Application d'un algorithme génétique à la synthèse d'un prétraitement non linéaire pour la segmentation et le regroupement du locuteur

Christophe Charbuillet, Bruno Gas, Mohamed Chetouani, Jean- Luc Zarader

Groupe Perception et Réseaux Connexionnistes (PRC)

Université Pierre et Marie Curie
Christophe.Charbuillet@lis.jussieu.fr, Bruno.Gas@upmc.fr, Mohamed.Chetouani@upmc.fr, zarader@ccr.jussieu.fr

ABSTRACT

Speech feature extraction plays a major role in a speaker recognition system. B. Gas & al. showed in [1] that a non linear filtering of speech can improve the feature extractor's ability. In this article we propose to use genetic algorithms to design a non-linear pre-processing of speech adapted to the speaker diarization task. The pre-processing system we present is based on artificial recurrent neural networks (ARNN). We used a genetic algorithm to find both the structure and the weights of the network.

Experiments are carried out using a state-of-the-art speaker diarization system. Results showed that the proposed method give significant improvements, reducing the diarization error rate from 17.38 % to 15.77 %.

1. INTRODUCTION

L'étape d'extraction de caractéristiques occupe une place fondamentale dans un système de reconnaissance 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, MFCC).
Depuis quelques années, un certain nombre de travaux se sont portés sur l'application de techniques non linéaires au problème de l'extraction de caractéristiques. Pitsikalis & Maragos [2] ainsi que Lindgren & al. [3] ont mis en évidence que les outils issus de la théorie du chaos sont applicables pour modéliser les phénomènes dynamiques non linéaires présents dans le signal de la parole. D'autre part, M. Chetouani & B. Gas [1] ont mis eux aussi en évidence qu'une transformation non linéaire du signal de parole permettait de faciliter l'extraction de caractéristiques discriminantes pour la reconnaissance de la parole.

Notre étude s'inscrit particulièrement dans la continuité de cette dernière approche. La transformation proposée par B. Gas & al. repose sur l'utilisation d'un réseau de neurones de type MLP (Multi Layer Perceptron) qui constitue un filtre non linéaire à réponse impulsionnelle finie. Nous proposons dans le présent article d'étendre ce principe à un filtrage non linéaire à réponse impulsionnelle infinie par l'utilisation d'un réseau de neurones récurrent. L'objectif étant d'élaborer un prétraitement adapté à la tâche de segmentation et de regroupement 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 Demirekler & Haydar [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 supervisé, 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.

Nous présentons ici l'application d'un AG à la synthèse d'un prétraitement du signal de parole basé sur un réseau de neurones récurrent. L'objectif étant d'explorer les potentialités de ce type de traitement pour la tâche de segmentation et regroupement du locuteur.

Nous abordons dans la section 2 une description de la tâche de segmentation et regroupement du locuteur ainsi que le système mis en œuvre. Dans la section 3 nous décrivons le filtre de prétraitement. La 4ème

section présente l'algorithme génétique utilisé. Les résultats obtenus sont exposés dans la section 5.

2. SEGMENTATION ET REGROUPEMENT DU LOCUTEUR

La tâche de segmentation et de regroupement du locuteur (SRL) consiste à identifier les segments de signal produits par le même locuteur. La phase de segmentation a pour objectif de détecter les moments de changement de locuteur. Elle est suivie d'une étape de regroupement qui consiste à étiqueter les segments obtenus en fonction du locuteur. Les applications de la SRL sont essentiellement tournées vers l'indexation de documents sonores.

Le système de segmentation et de regroupement mis en œuvre est basé sur le critère BIC (Bayesian Information Criterion), appliqué à une modélisation mono Gaussienne diagonale des vecteurs codes. Ces algorithmes sont implémentés dans l'outil logiciel audioseg [7]. Le principe du système de segmentation est de calculer la différence BIC entre deux fenêtres adjacentes du signal. Un changement de locuteur sera détecté à l'interface des fenêtres si la différence BIC est supérieure à zéro. Le fonctionnement de l'algorithme de regroupement est comparable au système de segmentation. Les segments ayant les différences BIC les plus faibles sont regroupés itérativement. L'algorithme s'arrête lorsque la distance minimale entre deux segments est supérieure à zéro. La différence BIC est donnés par:

$$dBIC(C_i, C_j) = -D_r(C_i, C_j) + \frac{\lambda}{2}\left(d + \frac{d(d+1)}{2}\right)\log\left(\mathrm{n_i} + \mathrm{n_j}\right)$$

avec:

$$D_r(C_i, C_j) = \frac{n_i + n_j}{2}.\log\left|\Sigma_{ij}\right| - \frac{n_i}{2}.\log\left|\Sigma_i\right| - \frac{n_j}{2}.\log\left|\Sigma_j\right|$$

Où C_i, C_j sont deux séquences de vecteurs codes représentant deux segments; $\mathrm{n_i}$ est la dimension de C_i et Σ_i sa matrice de covariance. λ est un paramètre pénalisant la complexité du modèle évalué.

3. RESEAUX DE NEURONES RECURRENTS & PRETRAITEMENT NON LINEAIRE

Les réseaux de neurones récurrents (RNR) sont des systèmes dynamiques non linéaires. La richesse dynamique de ces systèmes suscite un fort intérêt dans de nombreuses disciplines. Les applications dans le domaine du traitement du signal sont nombreuses. On pourra citer les travaux de B. Cessac & al. [8] qui ont proposé une méthode exploitant les propriétés dynamiques d'un RNR en mode chaotique pour transmettre un signal entre deux neurones distants, ou encore ceux de H. Jaeger [9] qui ont mis en évidence l'apport des RNR au problème de la prédiction de systèmes non linéaires. Nous proposons dans cet article, d'utiliser les RNR comme filtre de prétraitement du signal de parole.

Le modèle du filtre mis en œuvre est défini par:

$$u_i(t+1) = \tanh\left(\sum_{j=1}^{N} \mathbf{J}_{ij}.u_j(t) + \mathbf{E}_j.e(t) + \mathbf{B}_j\right)$$

$$s(t) = u_N(t)$$

où e et s représentent respectivement le signal d'entré et de sortie du filtre, N le nombre de neurones, $u_i(t)$ représente la sortie du neurone i au temps t, $\mathbf{J}_{ij}$ la connexion orientée du neurone j vers le neurone i, $\mathbf{E}_j$ la connexion de l'entrée vers le neurone j, $\mathbf{B}j$ la valeur du biais du neurone j.

4. ALGORITHME GENETIQUE

Un algorithme génétique est un outil d'optimisation. Son emploi permet de trouver les valeurs optimales d'un jeu de paramètres maximisant les performances du système. Les AG sont basés sur les théories Darwiniennes de l'évolution plus connues sous le nom de théorie de la sélection naturelle. Un AG opère sur une population d'individus. Dans notre application, les individus sont des filtres à RNR définis par leurs paramètres **J, E** et **B**. L'algorithme mis en œuvre est constitué de trois opérateurs: Sélection, Evaluation et Variation (SEV) [10]. Ces opérateurs sont appliqués à la population courante *p(t)*, produisant une nouvelle génération *p(t+1)*, par la relation *p(t+1)=SEV (p(t))*.

La première étape de l'algorithme consiste à initialiser aléatoirement les paramètres de chaque filtre de la population *p(0)*. Les opérateurs Sélection, Evaluation et Variation sont ensuite appliqués itérativement.

L'opérateur *Variation* consiste à faire varier les paramètres des filtres de la population. Dans notre application, un opérateur de création ou d'élimination de connexion est appliqué avec une probabilité P_{connex}, suivie par un opérateur de modification des poids des connexions, modifiant le poids d'une connexion ρ par**:** $\rho \leftarrow \rho + \alpha.\phi$ où α représente le facteur de variation du poids et ϕ est une variable aléatoire de distribution uniforme sur [-1 ; 1].

L'opérateur *Evaluation* est destiné à évaluer la performance de chaque individu de la population. Dans notre application, la performance sera quantifiée par le taux d'erreur de regroupement du locuteur du système intégrant le filtre de prétraitement à évaluer.

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 trois opérateurs aura pour

effet d'améliorer la performance moyenne de la population dans le temps, et donc de produire des filtres de prétraitement adaptés au problème du regroupement du locuteur. La figure n°1 illustre l'application de l'algorithme mis en oeuvre à l'optimisation du filtre de prétraitement.

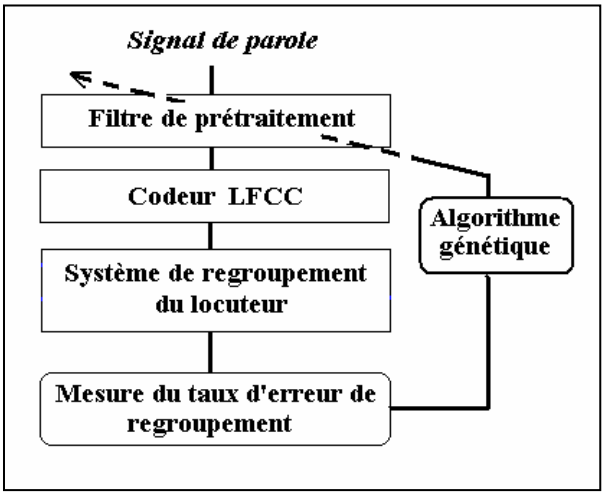

Figure 1 : Optimisation du filtre de prétraitement par algorithme génétique.

5. EXPERIENCES ET RESULTATS

Les expérimentations ont été menées sur la tache de segmentation et de regroupement du locuteur (SRL) de la campagne d'évaluation ESTER [11].

5.1. Bases de données

La base de donnée ESTER est composée de 40 heures de signaux audio provenant de l'enregistrement de 4 radios différentes échantillonnés à 16 kHz. Cette base est très riche: elle contient des signaux de parole enregistrés en studio, des interventions téléphoniques, de la parole sur font musical, avec des qualités d'enregistrement variables. Les fichiers audio ont d'une durée comprise entre 15 et 60 minutes. Le nombre de locuteur intervenant dans un fichier est compris entre 1 et 39.

La métrique de mesure des performances SRL est celle utilisée pour la campagne Nist RT 2003. Cette mesure est basée sur la qualité d'appariement du regroupement des locuteurs.

5.2. Bases d'évolution

Deux bases distinctes sont utilisées pour l'évolution des filtres. La première nommée "*base d'évolution"*, composée de 4 heures de signaux, est destinée à l'évaluation et à la sélection des filtres. La seconde nommée "*bases de cross validation",* composée de 8 heures d'enregistrement, a pour fonction de mesurer la capacité de généralisation des filtres obtenus.

5.3. Protocole expérimental

Les expériences d'évolution du filtre de prétraitement ont été menées sur la tâche de regroupement seul. La tâche de segmentation ayant été effectuée préalablement à partir d'un codage LFCC à 24 filtres et 16 coefficients) sans prétraitement. Le système de regroupement mis en œuvre comprend le filtre de prétraitement élaboré, suivi d'un codage LFCC 24/16 ainsi que du module de regroupement du locuteur. Les paramètres de l'algorithme génétique utilisés pour cette expérience sont: Np = 50; Ns = 15; N = 4; $\alpha = 0.05$; $P_{connex} = 0.01$;

L'initialisation de la population *p(0)* consiste à connecter toutes les connexions des réseaux et à initialiser leurs poids aléatoirement entre -0.6 et +0.6 pour les matrices **J** et **B** et entre -3500 et +3500 pour la matrice **E**. Ces valeurs d'initialisation étant déterminées empiriquement.

5.4. Evolution des filtres

La figure n°2 représente l'évolution du taux d'erreur de regroupement du meilleur individu de la population de filtres sur les bases d'évolution et de cross validation.

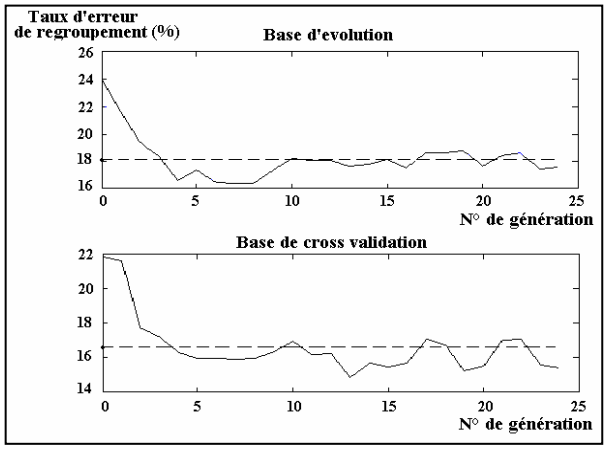

Figure 2 : Evolution des performances des filtres de prétraitement. Trait plein: meilleur individu de la population. Trait discontinu: système de référence.

On peut observer que les performances des filtres dépassent rapidement celle du système de référence. Le filtre sélectionné à l'issus de cette phase d'évolution est celui qui présente les meilleures performances sur la base de cross validation (génération n°13).

5.5. Résultats obtenus

L'évaluation des performances du filtre de prétraitement obtenu a été effectuée sur les bases de développement (DEV2) et de test (TEST2) de la campagne d'évaluation ESTER phase 2. La première base est composée de 8 heures d'enregistrements radiophoniques. La composition de cette base est identique à celles des bases d'évolution et de cross validation. La seconde base est celle utilisée pour l'évaluation des soumissions de la campagne. Cette

base est également composée de 8 heures d'enregistrements audio. La particularité de cette base est la présence de deux heures d'enregistrements provenant de deux radios inconnues. Le tableau n°1 présente les résultats obtenus sur ces bases avec le système de référence et le système intégrant le filtre de prétraitement élaboré.

Tableau 1 : Taux d'erreur de regroupement sur les bases DEV2 et TEST2.

	Base DEV2	**Bases TEST2**
Système de référence	**17.38%**	**21.52%**
Système avec prétraitement	**15.77%**	**21.59%**

Les résultats font apparaître une amélioration significative des performances sur la base DEV2. Cependant, cette amélioration ne se répercute pas sur la base TEST2 où les performances du système proposé sont comparables à celles du système de référence. Ceci peut s'expliquer par le fait que cette base intègre deux heures de signaux provenant de deux nouvelles radios non incluses dans les autres bases.

7. CONCLUSION

Nous avons proposé dans cet article d'utiliser un algorithme génétique pour la synthèse d'un prétraitement non linéaire de la parole adapté au problème de la segmentation et du regroupement du locuteur. Le système obtenu nous a permis d'obtenir une amélioration significative des performances de regroupement du locuteur, réduisant le taux d'erreur de 17,38% à 15,77%.

Les systèmes dynamiques non linéaires de part leur complexité et leur richesse permettent d'envisager de nouvelles méthodes de traitement du signal. Nos perspectives de recherche s'orientent vers l'étude de la capacité de ce type de système à l'extraction de caractéristiques discriminantes pour la reconnaissance du locuteur.

8. REMERCIMENTS

Nous tenons à remercier particulièrement Guillaume Gravier pour nous avoir fournis les outils logiciels de segmentation et de regroupement du locuteur présentés dans cet article.

BIBLIOGRAPHIE

[1] B. Gas, JL. Zarader, C. Chavy, M. Chetouani. Discriminant neural predictive coding applied to phonem recognition. *Neurocomputing*, 56:141-166, 2004.

[2] Pitsikalis V. and Maragos, P. Speech analysis and feature extraction using chaotic models. In *Proc. Intl. IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 533-536, 2002

[3] Lindgren A.C. Johnson M.T. and Povinelli, R.J. Speech recognition using reconstructed phase space features. In *Proc. Con.f Intl. IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 60-63, 2003

[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.*volume 8, pages 664-675, 2000.

[5] Demirekler, M. and Haydar, A. Feature selection using genetics-based algorithm and its application to speaker identification. In *Proc. Con.f Intl. IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 329-332, 1999.

[6] C. Charbuillet, B. Gas, M. Chetouani and J.L. Zarader. Filter bank design for speaker diarization based on genetic algorithms. To be pub. In *Proc. Conf. Intl. IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2006.

[7] Gravier G. and Betser M., Audioseg: Audio Segmentation Toolkit, 2005 *http://www.irisa.fr/metiss/guig/index-en.html*

[8] B. Cessac and J-A Sepulchre, Stable resonances and signal propagation in a chaotic network of coupled units (2004). *Physical Review E,* volume 70, 056111, 2004.

[9] Jaeger, H, Adaptive nonlinear system identification with echo state networks, in *S. T. S. Becker & K. Obermayer, eds, Advances in Neural Information Processing Systems 15, MIT Press, Cambridge, MA*, page. 593-600, 2003.

[10] F. Pasemann, U. Dieckmann, and U. Steinmetz, Evolving Structure and Function of Neurocontrollers, In *Proc. Conf. Congress on Evolutionary Computation Journal*, IEEE Press US, Piscataway, page. 1937-1978, 1999.

[11] S. Galliano, E. Geoffrois, D. Mostefa, K. Choukri, J.F. Bonastre, and G. Gravier, The Ester Phase II Evaluation Campaign for the Rich Transcription of French Broadcast News, *Proceedings of Eurospeech/Interspeech'05*, pages 1149-1152, 2005.