Étude de la réduction non linéaire de la dimension du signal de parole

José Anibal Arias, Régine André-Obrecht, Jérôme Farinas et Julien Pinquier

IRIT - Équipe SAMOVA
Université Paul Sabatier, 118 route de Narbonne, 31062 Toulouse Cedex 9, FRANCE
{arias,obrecht,jfarinas,pinquier}@irit.fr
http ://www.irit.fr/recherches/SAMOVA

ABSTRACT

In this article we study some results of the non-linear dimensionality reduction of speech vectors. Spectral clustering, Kernel PCA, Isomap, Laplacian eigenmaps and Locally Linear Embedding are related non-supervised methods that help to discover important caracteristics from data such as high-density regions or low-dimensional surfaces (manifolds). This reduction of dimension is a necessary step when we want to model speech sequences with discriminative/generative functions such as Support Vector Machines or Gaussian Process.

1. INTRODUCTION

Le signal de parole est une source d'information ne pouvant pas être modélisée de façon exhaustive sans prendre en compte sa dimension temporelle. On ne peut se limiter à le considérer comme une variable aléatoire représentée par un vecteur de paramètres extrait en utilisant un fenétrage fixe (centiseconde par exemple). Pour le signal de parole, les modèles de Markov cachés (HMM) ont par exemple permis d'étendre la modélisation statistique par mélange de lois gaussiennes (GMM), en prenant en compte les enchaînements temporels.

Nous souhaitons étudier l'adéquation de méthodes discriminantes telles les machines à vecteur support (SVM) et génératives telles les Processus Gaussiens (GP) avec le signal de parole. Pour pouvoir prendre en compte l'évolution temporelle du signal, il est courant avec les SVM de modéliser des séquences d'observations [12]. La difficulté va alors provenir de la dimension d des vecteurs d'entrée dans le système et de la taille N de l'ensemble des données d'entrainement. En effet, ces deux paramètres crutiaux influent sur la dimension des matrices de traitements internes aux méthodes, il est nécessaire de les contraindre pour pouvoir obtenir des solutions réalisables.

Nous proposons dans cet article d'étudier les variétés ainsi que le regroupement permettent de réduire le nombre des variables pour le signal de parole, dans le but d'appliquer par la suite des modélisations par SVM et GP. Cela devrait introduire des solutions plus robustes (moins sensibles au bruit et aux données aberrantes) et permettre l'analyse visuelle de la structure de l'information.

Dans la section 2 nous présentons synthétiquement plusieurs méthodes que nous avons étudiées, basées sur une décomposition spectrale. Dans la section suivante nous détaillons les expériences menées ainsi que analyse des ces résultats.

2. MÉTHODES SPECTRALES

Les méthodes basées sur une décomposition spectrale estiment de manière non-supervisée les principales fonctions propres d'un opérateur qui dépend d'une densité de données inconnue. On est capable d'utiliser leurs résultats pour généraliser les fonctions propres à des données externes à l'ensemble d'entraînement [3]. L'hypothèse est que, en dépit de la haute complexité du sujet, les sons de la parole sont groupés en variétés non-linéaires de relativement faible dimension liées au processus de production du signal acoustique.

Les algorithmes spectraux d'estimation de variétés s'appuient, pour un ensemble de vecteurs d'entrée $x_{i,\ i=1...N}$, $x_i \in \boldsymbol{R}^d$, sur une matrice de similarité $\overline{K}_{N\times N}$ et conduisent à rechercher ses principaux vecteurs et valeurs propres. La représentation en faible dimension de chaque vecteur x_i en entrée est obtenue en utilisant les j premiers vecteurs propres de $\overline{K}(j \ll d)$. Si l'on veut calculer tenir compte d'un nouveau vecteur x, on utilise la formule de Nyström [5] pour évaluer l'extension des vecteurs propres.

En traitement automatique de la parole, les vecteurs d'entrée sont généralement issus d'une paramétrisation MFCC ou LPC avec d inférieur à la cinquantaine. Les algorithmes spectraux projettent ces données en vecteurs $y_{i,\ i=1...N}$ de dimension très inférieure, idéalement 2 ou 3, pour pouvoir les analyser visuellement.

2.1. Kernel PCA

Schématiquement l'analyse en composantes principales (PCA) est un changement de repère qui vise à privilégier les axes de variance maximale par rapport à un ensemble de données. Les axes où la variance des données est réduite peuvent être éliminés pour atteindre une réduction de la dimensionalité avec une perte minimale d'information. La transformation est, par essence, linéaire (matrice de passage orthogonale). Or, pour les vecteurs de la parole, il est souhaitable de pouvoir atteindre des relations non-linéaires ; la méthode Kernel PCA est une première extension de PCA qui l'envisage.

Kernel PCA réalise une analyse en composantes principales dans l'espace $\boldsymbol{K}$ appelé «espace de caractéristiques» généré à l'aide d'une fonction noyau $k(x_i, x_j)$ tel que $x_i, x_j \in \boldsymbol{R}^d,\ k(x_i, x_j) = \Phi(x_i) \cdot \Phi(x_j)$ [11]. La transformation non-linéaire $\Phi(x)$ implicite dans la fonction noyau permet au Kernel PCA de trouver un sous-espace qui, plus qu'une réduction de dimensionalité, est le résultat d'un processus d'extraction d'information.

Si X est la matrice de l'ensemble des transformées des données d'apprentissage, dans l'espace des caractéristiques, centrées, la matrice C de covariance de ces transformées vérifie : $NC = X'X$. Si $\overline{K}$ est la matrice « kernel », par définition, $\overline{K} = XX'$.

Il vient que si Xu est vecteur propre de NC associé à la valeur propre λ, X est un vecteur propre de $\overline{K}$ associé à la même valeur propre :

$$\begin{aligned} u &= \lambda^{-1/2} \sum_{i=1}^{N} v_i \Phi(x_i) \\ v &= \frac{Xu}{\sqrt{\lambda}} \end{aligned} \tag{1}$$

En associant de cette manière à tout vecteur propre u_j de NC, le vecteur v_j et la valeur propre λ_j, la projection d'un nouveau vecteur $\Phi(x)$ sur la direction u_j est donnée par :

$$\begin{aligned} P(x)_{u_j} &= u_j' \Phi(x) = \left\langle \sum_{i=1}^{n} \alpha_i^j \Phi(x_i), \Phi(x) \right\rangle \\ &= \sum_{i=1}^{n} \alpha_i^j k(x_i, x) \end{aligned} \tag{2}$$

où $\alpha^j = \lambda_j^{-1/2} v_{j,\, j=1\ldots d'}$.

Une réduction d'information s'obtient en conservant les valeurs et vecteurs propres les plus élevés.

2.2. Isomap

Isomap [13] est une généralisation non-linéaire de l'algorithme d'échelle multidimensionnelle (MDS). MDS permet à partir des distances euclidiennes entre points, de déterminer un système de coordonnées réduit qui préserve les distances. L'idée fondamentale du MDS est la définition d'un produit à partir de la distance entre les vecteurs.

Cette définition nécessite de centrer ces vecteurs [4] : l'expression $x_i \cdot x_j$ dépend non seulement des distances 2-à-2 d_{ij}^2, d_{ki}^2, et d_{kj}^2 mais de toutes les autres distances entre points :

$$x_i \cdot x_j = -\frac{1}{2}(d_{ij}^2 - \frac{1}{n}\sum_{k=1}^{n} d_{kj}^2 - \frac{1}{n}\sum_{l=1}^{n} d_{il}^2 + \frac{1}{n^2}\sum_{k=1}^{n}\sum_{l=1}^{n} d_{kl}^2) \tag{3}$$

L'étude des vecteurs et valeurs propres de la matrice des produits scalaires permet une réduction de la dimension de l'espace d'observations.

Isomap construit un graphe dont les sommets sont les points et les arêtes les distances entre eux. Un sommet est adjacent à un autre seulement s'ils sont proches. On estime la distance géodésique entre chaque paire de données par la distance la plus courte parcourue sur le graphe (algorithme de Floyd ou Djikstra). On applique MDS à partir de ces distances géodésiques pour obtenir le nouveau système de coordonnées.

2.3. Locally Linear Embedding

L'algorithme LLE [10] modélise une variété comme une union de petits espaces linéaires. Il exploite la géométrie locale des points x_i dans l'espace original pour la reproduire dans un espace de plus faible dimension. Chaque point x_i a un voisinage $N(i)$ et l'idée consiste à exprimer x_i comme une combinaison linéaire de ses voisins $N(i)$ et de construire son image dans le nouvel espace y_i en respectant cette relation.

Les combinaisons linéaires sont obtenues en minimisant l'erreur quadratique globale :

$$\sum_i \| x_i - \sum_{j \in N(i)} W_{ij} x_j \|^2 \tag{4}$$

sous les contraintes $\sum_{j \in N(i)} W_{ij} = 1, \forall i$. Ces contraintes assurent l'invariance par translation des points et de ses voisins. Une procédure de résolution consiste à utiliser le Lagrangien en décomposant le problème en N sous problèmes.

On cherche alors un espace Y de dimension d' ($d' \ll d$) et un ensemble de points $y_{i,\, i=1\ldots N}$, $y_i \in Y$ tels que l'équation suivante soit minimale, W_{ij} étant donné :

$$\sum_i \| y_i - \sum_{j \in N(i)} W_{ij} y_j \|^2 \tag{5}$$

Y doit pouvoir être translaté sans affecter la fonction de coût, donc $\sum_i y_i = 0$. Pour éviter des solutions dégénérées, la covariance de Y est diagonale et $\frac{1}{N}\sum_i y_i y_i' = I$. Cette contrainte lie toutes les variables et le problème d'optimisation ne peut pas être décomposé pour chaque i comme précédemment.

La solution est de la forme :

$$(I - W)'(I - W)Y = \frac{1}{N} Y \Lambda \tag{6}$$

La réduction de la dimension à d' est reliée aux d' plus petites valeurs propres non nulles de $(I - W)'(I - W)$.

2.4. Spectral clustering

Les résultats du regroupement de données basé sur une décomposition spectrale sont proches de ceux subjectivement perçus les humains. Contrairement à l'algorithme des k-means, ils sont capables de trouver des classes de structure non convexe. Ces méthodes utilisent les vecteurs propres d'une matrice dérivée de la distance entre les vecteurs $x_i,\ i = 1 \ldots N$ pour déterminer les groupes.

L'algorithme « spectral clustering » proposé par [8] est une approximation de la solution au problème (NP-complet) de la séparation d'un graphe en k-groupes. À partir d'une matrice d'affinité A non négative et symétrique qui représente les distances entre points, est définie une matrice D diagonale dont la valeur (i, i) est la somme de la ligne i de A.

Les vecteurs propres, associés aux plus grandes valeurs propres du Laplacien $L = D^{-1/2} A D^{-1/2}$, permettent de construire une représentation de faible dimension des points originaux. Une procédure k-means aide à déterminer les classes de données.

2.5. Laplacian Eigenmaps

Cet algorithme [2] est une variante et mélange les méthodes précédentes. Une représentation graphique des données est obtenue en considérant comme nœuds les points x_i et comme poids sur les arêtes, W_{ij}, les distances calculés avec un noyau Gaussien ou un noyau de type les k-plus proches voisins. Si D est la matrice diagonale avec éléments $D_{ii} = \sum_j W_{ij}$, la fonction à minimiser est :

$$\sum_{ij} (y_i - y_j)^2 W_{ij} \tag{7}$$

De manière similaire au LLE, la minimisation est forte sous les contraintes d'une projection centrée, de variance unitaire. La solution est trouvée grâce aux plus faibles valeurs propres.

2.6. Comparaison des méthodes

Les algorithmes des méthodes de réduction de dimension se déroulent selon des schémas comparables : des suites d'optimisations et de décompositions spectrales sont effectuées à chaque fois.

MDS approche une matrice de Gram de produits scalaires. Cette matrice possède les mêmes valeurs propres que celle de la matrice de covariance de l'ACP : les sorties des deux procédures sont équivalentes.

Selon [6, 9], Isomap, LLE, Laplacian eigenmaps et spectral clustering peuvent être considérés comme des instances de Kernel PCA où la matrice de Gram a été calculée à partir des graphes pondérés au lieu d'une fonction prédéfinie. Ces noyaux sont dits « dépendants des données ». Les graphes reflètent les relations de voisinage des données d'entrée.

3. EXPÉRIENCES

Nous avons effectué des expériences de réduction non linéaire de la dimensionalité, de regroupement et de déroulement de variétés sur des séquences de parole dans un esprit de « fouille de données ». On a utilisé des séquences de vecteurs MFCC du corpus OGI multilangues [7]. Seul les extraits de parole spontanée ont été traités (sous corpus « story-bt », segments de 45 secondes).

3.1. Visualisation des variétés

La visualisation des variétés associées aux séquences de parole obtenues par Isomap et LLE permettent de distinguer une distribution et un regroupement selon les unités phonétiques.

Sur les figures 1 et 2, chaque y, image d'une donnée x, est étiqueté manuellement de la façon suivante : #=pause, #bn=bruit de fond, c=silence occlusive, 0=occlusives, F=fricatives, N=nasales, Vx=voyelles.

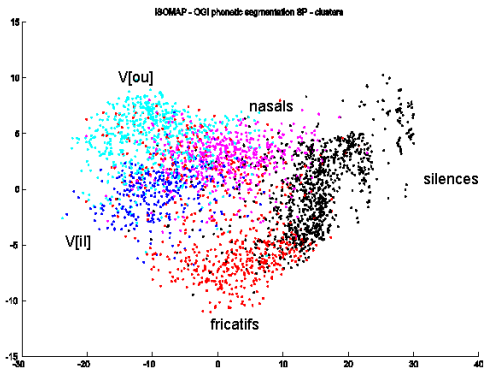

FIG. 1: Représentation d'une séquence de parole obtenue avec l'algorithme Isomap.

Sur la figure 1, on retouve une répartition intéressante des classes. Au milieu, on aperçoit les consonnes avec des regroupements relativement homogènes en silences avant occlusion, nasales, fricatives, occlusives. Et sur la partie gauche de la figure sont regroupées les voyelles.

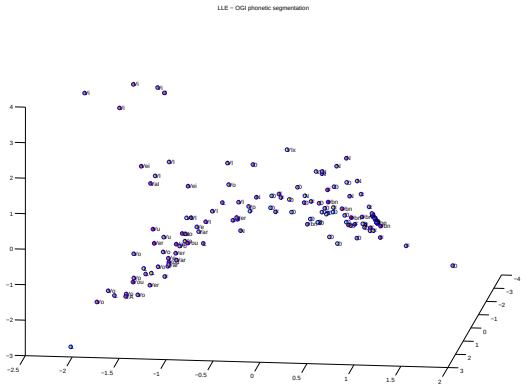

FIG. 2: Représentation d'une séquence de parole obtenue avec la méthode LLE.

Sur la figure 2, les zones de regroupement sont un peu moins homogènes que sur la figure précédente, mais on note tout de même un clivage entre consonnes (à droite) et voyelles (à gauche).

Ces projections permettent d'envisager de réaliser des classifications automatiques. Dans le paragraphe suivant nous allons envisager une telle utilisation.

3.2. Détection automatique de classes

Le deuxième test illustre l'application d'une méthode de classification au sous-espace optenu par un « spectral clustering ». Cette méthode, présentée en [1], permet d'obtenir une classification automatique de régions.

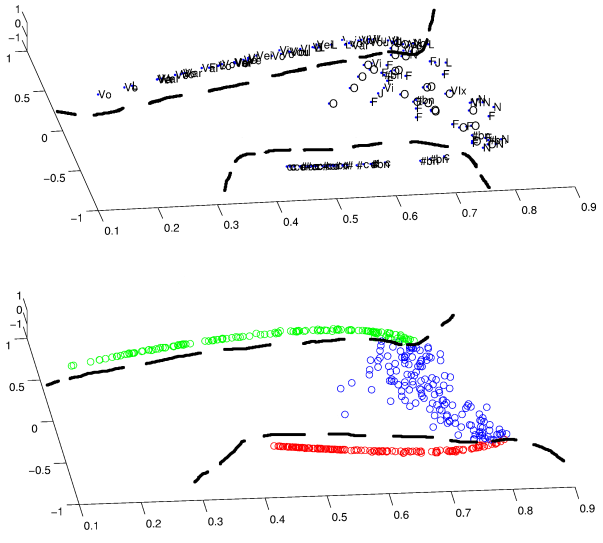

FIG. 3: Détections manuelle et automatique des classes phonétiques par « spectral clustering ». On distingue les trois principales classes phonétiques : silences (en bas), consonnes (au milieu) et voyelles (en haut).

La figure 3, sur la partie haute, est une projection des trois premières dimensions optenu après « spectral clustering », avec affichage des classes avec le même étiquetage que précédemment. La projection en dessous, représente le résultat obtenu après une classification automatique. Les trois figures sont très similaires : la classification automatique nous permet de discriminer les différentes classes.

3.3. Études des variétés

La figure 4 représente les projections sur trois langues de séquences de la parole après réduction non linéaire de la dimension. Seuls les segments consonantiques (détectés automatiquement) sont représentés. Chaque projection laisse apparaitre une forme spécifique à chaque langue. Il est à noter que ces formes apparaissent au bout de quelques secondes : très peu de représentants sont nécessaires pour commencer à les analyser.

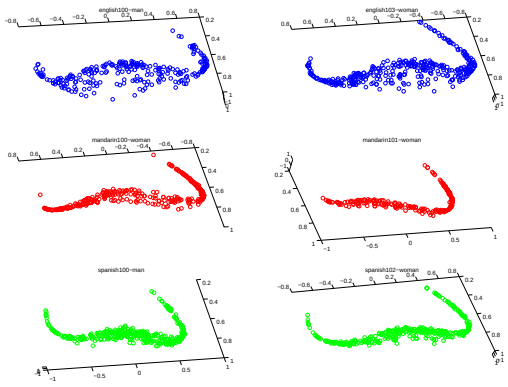

FIG. 4: Répresentation multilangues de séquences de parole par spectral clustering. Les langues étudiées sont l'anglais (en haut), le mandarin (au milieu) et l'espagnol (en bas).

4. CONCLUSIONS ET PERSPECTIVES

Nous avons présenté plusieurs méthodes de réduction non linéaire de séquences de parole, basées sur la décomposition spectrale. Ces méthodes nous ont permis de visualiser en dimension 2 ou 3 l'espace des caractéristiques (section 3.1), en faisant apparaitre des regroupements de grandes classes phonétiques. Dans la section 3.2 nous avons appliqué une classification automatique travaillant dans cet espace projeté, ce qui a permis de les discriminer. La projection de certaines classes phonétiques sur plusieurs langues fait apparaître différentes variétés (formes) associées au processus de production de la parole.

Notre objectif maintenant consiste à découvrir les géométries de ces différentes variétés afin de pouvoir les détecter et les identifier. La modification des contraintes d'optimisation de certaines méthodes, telles LLE ou Isomap, doit nous permettre de reconstituer les formes des distributions des données acoustiques afin de les caractériser et de les comparer.

Nous nous interessons également à la méthode de Nyström [3]. Celle-ci évite de recalculer systématiquement les vecteurs propres des matrices semi-définies positives que l'on a trouvés comme solution avec les méthodes de réduction spectrales non-linéaire en les généralisant aux nouveaux vecteurs d'entrée.

Nous pensons que, appliqués à la parole ces méthodes pourront être utiles pour la caractérisation de certaines modes de production acoustiques, en utilisant des méthodes discriminantes (type SVM) ou génératives (GP) pour des tâches d'indexation sonore, d'identification de langues ou de reconnaissance de locuteur.

RÉFÉRENCES

[1] A. Arias. Unsupervised identification of speech segments using kernel methods for clustering. In *European conference on Speech Communication and Technology*, 2005.

[2] M. Belkin. and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation*, 15(6) :1373–1396, 2003.

[3] Y. Bengio, O. Delalleau, N. Le Roux, and J.F. Paiement. Spectral dimensionality reduction. *Centre interuniversitaire de recherche en analyse des organisations (CIRANO)*, 27, 2004.

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

[5] C. Fowlkes, S. Belongie, F. Chung, and J. Malik. Spectral grouping using the nyström method. In *IEEE Transactions on pattern analysis and machine intelligence*, volume 26, 2004.

[6] J. Ham, D. Lee, S. Mike, , and B. Schölkopf. A kernel view of the dimensionality reduction of manifolds. In *Proceedings of the Twenty First International Conference on Machine Learning*, pages 369–376, 2004.

[7] Yeshwant Kumar Muthusamy, Ronald A. Cole, and B. T. Oshika. The ogi multilanguage telephone speech corpus. In *International Conference on Speech and Language Processing*, volume 2, pages 895–898, October 1992.

[8] A. Ng, M. Jordan, and Y. Weiss. On spectral clustering : Analysis and an algorithm. In *Advances in Neural Information Processing Systems*, volume 13, 2001.

[9] J. Platt. Fast embedding of sparse similarity graphs. *Advances in Neural Information Processing Systems*, 2004.

[10] S.T. Roweis and L.K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(22) :2323–2326, 2000.

[11] B. Schölkopf and A. Smola. *Learning with Kernels*. MIT Press, Cambridge, MA, 2002.

[12] H. Shimodaira, K.I. Noma, M. Nakai, and Sagayama S. Support vector machine with dynamic time-alignment kernel for speech recognition. In *Proc. Interspeech*, 2001.

[13] J.B. Tenenbaum, V. De Silva, and J. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(22) :2319–2322, 2000.