Représentations de séquences de parole en espaces de faible dimensionalité

J. Arias, R. André-Obrecht, J. Farinas

IRIT/SAMOVA, Université Paul Sabatier
118 route de Narbonne, 31062 Toulouse, France
{arias,obrecht,jfarinas}@irit.fr
http ://www.irit.fr/-Equipe-SAMoVA

ABSTRACT

In this article we study several low-dimensional representations of speech sequences. Using the cepstra parameters of a speech sequence we create a Gaussian mixture model (GMM) that represents the sequence in several systems. In the first approach, symmetric Kullback-Leibler divergence between models is approximated and then projected in Euclidian space using Multidimensional Scaling. In the second system, we propose a decomposition of the speech sequences to model separately stationary and transitory speech units. Finally, GMMs created from a UBM MAP-adaptation are used to define supervectors, which are visualized with the aid of Principal Components Analysis. We use spectral clustering to analyze the grouping properties of each system.

Keywords: spectral clustering, dimensionality reduction, speech segmentation

1. Introduction

L'explosion des collections de données audio et la nécessité d'analyser leur contenus et de les classifier a motivé un intérêt certain vers les algorithmes d'apprentissage automatique non supervisés [1]. Ces études nécessitent de définir des mesures de similarité entre séquences de vecteurs acoustiques, chaque vecteur correspondant à l'analyse d'un trame de signal de l'ordre de la centiseconde. Pour faire face à la dimension élevée de l'espace de représentation des trames et à la longueur variable des suites de vecteurs, il est d'usage de représenter une séquence de parole par les paramètres d'une distribution probabiliste, le plus souvent un mélange de lois gaussiennes. La dimension de l'espace de représentation de la suite est alors fixée, mais elle reste élevée. La réduction de la dimensionalité s'avère nécessaire pour ne traiter que l'information utile et atteindre des signatures robustes, comme dans tout problème d'analyse des données. Une très forte réduction peut également aider à visualiser une collection de séquences et à relier *a posteriori* des regroupements trouvés automatiquement à des connaissances que l'on avait *a priori*.

Beaucoup de techniques de paramétrisation, transformation et sélection de données appropriées à des tâches spécifiques [2] ainsi que des méthodes de réduction de la dimensionalité de données sont actuellement approfondies. Au cours de notre étude nous recherchons si une réduction très forte de la dimensionalité peut avoir un intérêt en traitement automatique de la parole et guider un traitement ultérieur.

Partant de l'hypothèse énoncée précédemment à savoir qu'une séquence s est décrite par un mélange de lois gaussiennnes GMM_s, nous recherchons un nouvel espace dans lequel seules quelques coordonnées pourraient être significatives en terme de regroupement. Trois espaces sont ainsi obtenus :

- Deux d'entre eux s'obtiennent en considérant qu'un GMM est avant tout une loi et que l'espace recherché doit rendre compte des distances entre lois. L'algorithme d'échelle multidimensionnelle est utilisé dans ce but.
- Le troisième est plus classique puisqu'il fait appel à une analyse en composantes principales (PCA) traditionnel dès lors qu'un GMM est assimilé à un supervecteur.

Pour confronter ces trois types d'analyse, nous avons traité une collection de séquences de parole, avec comme objectif le regroupement en locuteur :

1. Le corpus est formé de 180 fichiers audio (30 séquences de 6 locuteurs différents, 3 hommes et 3 femmes) obtenus de la base de données ANITA. Les séquences sont des segments d'une durée d'environ 7 secs de parole 'phonétiquement équilibrée' et échantillonnées à 16 kHz. Une paramétrisation MFCC de ces séquences est réalisée avec une fenêtre d'analyse de 16 ms et un chevauchement de 8 ms entre trames : 15 coefficients cepstraux avec ses dérivées respectives sont utilisés.

2. Le regroupement spectral est utilisé pour mettre en évidence des clusters dans le nouvel espace. Le regroupement de données basé sur une décomposition spectrale est proche en sémantique de celui perçu par les humains. En contraste avec l'algorithme des k-means, il est capable de trouver des classes de structure non convexe. L'algorithme de regroupement spectral proposé par [3] est une approximation de la solution au problème (NP-complet) de la séparation d'un graphe en q-groupes. Si A est une matrice d'affinité non négative et symétrique qui représente les distances entre les points y_n et D est une matrice diagonale dont la valeur (n, n) est la somme de la ligne n de A, les valeurs propres du Laplacian $L = D^{-1/2}AD^{-1/2}$ donnent une indication de la structure et du nombre de groupes de l'ensemble Y. L'examen des valeurs propres et en particulier l'apparition d'une différence

importante entre elles (eigengap) indique le nombre de groupes pertinents pour les données observées.

Chacun des trois paragraphes suivants est dédiée à la définition de ces trois espaces et à leur analyse à l'aide du regroupement spectral. Compte tenu du corpus étudié, nous espérons pouvoir associer aux clusters ainsi mis en évidence un locuteur. Une synthèse des résultats amènera naturellement à quelques perspectives.

2. Représentation de distances statistiques entre distributions de probabilité

2.1. Système KL

Comme dit précédemment, ce premier système résulte de la paramétrisation d'une séquence de parole par un GMM et du fait qu'un GMM est avant tout une loi de probabilité. La dissimilarité statistique $\delta_{ij},\ i,j = 1,\ldots,N$ entre deux lois GMM_i et GMM_j est estimée avec la divergence symétrique de Kullback-Leibler [4]. Cette divergence est la somme de deux divergences orientées $KL(GMM_i/GMM_j)$ et $KL(GMM_j/GMM_i)$:

$$\delta_{ij} = \frac{1}{2}(KL(GMM_i/GMM_j)+KL(GMM_j/GMM_i)) \tag{1}$$

Pour déterminer $KL(GMM_i/GMM_j)$ nous utilisons l'échantillonnage de Monte-Carlo pour générer un ensemble aléatoire X à partir de GMM_i et puis nous calculons la moyenne du log des taux de vraisemblance $GMM_i(X)/GMM_j(X)$. On fait du même avec $KL(GMM_j/GMM_i)$ pour obtenir la distance statistique δ_{ij}. L'algorithme d'échelle multidimensionnelle

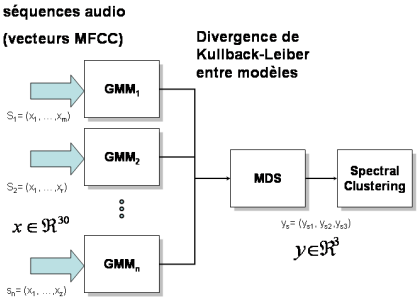

Fig. 1: Système KL. Un GMM modèlise une distribution de paramètres acoustiques. La distance entre modèles est approchée par la divergence symétrique de Kullback-Leiber et représentée dans un espace euclidien au moyen de l'algorithme MDS. La configuration résultante est analysée par le regroupement spectral.

(MDS) est alors appliqué. Cet algorithme permet, à partir des distances entre points $X \in \mathcal{R}^P$, de déterminer un système de coordonnées $Y \in \mathcal{R}^Q$ qui préserve ces distances. L'idée fondamentale du MDS est le calcul d'un produit scalaire en fonction de la distance entre les vecteurs. Néanmoins, cette relation est valable pour des vecteurs centrés [5], c'est pour quoi l'expression $\langle y_i \cdot y_j \rangle$ dépend non seulement des distances 2-à-2 d^2_{ij}, d^2_{ki}, et d^2_{kj} mais de toutes les autres distances entre points :

$$\langle y_i \cdot y_j \rangle = -\frac{1}{2}(d^2_{ij} - \frac{1}{n}\sum_{i=1}^{n} d^2_{ij} - \frac{1}{n}\sum_{j=1}^{n} d^2_{ij} + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} d^2_{ij}) \tag{2}$$

Si l'on décompose la matrice de produits scalaires $Y = V\Lambda V'$, les vecteurs propres $\sqrt{\lambda_q} v_q,\ q = 1,\ldots,Q$ peuvent être utilisés comme une approximation de Y. Pour le Système KL, les distances $d_{ij}, i,j = 1,\ldots,n$ utilisées dans l'éq. 2 sont en fait les dissimilarités δ_{ij} de l'éq. 1. La sortie de MDS est l'ensemble Y de vecteurs 3-dimensionnels, et chaque vecteur y_n représente une séquence audio de la base de données (voir fig. 1).

2.2. Expériences et interprétation des résultats

La représentation $Y \in \mathcal{R}^3$ des séquences est montrée dans la figure 2. Les meilleurs résultats sont obtenus avec 32 composantes par GMM. L'échantillonnage de Monte-Carlo pour estimer la divergence KL est fixé à 5000 vecteurs par modèle. Pour avoir une idée plus précise du regroupement des 6 locuteurs de test dans l'espace, la matrice A est obtenue à l'aide de la fonction RBF ($\sigma = 100$) et la matrice Laplacienne est diagonalisée pour observer ses valeurs propres. L'eigengap signale 6 groupes dans l'ensemble, mais leur identification dans l'espace Y ne correspond pas exactement aux groupes de locuteurs du corpus. La configuration des paramètres est robuste, elle est tolérante aux variations du nombre de composantes des GMM et à la taille de la fenêtre d'analyse MFCC.

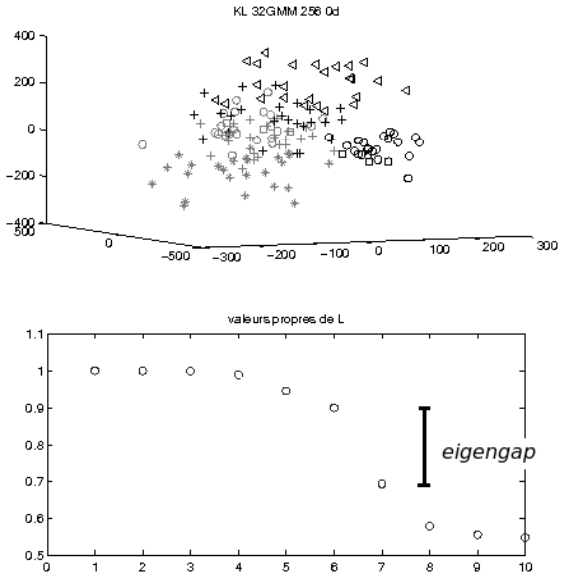

Fig. 2: Représentation en faible dimension de la divergence de Kullback-Leiber entre GMM. En haut, chaque point représente une séquence de parole et chaque symbole signale un locuteur. En bas, les principales valeurs propres du Laplacian montrent la présence de 6 clusters dans l'ensemble.

3. Modélisation différenciée Consonne - Voyelle

Dans cette approche nous ajoutons une étape de pré-traitement au système précèdent pour extraire des unités élémentaires différenciées et les caractériser par autant de GMM que de catégories d'unités. Ce pré-traitement s'apparente à une classification grossière du signal de parole. Notre première expérience utilise les classes consonnes et voyelles.

3.1. Système KL-CV

La segmentation et l'étiquetage SCV [6] utilise la divergence forward-backward (Fbd) pour réaliser une segmentation temporelle du signal. La Fbd localise des changements dans le modèle autoregressif de deux fenêtres glissantes d'analyse pour définir les frontières de segments stationnaires ou quasi-stationnaires dans le signal. Les segments ainsi déterminés sont ensuite classifiés comme 'noyau vocalique' (V), 'zone consonantique' (C) ou silences, grâce à l'inspection de l'énergie dans la bande spectrale de 0.35-1kHz [7].

A partir des unités C et des unités V extraites sur chaque séquence s_n de parole, sont appris deux modèles GMM (GMM_sC, GMM_sV). Les calculs de la distance de KL sur chaque sous-ensemble GMM_SC et GMM_SV permettent par la méthode MDS de projeter l'ensemble de séquences dans deux espaces différents Y_C et Y_V. Un synoptique du système est reproduit en figure 3.

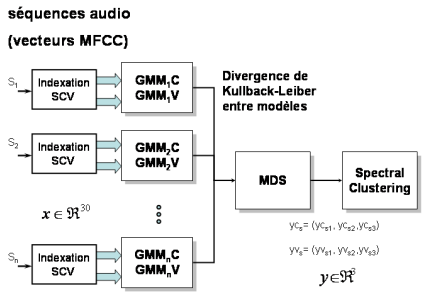

Fig. 3: Système KL-CV. Une étape de pré-traitement est ajoutée au Système KL pour modéliser séparément les unités phonétiques 'vocaliques' (V) et 'consonantiques' (C).

3.2. Expériences

Etant donné la 'spécialisation' des GMM, il est désormais suffisant d'utiliser 8 composantes par modèle, au lieu de 32 dans le Système KL, ce qui accélère les calculs de la divergence de Kullback-Leibler. Les meilleurs résultats sont obtenus avec une fenêtre d'analyse MFCC de 8ms et la matrice A est calculée avec le paramètre $\sigma = 30$.

La projection Y_C des $GMMC$ (voir figure 4) est similaire à celle du Système KL. Les eigengaps des deux systèmes se ressemblent. Visuellement, les résultats de la projection Y_V montrent une meilleure séparation de locuteurs que le Système KL ; néanmoins, deux locuteurs font pratiquement partie du même groupe, et deux autres sont assez proches. L'eigengap est mieux défini, il indique 4 clusters.

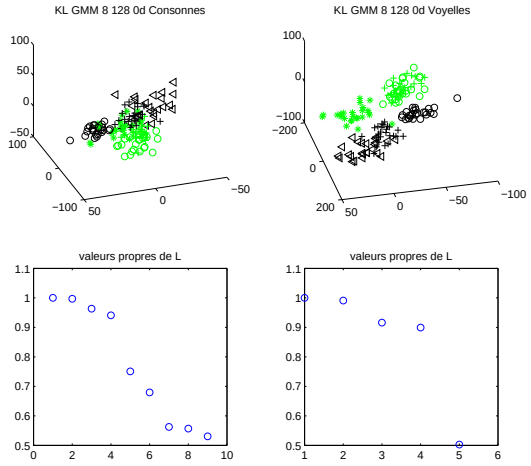

Fig. 4: Modélisation GMM différenciée. Á gauche, les modèles $GMMC$ sont appris avec les segments 'consonantiques' des séquences de parole. Á droite, les $GMMV$ modélisent les segments 'vocaliques'. Un eigengap mieux défini est obtenu par les GMMV.

4. Les supervecteurs GMM

Le calcul de la divergence KL est très coûteux en temps d'exécution et, si le nombre d'échantillons de l'estimation de Monte-Carlo est limité, il est peu robuste. Une alternative d'utilisation des GMM comme représentants de séquences acoustiques est de concaténer les vecteurs moyennes et créer ainsi un 'supervecteur' par GMM. Plusieurs mesures de similarité sont proposées pour comparer ces vecteurs [8], mais une simple analyse PCA nous donne une information pertinente sur la position de chaque vecteur dans l'ensemble.

4.1. Système SV

Ce troisième système basé sur des supervecteurs GMM et une analyse PCA est illustré dans la figure 5. Néanmoins, la méthode pour estimer les GMM a été modifiée pour rendre les GMM compatibles entre eux en terme de supervecteurs. Un modèle GMM global caractérise la répartition 'totale' des données acoustiques d'entrée ; il est appelé Modèle du Monde (UBM-GMM). La technique d'adaptation la plus répandue du modèle UBM-GMM aux données d'entrée est l'adaptation par *Maximun a posteriori* [9]. Nous utilisons cette technique dans le but d'adapter le modèle UBM-GMM à chaque séquence s_n de parole analysée et fournir leur modèle GMM_n : seules les moyennes des lois gaussiennes sont adaptées. Dans l'espace des supervecteurs est appliqué une analyse PCA pour visualiser les séquences dans un espace de plus faible dimension.

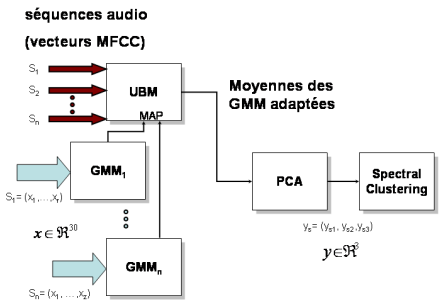

Fig. 5: Système SV. Un GMM issue de l'adaptation MAP d'un modèle universel apporte ses moyennes pour créer un supervecteur qui le représente. Une analyse PCA permet leur visualisation.

4.2. Expériences

Plusieurs tailles du modèle UBM-GMM sont testées (de 16 à 128 composantes), mais les meilleurs résultats sont obtenus avec un modèle UBM-GMM à 32 composantes. L'analyse PCA a été comparée à d'autres méthodes (Projection de Sammon, Isomap) pour réduire la dimensionalité des supervecteurs GMM avec les mêmes résultats. La projection montre une bonne séparation de locuteurs, proche de celle obtenue avec le Système KL-CV($GMMV$) (voir figure 6). Les valeurs propres du Laplacian ($\sigma = 0.055$) reflètent cette configuration, néanmoins l'eigengap est plus difficile à établir.

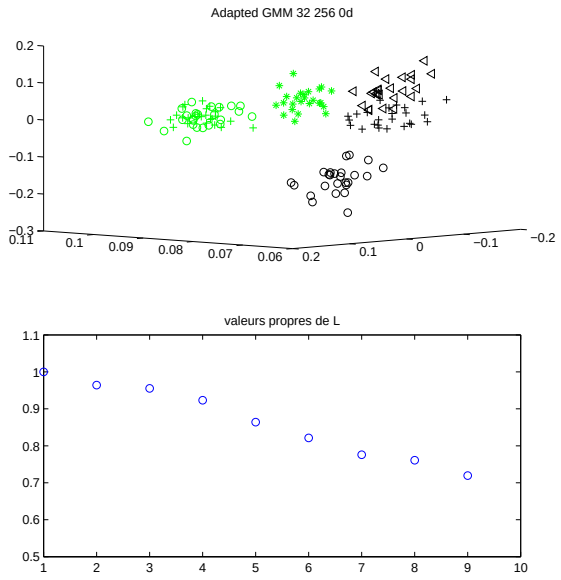

Fig. 6: Représentation en 3D des supervecteurs GMM à l'aide d'une analyse PCA.

5. Analyse des résultats

Nous présentons divers possibilités de visualiser des séquences de parole en espaces 3D. Nous considérons cette perspective intéressante car des traitements ultérieures à cette projection (regroupement automatique, classification supervisé, etc) peuvent être facilement contrôlés par l'utilisateur. Pour avoir une idée plus précise de la qualité des résultats, on propose le calcul d'une matrice Laplacienne ainsi que l'étude de la variation de ses valeurs propres de manière à identifier le plus grand écart entre elles -l'eigengap- et le nombre de clusters stables dans l'ensemble. Plus grand est l'eigengap, mieux les clusters sont définis. Les résultats montrent que la projection du Système KL-CV($GMMV$) présente l'eigengap le mieux défini de toutes les approches. Le Système KL présente un eigengap précis, mais ses clusters ne correspondent pas exactement aux locuteurs de la base de données. En fait, les projections Y issues du Système SV et du Système KL-CV($GMMV$) sont les plus appropriées à l'identification de locuteurs. Il est aussi à considérer le temps d'exécution des systèmes, car le calcul de la distance de Kullback-Leibler dans le Système KL est extrêmement coûteux (il est allégé dans le Système KL-CV grâce à la réduction d'information du pré-traitement) tandis que pour l'analyse PCA du Système SV il est insignifiant.

6. Remerciements

Le travail de José Arias est soutenu par une bourse du Programme de Coopération Scientifique et Technique franco-mexicain SFERE-CONACyT.

Références

[1] T. Zhang and C. Kuo, "Hierarchical system for content-based audio classification and retrieval," in *Conference on Multimedia storage and Archiving Systems*, Nov. 1998, vol. 3527, pp. 398–409.

[2] E. Sungur, "Overview of multivariate statistical data analysis," `http://mrs.umn.edu/~sungurea/multivariatestatistics/overview.html`, 2007.

[3] A. Ng, M. Jordan, and Y. Weiss, "On spectral clustering : Analysis and an algorithm," in *NIPS*, 2001, vol. 13.

[4] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, Academic Press, 1990.

[5] I. Borg and P. Groenen, *Modern Multidimensional Scaling : Theory and Applications*, Springer, 1997.

[6] R. André-Obrecht, "A new statistical approach for automatic speech segmentation," *Transactions on Audio, Speech, and Signal Processing*, vol. 36, no. 1, Jan. 1988.

[7] F. Pellegrino, J. Farinas, and R. André-Obrecht, "Comparaison of two phonetic approaches to language identification," in *EUPSICO*, Budapest, Hongrie, Sept. 1999, pp. 399–402, 5-9 sep.

[8] W. Campbell, D. Sturim, and D. Reynolds, "Support vector machines using gmm supervectors for speaker verification," *Signal processing letters*, vol. 13, no. 15, May 2006.

[9] J. Gauvain and C. Lee, "Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains," *Transactions on speech and audio processing*, vol. 2, no. 2, pp. 291–298, 1994.