

# Representaciones de secuencias de voz en espacios de baja dimensionalidad

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

SAMOVA - IRIT  
Université Paul Sabatier

XXVII<sup>mes</sup> Journées d'Etude sur la Parole  
Juin 2008



# Plan

- 1 Introduction
  - Estado del arte
  - Notre approche
- 2 Représentation de la distance entre distributions de probabilité
  - Systeme KL
  - Systeme KL-CV
  - Systeme SV

# Plan

- 1 Introduction
  - Estado del arte
  - Notre approche
- 2 Représentation de la distance entre distributions de probabilité
  - Systeme KL
  - Systeme KL-CV
  - Systeme SV

# Traitement automatique de l'information audio

- Les collections de données audio se développent de plus en plus
- La recherche se concentre en :
  - 1 La discrimination parole/musique
  - 2 La classification automatique ou semi-automatique de la musique ou de la vidéo
  - 3 L'identification des langues
  - 4 La vérification des locuteurs
  - 5 L'identification des émotions dans la parole

# Paramétrisation - Classification

- Extraction de paramètres
  - Coefficients cepstraux, analyse de prédiction linéaire, taux de passage par zéro, entropie, flux spectral, modulation de l'énergie dans certaines bandes de fréquence, etc,
- Méthodes de classification
  - Modèles de Markov Cachés (HMM), Modèles de Mélanges de Lois Gaussiennes (GMM), Séparateurs de Vaste Marge (SVM) ou des combinaisons de ces modèles

# Plan

- 1 Introduction
  - Estado del arte
  - Notre approche
- 2 Représentation de la distance entre distributions de probabilité
  - Systeme KL
  - Systeme KL-CV
  - Systeme SV

# Objectifs

- Transformer des séquences de taille variable en vecteurs de taille fixe de faible dimensionalité
- Estimer le nombre de groupes de la base de données
- Appliquer des algorithmes à noyau et de regroupement
- Fixer la paramétrisation. On utilise les coefficients cepstraux (séparation filtre-source ou les premiers coefficients correspondent à l'information relative au seul conduit vocal)

# Algorithmes utilisées

- Algorithme d'échelle multidimensionnelle : À partir des distances entre points, on peut trouver un système de coordonnées qui préserve ces distances
- Regroupement spectral : Approximation du problème de séparation d'un graphe en  $k$ -groupes. Nous l'utilisons pour mettre en évidence le nombre de clusters dans le nouvel espace
- Kernel PCA : Analyse en composantes principales dans le «espace de caractéristiques»

# Corpus

- ANITA (Audio eNhancement In secured Telecom Applications)
- 180 enregistrements monocanaux en studio (haute qualité)
- Exemple : “Je ne sentis ni le coup ni la chute ni rien de ce qui s'ensuivit jusqu'au moment ou je revins a moi”

# Plan

- 1 Introduction
  - Estado del arte
  - Notre approche
- 2 Représentation de la distance entre distributions de probabilité
  - Systeme KL
  - Systeme KL-CV
  - Systeme SV

# Description

## Dissimilarité de Kullback-Leibler entre distributions de probabilité

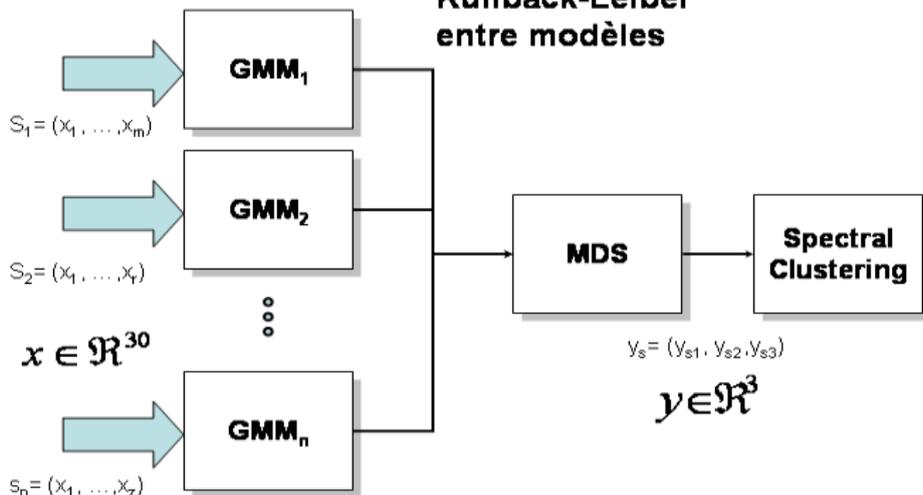
- Chaque séquence  $s_i$  de la base est décrite par un mélange de lois gaussiennes  $GMM_i$
- La dissimilarité statistique  $\delta_{ij}$ ,  $i, j = 1, \dots, N$  entre deux lois  $GMM_i$  et  $GMM_j$  est obtenue avec la divergence symétrique de Kullback-Leibler. Ces dissimilarités sont soumis a l'algorithme MDS
- Le résultat de MDS est un lot de vecteurs de faible dimensionnalité qui représentent les séquences audio de la base de données (un vecteur par séquence)
- Les valeurs propres du regroupement spectral indiquent le nombre de clusters dans l'ensemble
- Kernel PCA,  $k$ -means, ou le regroupement agglomerative peuvent ensuite être utilisés

# Systeme

## Dissimilarité de Kullback-Leibler entre distributions de probabilité

séquences audio

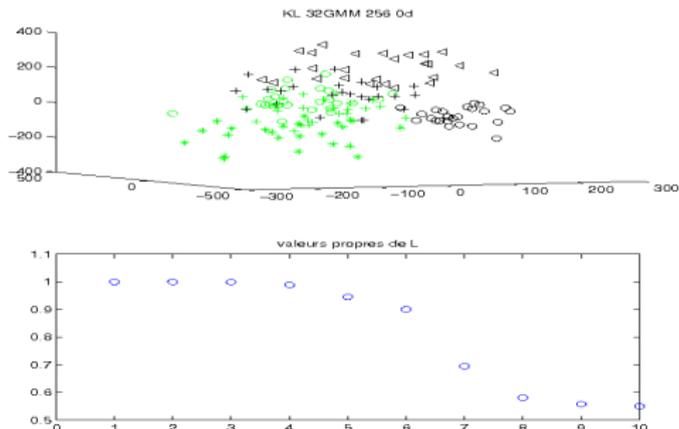
(vecteurs MFCC)



# Résultats

## Dissimilarité de Kullback-Leibler entre distributions de probabilité

- Chaque point représente une séquence de parole et chaque symbole signale un locuteur. Les principales valeurs propres du Laplacien montrent la présence de 6 clusters dans l'ensemble



# Plan

- 1 Introduction
  - Estado del arte
  - Notre approche
- 2 Représentation de la distance entre distributions de probabilité
  - Systeme KL
  - **Systeme KL-CV**
  - Systeme SV

# Description

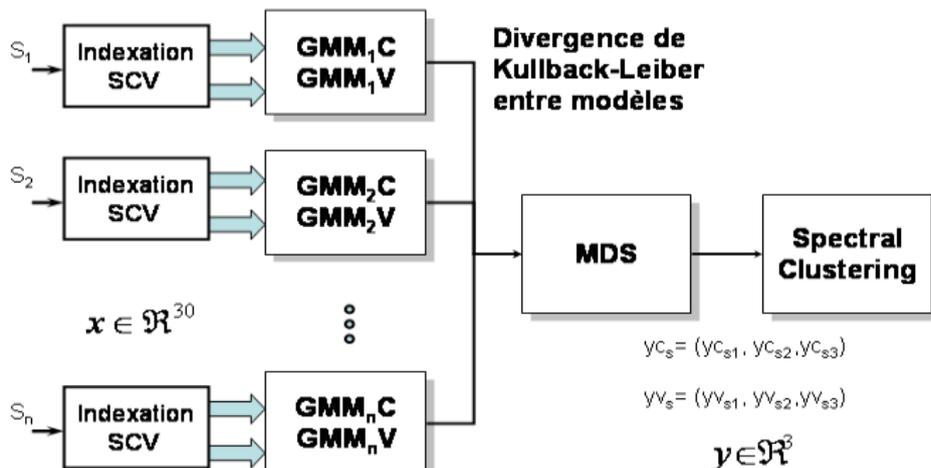
## Modélisation différenciée Consonne - Voyelle

- Une étape de pré-traitement est ajoutée au Systeme KL pour modéliser séparément les unités phonétiques 'vocaliques' (V) et 'consonantiques' (C)
- À partir des unités C et des unités V extraites sur chaque séquence  $s_i$  de parole, sont appris deux modes GMM ( $GMM_iC$ ,  $GMM_iV$ )
- Les calculs de la distance de KL sur chaque sous-ensemble  $GMM - C$  et  $GMM - V$  permettent par la méthode MDS de projeter l'ensemble de séquences dans deux espaces différents  $Y_C$  et  $Y_V$

# Système

## Modélisation différenciée Consonne - Voyelle

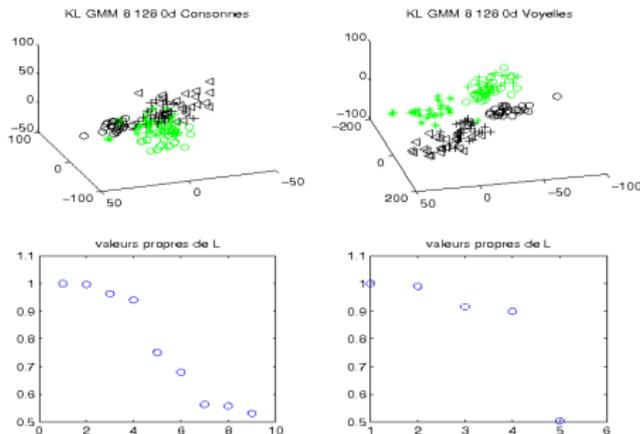
séquences audio  
(vecteurs MFCC)



# Résultats

## Modélisation différenciée Consonne - Voyelle

- On utilise moins de composantes par modèle
- Les résultats de la projection  $Y_V$  montrent une meilleure séparation de locuteurs
- L'eigengap indique 4 clusters



# Plan

- 1 Introduction
  - Estado del arte
  - Notre approche
- 2 Représentation de la distance entre distributions de probabilité
  - Systeme KL
  - Systeme KL-CV
  - Systeme SV

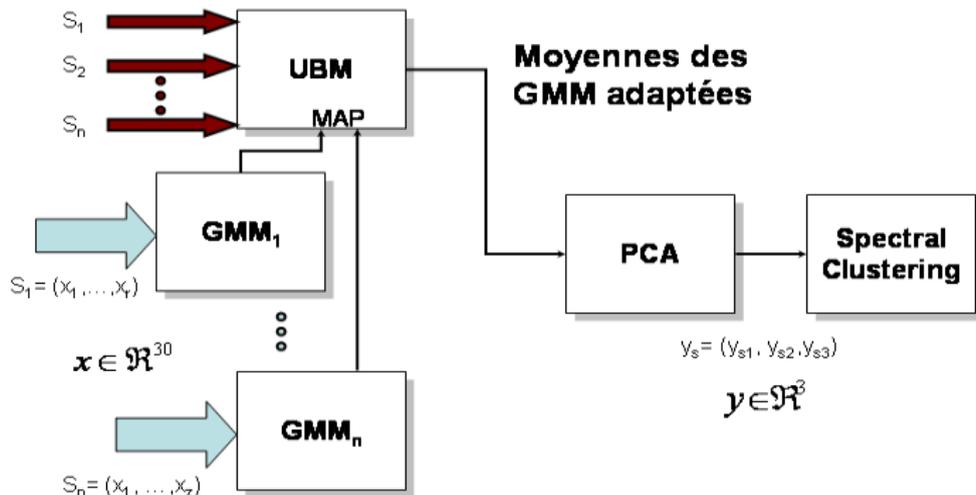
# Description

## Supervecteurs GMM

- Le calcul de la divergence KL est très coûteux en temps d'exécution
- Une alternative d'utilisation des GMM comme représentants de séquences acoustiques est de concaténer les vecteurs moyennes et créer ainsi un "supervecteur" par GMM
- La méthode pour estimer les GMM a été modifiée pour rendre les GMM compatibles entre eux
- Nous utilisons l'adaptation MAP d'un modèle universel à chaque séquence  $s_i$  de parole analysée pour fournir leur modèle  $GMM_i$

# Systeme Supervecteurs GMM

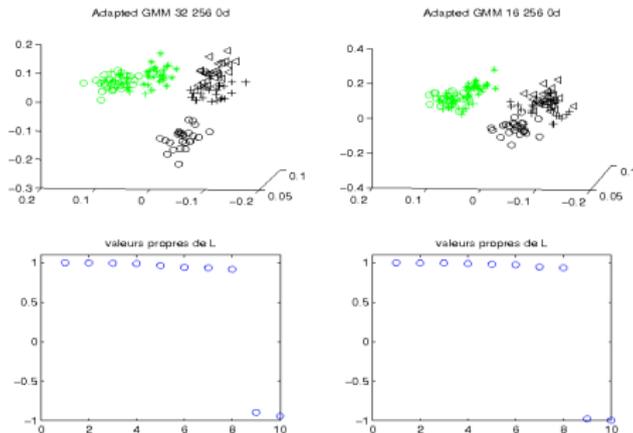
**séquences audio  
(vecteurs MFCC)**



# Résultats

## Supervecteurs GMM

- La projection montre une bonne séparation de locuteurs, proche de celle obtenue avec le Systeme KL-CV(*GMMV*)
- L'eigengap indique 8 clusters



# Conclusions

- Nous présentons diverses possibilités pour visualiser des séquences de parole en espaces 3D
- On propose l'étude des valeurs propres d'une matrice Laplacienne pour identifier le nombre de clusters stables dans l'ensemble
- Les projections  $Y$  issues du Systeme SV et du Systeme KL-CV( $GMMV$ ) sont les plus appropriées pour l'identification des clusters
- Perspectives
  - Utilisation des methodes a noyau
  - Application d'autres bases de données (parole/musique, langues, émotions)

# Bibliographie

-  C. Bishop  
*Pattern Recognition And Machine Learning.*  
Springer, 2006.
-  I. Borg and P. Groenen  
*Modern Multidimensional Scaling: Theory and Applications.*  
Springer, 1997.
-  A. Ng and M. Jordan and Y. Weiss  
On Spectral Clustering: Analysis and an algorithm  
*Advances in Neural Information Processing Systems*, 2001.
-  P. Knees and M. Schedl and T. Pohle and G. Widmer  
Exploring music collections in virtual landscapes  
*IEEE Multimedia*, 2007.