optimiser les systèmes d'extraction de caractéristiques à une tâche spécifique. Ces méthodes consistent à apprendre simultanément les paramètres du codeur et du classifieur [1]. La technique employée repose sur l'optimisation d'un critère qui peut être la maximisation de l'information mutuelle (MMI) [2] ou encore la minimisation du taux d'erreur de classification (MCE) [3].

Nous proposons dans cet article d'utiliser un algorithme génétique pour optimiser un système d'extraction de caractéristiques pour la tâche de vérification du locuteur.

Les algorithmes génétiques (AG) ont été proposés par Holland en 1975 et sont aujourd'hui couramment utilisés dans divers domaines pour l'optimisation de systèmes complexes. L'application de cette famille d'algorithmes au domaine du traitement automatique de la parole a connu ces dernières années un succès grandissant. On pourra citer les travaux de Chin-Teng Lin & al. [4] portants sur l'application des AG au problème de la transformation de caractéristiques pour la reconnaissance de la parole, ainsi que ceux de M. Zamalloa & al. [5] qui proposent d'utiliser ces algorithmes pour la sélection de caractéristiques destinées à la reconnaissance du locuteur.

Notre étude se base sur la capacité des AG à optimiser un système de façon non supervisée, sans connaissance à priori de son fonctionnement. L'algorithme possède donc une certaine autonomie dans ses moyens de résoudre le problème d'optimisation. Notre approche consiste à utiliser cette autonomie comme un outil d'exploration. Dans une précédente étude [6], cette méthodologie nous a permis de mettre en évidence l'importance de certaines informations spectrales pour la tache de segmentation et de regroupement du locuteur.

Les systèmes état de l'art de vérification du locuteur reposent sur un codage cepstral du signal (MFCC, LFCC, LPCC) suivi d'un classifieur de type mixture de gaussiennes (GMM). Une alternative aujourd'hui de plus en plus utilisée consiste à fusionner plusieurs systèmes de compositions différentes. Cette pratique peut être divisée

en deux catégories selon que cette différence porte sur le classifieur ou bien sur l'extracteur de caractéristiques. Notre étude repose sur ce second principe. On pourra citer les travaux de M. Zhiyou & al. [7] qui consistent à combiner les codeurs LPCC et MFCC, ceux de Poh Hoon Thian & al. [8] qui proposent de fusionner des informations relatives aux formants avec celles fournies par un codage conventionnel, et enfin l'étude de J. Campbell [Campbell2003] qui porte sur la fusion d'informations bas niveau (codage conventionnel) avec des informations haut niveau (prosodie, fréquence fondamentale, modèle de prononciation, etc.).

Dans cet article, nous proposons de fusionner trois systèmes basés sur des codeurs différents. Un algorithme génétique est mis en œuvre dans le but d'optimiser la complémentarité de ces codeurs. La figure 1 illustre ce principe.

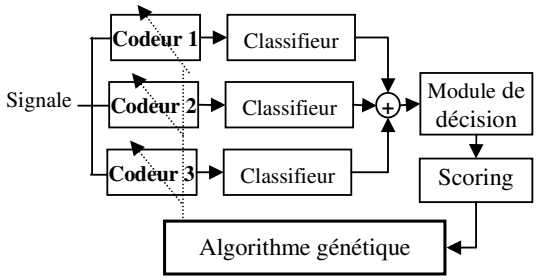

FIG. 1 : Optimisation du codage par algorithme génétique

Dans la section 2, une description du type de codage utilisé est présentée. L'algorithme génétique mis en œuvre est décrit dans la section 3. La section 4 présente les expériences réalisées ainsi que les résultats obtenus sur la base Nist SRE 2005.

2. Codage mis en œuvre

Les codages MFCC (Mel Frequency Cepstral Coefficients) et LFCC (Linear Frequency Cepstral Coefficients) reposent sur une analyse à court terme du signal par banc de filtres. La figure 2 représente un banc linéaire de filtres triangulaires. Ce processus d'extraction de caractéristiques comporte quatre étapes :

- Calculer le spectre en puissance de la trame analysée ;
- Calculer l'énergie correspondant à chaque filtre du banc ;
- Appliquer un logarithme aux coefficients obtenus ;
- Appliquer une transformée en cosinus discret (DCT).

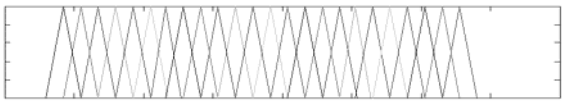

FIG. 2 : Banc de filtres linéaires

Pour la tâche spécifique de vérification du locuteur à partir de signaux téléphoniques, le codeur LFCC est réputé être le plus robuste.

Les codeurs mis en œuvre dans notre application sont des adaptations du codage LFCC. L'algorithme génétique sera utilisé pour optimiser le nombre de filtres du banc, le nombre de coefficients cepstraux extraits, ainsi que la position centrale et la largeur de bande de chaque filtre du banc.

3. Algorithme génétique

Un algorithme génétique (AG) est un outil d'optimisation. Son emploi permet de trouver les valeurs optimales d'un jeu de paramètres maximisant les performances du système. Un AG opère sur une population d'individus. Dans notre application, les individus sont des codeurs, définis par un jeu de paramètres appelés gènes. Ces gènes constituent une représentation condensée et adaptée des paramètres opérationnels de notre codeur. Le principe général de l'algorithme repose sur l'application itérative des opérations suivantes:

- **Muter** aléatoirement les gènes des individus de la population.
- **Décoder** ces gènes dans le but d'obtenir les paramètres opérationnels des codeurs.
- **Evaluer** les performances de chaque codeur.
- **Sélectionner** les meilleurs codeurs et les dupliquer pour revenir à une population de taille initiale.

Ces opérateurs seront définis dans la section 3.2.

3.1 Encodage des paramètres

La méthode d'encodage des paramètres occupe une place importante. Elle permet d'augmenter considérablement la vitesse de convergence de l'algorithme. De plus, elle permet de travailler dans un espace de paramètres réduit, qui diminue les risques de sur-apprentissage.

Les paramètres que nous avons choisis d'optimiser sont les suivants :

- Nf : Nombre de filtres du banc
- Nc: Nombre de coefficients cesptraux
- C_i: Fréquence centrale du $i^{ème}$ filtre du banc
- B_i: Largeur de bande du $i^{ème}$ filtre du banc

Les paramètres C et B sont encodés à l'aide de deux fonctions polynomiales décrites par les équations (1) et (2). Les gènes associés aux paramètres C et B sont les coefficients $\{gc_0,\ldots, gc_N\}$ et $\{gb_0,\ldots, gb_N\}$ de ces fonctions.

$$C_i = gc_0 + gc_1 \cdot \frac{i}{Nf} + gc_2 \cdot \frac{i}{Nf}^2 + ... + gc_N \cdot \frac{i}{Nf}^N \quad (1)$$

$$B_i = gb_0 + gb_1 \cdot \frac{i}{Nf} + gb_2 \cdot \frac{i}{Nf}^2 + ... + gb_N \cdot \frac{i}{Nf}^N \quad (2)$$

Ce type d'encodage permet de réduire considérablement le nombre de paramètres à optimiser tout en garantissant la régularité des bancs de filtres. Les paramètres *Nf* et *Nc* ne sont pas encodés et seront mutés directement.

3.2 Description de l'algorithme

L'algorithme mis en œuvre est constitué de quatre opérateurs: Mutation, Décodage, Evaluation, et Sélection. Ces opérateurs sont appliqués à la population courante *p(t)*, produisant une nouvelle génération *p(t+1)* par la relation *p(t+1)*=SEDM *(p(t))*.

La première étape de l'algorithme consiste à initialiser aléatoirement les gènes de chaque codeur de la population *p(0)*. Les opérateurs sont ensuite appliqués itérativement.

L'opérateur *Mutation* consiste à appliquer une petite variation aléatoire aux gènes.

L'opérateur *Décodage* a pour fonction de décoder les gènes afin d'obtenir les paramètres opérationnels des codeurs.

L'opérateur *Evaluation* est destiné à évaluer la performance de chaque individu de la population. Le critère d'évaluation utilisé sera présenté dans la section 3.3.

L'opérateur *Sélection* a pour fonction de sélectionner les *Ns* meilleurs individus de la population en fonction de leur performance. Ces individus seront ensuite dupliqués pour former une nouvelle population *p(t+1)* de *Np* individus.

L'application itérative de ces quatre opérateurs aura pour effet d'améliorer la performance moyenne de la population dans le temps.

3.3 Application à l'extraction de caractéristiques complémentaires

Notre objectif est d'optimiser un jeu de trois codeurs complémentaires (c.f. figure 1). La méthode mise en œuvre consiste à faire évoluer parallèlement trois populations de codeurs et de sélectionner à chaque génération les meilleurs combinaisons. Le critère d'évaluation consiste à mesurer le taux de reconnaissance (EER, Equal Error Rate) associé à chaque combinaison de codeur. La performance associée à un codeur est alors définie comme le taux d'erreur correspondant à la meilleure combinaison à laquelle ce codeur a participé

4. Résultats expérimentaux

4.1 Bases de données

Les bases de données utilisées sont extraites du corpus de la campagne d'évaluation du National Institute of Standart and Technologie 2005 (Nist SRE 2005) [9]. Cette base est constituée d'enregistrements de conversations téléphoniques passés au travers de différents canaux de transmission, et échantillonnés à 8kHz. La durée utile des signaux est en moyenne de 2 minutes 30 par locuteur pour les fichiers d'apprentissage et de test. La base utilisée pour l'évolution des codeurs est constituée de 10 locuteurs hommes et 10 locuteurs femmes. Le nombre de tests modèle x locuteur correspondant est de 2052. Une base de cross-validation est nécessaire pour le critère d'arrêt de l'algorithme. Cette base est constituée de 40 locuteurs répartis en 20 hommes et 20 femmes. Le nombre de tests correspondants est de 18947. La base de test est quant à elle composée de 50 locuteurs et de 50 locutrices, totalisant 116942 tests.

4.2 Système de vérification du locuteur

Le système de vérification utilisé est le système LIA SpkDet [10] distribué par le laboratoire d'Avignon. Ce système état de l'art est basé sur une modélisation à mixture de gaussiennes utilisant un modèle du monde (GMM-UBM). La modélisation choisie est une modélisation à 16 gaussiennes à matrice de covariance diagonale.

4.3 Paramètres de l'algorithme génétique

Les gènes relatifs aux fréquences centrales et aux largueurs de bande sont initialisés aléatoirement suivant une loi normale réduite. Les gènes relatifs au nombre de filtres sont initialisés à 24. Ceux relatifs au nombre de coefficients sont initialisés à 16.

Les paramètres utilisés pour l'évolution des codeurs sont les suivants :

- *Taille de chaque population* ***Np*** *: 20*
- *Nombre d'individus sélectionnés* ***Ns*** *: 5*
- *Ordre des polynômes codant pour le banc de filtre: 5*
- *Mutation des polynômes: variation aléatoire normale de +-0.1*
- *Mutation du nb de filtres : variation aléatoire uniforme de +-5.*
- *Mutation du nb de coefficients : variation aléatoire uniforme de +-3.*

4.4 Résultats

Dans cette section, les codeurs obtenus par évolution sont présentés et interprétés. Les taux d'erreurs associés sont détaillés et comparés à des systèmes de référence basés sur un codage LFCC et MFCC.

La figure 4 représente les bancs de filtres obtenus ainsi qu'une analyse statistique de la fréquence fondamentale et des formants. La table 1 détaille les caractéristiques des codeurs ainsi que les différents taux d'erreur (EER) obtenus sur la base de test. La méthode employée pour combiner les différents systèmes est une fusion arithmétique des vraisemblances (cf. figure 1).

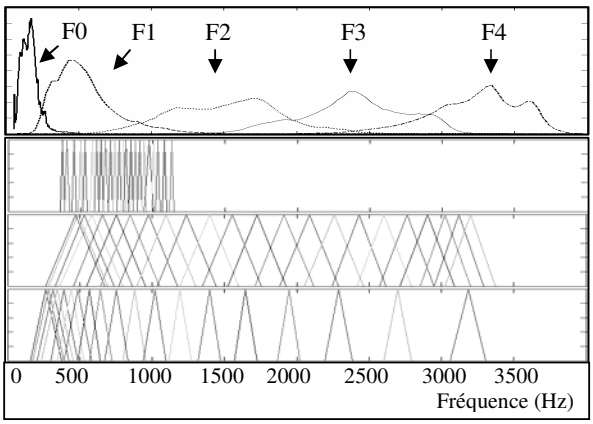

FIG. 3 : a - Distribution de la fréquence fondamentale et des formants (haut), b - Banc de filtres obtenus pour les codeurs C1, C2 et C3 (bas).

TAB. 1 : Résultats comparatifs : *Nf* = nombre de filtres par banc ; *Nc* nombre de coefficients cepstraux ; *Fmin, Fmax* = domaine fréquentiel couvert.

Codage	*Nf*	*Nc*	*F min* (Hz)	*F max* (Hz)	EER
LFCC	24	16	300	3400	**14.44%**
MFCC	24	16	300	3400	**14.88%**
C1	23	15	360	1145	22.90%
C2	25	20	266	3372	14.79%
C3	19	19	156	3309	16.07%
C1+C2	- -	- -	- -	- -	13.21%
C2+C3	- -	- -	- -	- -	13.45%
C1+C3	- -	- -	- -	- -	15.39%
C1+C2+C3	- -	- -	- -	- -	**12.69**%

Nous pouvons constater que le codeur C2 présente un banc de filtres couvrant l'ensemble utile du spectre, et permet d'obtenir un taux d'erreur proche de celui des codeurs conventionnels (LFCC, MFCC). Les codeurs C1 et C3 quant à eux semblent se focaliser sur des zones spécifiques du spectre. Dans le but de donner une interprétation fréquentielle aux bancs de filtres obtenus, une extraction de la fréquence fondamentale ainsi que des formants a été effectuée sur 40 locuteurs (20 hommes et 20 femmes). La figure 3a illustre les distributions de probabilité des fréquences centrales des formants et de la fréquence fondamentale.

Il apparaît d'une part que les informations relatives à la fréquence fondamentale n'ont pas été exploitées. D'autre part, que le codeur C1 semble se focaliser exclusivement sur les informations portées par le premier formant. Le codeur C3 quant à lui, présente également une forte densité de filtres centrés sur le premier formant tout en prenant en compte l'intégralité du spectre.

Les résultats obtenus par fusion montrent que les codages proposés sont complémentaires et qu'ils permettent une amélioration relative du taux d'erreur de 12% par rapport à un codage LFCC classique.

5. Conclusion

Nous avons proposé dans cet article d'utiliser un algorithme génétique pour optimiser un système d'extraction de caractéristiques adapté à la tâche de vérification du locuteur. Le principe d'extraction repose sur la combinaison de trois codeurs complémentaires. Le système obtenu nous a permis d'obtenir une amélioration significative des taux d'erreur de reconnaissance. Les solutions générées ont également permis de mettre en évidence la présence d'informations discriminantes dans la zone spectrale correspondant au premier formant.

Nos perspectives de recherche s'orientent vers l'étude de la robustesse de l'algorithme génétique relativement aux conditions initiales ainsi qu'à la base utilisée pour la phase d'évolution.

Références

[1] M. Chetouani, M. Faundez, B. Gas and J.L. Zarader , "Non-linear Speech Feature Extraction for Phoneme Classification and Speaker Recognition", *Nonlinear speech processing : Algorithms and Analysis*, Springer Verlag, 2005.

[2] K. Torkkola, "Feature Extraction by Non-Parametric Mutual Information Maximization", *Journal of Machine Learning Research, Vol. 3*, MIT Press, Cambridge, MA, pp. 1415-1438, 2003.

[3] Katagiri, S., *Handbook of Neural Networks for Speech Processing*, Artech House, Boston, 2000

[4] Chin-Teng L. Hsi-Wen N. and Jiing-Yuan H, "GA-based noisy speech recognition using two-dimensional cepstrum," In *Proc. Conf. Intl. IEEE Transactions on Speech and Audio Processing, .vol. 8*, pp 664-675, 2000.

[5] M. Zamalloa, G. Bordel, L.J. Rodriguez, M. Penagarikano, "Feature Selection Based on Genetic Algorithms for Speaker Recognition," *Conf. Intl. IEEE Odyssey, vol.*, no.pp.1-8, 2006.

[6] C. Charbuillet, B. Gas, M, Chetouani and J.L. Zarader, "Filter bank design for speaker diarization based on genetic algorithms," In *Proc. Conf. Intl. IEEE International Conference on Acoustics, Speech, and Signal Processing*, *vol. 1* pp 673-676, 2006.

[7] M. Zhiyou, Y. Yingchun, W Zhaohui, "Further feature extraction for speaker recognition," *IEEE International Conference on Systems, Man and Cybernetics*, *vol.5*, pp. 4153- 4158, 2003

[8] N. Poh Hoon Thian, C. Sanderson, S. Bengio, D. Zhang K. Jail Anil, "Spectral subband centroids as complementary features for speaker authentication" *Lect. notes comput. sci., vol. 3072*, 631-639, 2004

[9] 2005 NIST Speaker Recognition Evaluation site, www.nist.gov/speech/tests/spk/2005/

[10] LIA SpkDet system web site, http://www.lia.univ-avignon.fr/heberges/ALIZE/LIA RAL