Présentation de l'article

"Speaker Verification Using Sequence Discriminant Support Vector Machines"

de V. Wan et S. Renals, IEEE Trans. on Speech and Audio Processing, Vol. 13, No. 2, Mars 2005

Manuel Davy

21 octobre 2005

CNRS/LAGIS – Lille

Laboratoire d'Automatique, de Génie Informatique et de Signal, Lille

Manuel.Davy@ec-lille.fr

Vérification de locuteurs

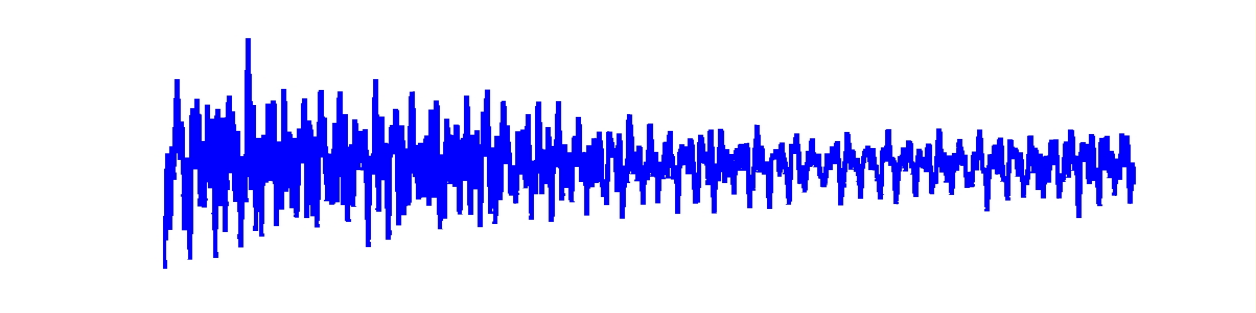
– Le contexte : Pour des raisons de sécurisation, on souhaite vérifier l'indentité d'un interlocuteur par sa voix.

Vérification de locuteurs

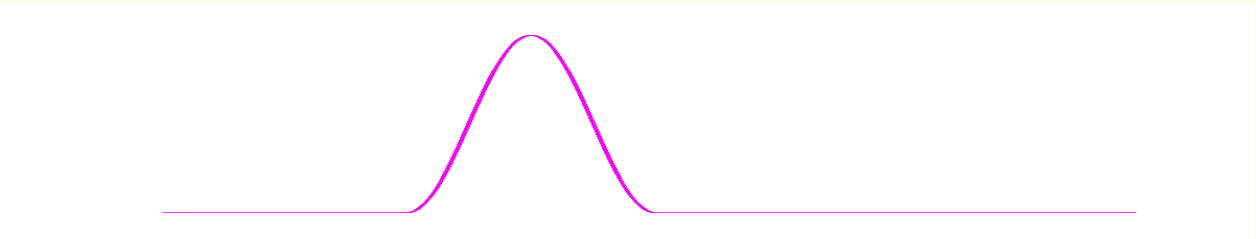
- Le contexte : Pour des raisons de sécurisation, on souhaite vérifier l'indentité d'un interlocuteur par sa voix.
- Exemples : transactions bancaires par téléphone, sécurisation d'accès, etc.
- Contrainte : nécessite d'avoir appris la voix du locuteur auparavant
- Deux approches :
  1. Dépendante du texte : le locuteur est reconnu sur la base d'une phrase (ou de mots) convenus par avance
  2. Indépendante du texte : aucun texte n'est convenu d'avance

L'approche standard (1/4)

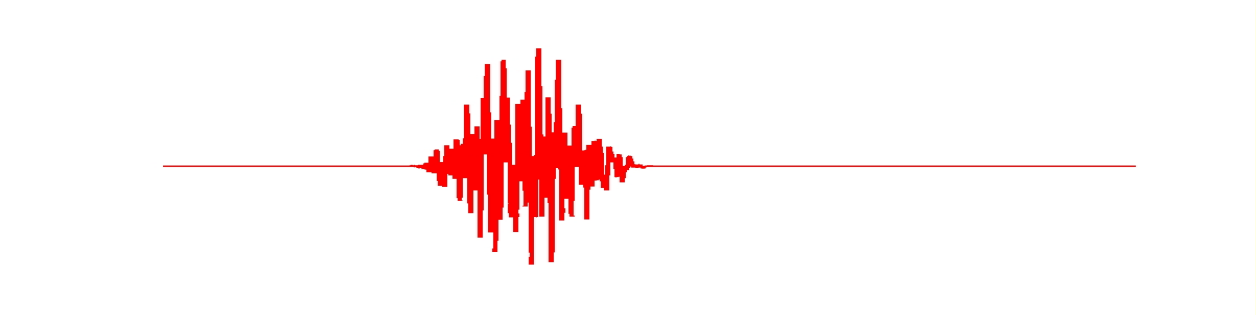
– Analyse du signal par trames $x_n(t) = s(t)w(t - n\Delta_t)$, où n est la position de la trame



$s(t)$



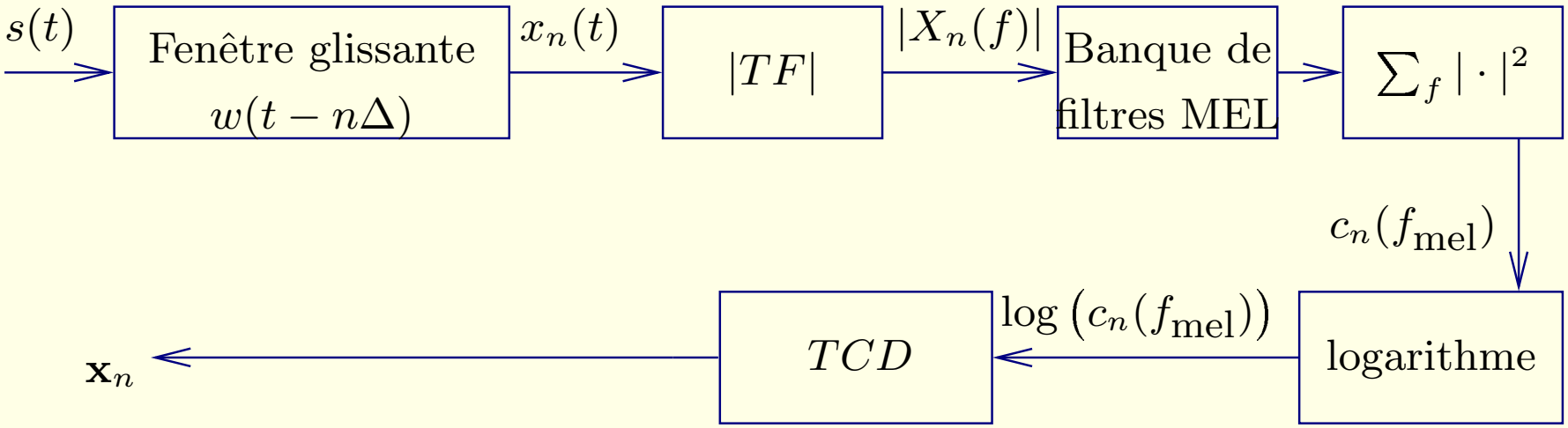
$w(t - n\Delta_t)$



$x_n(t)$

L'approche standard (2/4)

– A partir des trames, on calcule les coefficients mel-cepstraux (*Mel frequency cepstral coefficients*)



– C'est un vecteur de dimension 12 ou 16 à chaque instant n

L'approche standard (3/4)

- Pour vérifier le locuteur, on considère la séquence des MFCCs $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$, comme étant i.i.d.
- Alors, on apprend la distribution les ayant engendrés, notée $p(\mathbf{x}|M, \boldsymbol{\theta}_M)$ où M est le modèle et $\boldsymbol{\theta}_M$ désigne l'ensemble des paramètres du modèle
- L'approche standard considère des modèles de mélanges de Gaussiennes (GMM)

$$p(\mathbf{x}_n|M, \boldsymbol{\theta}_M) = \sum_{k=1}^{K_M} \beta_{M,k}\, \mathcal{N}\left(\mathbf{x}_n\,;\, \boldsymbol{\mu}_M, \boldsymbol{\Sigma}_M\right) \tag{1}$$

et la log-vraisemblance de toute la séquence $\mathbf{X}$ est

$$\log\left[p(\mathbf{X}|M, \boldsymbol{\theta}_M)\right] = \frac{1}{N}\sum_{n=1}^{N} \log p(\mathbf{x}_n|M, \boldsymbol{\theta}_M) \tag{2}$$

- Apprentissage des paramètres par l'algorithme EM

L'approche standard (4/4)

– Le modèle du locuteur à vérifier étant appris, on définit le score d'une séquence $\mathbf{X}$

$$\begin{aligned} S(\mathbf{X}) &= \frac{1}{N}\sum_{n=1}^{N}\log\frac{p(\mathbf{x}_n|M,\boldsymbol{\theta}_M)}{p(\mathbf{x}_n|\Omega,\boldsymbol{\theta}_\Omega)} \\ &= \log p(\mathbf{X}|M,\boldsymbol{\theta}_M) - \log p(\mathbf{X}|\Omega,\boldsymbol{\theta}_\Omega) \end{aligned} \tag{3}$$

où Ω est le modèle du monde (appris sur un grand nombre de locuteurs)

Performances

- Cet algorithme donne environ 6% d'erreur
- Idée des auteurs – et de Smith & Gales (NIPS 2002) avant eux : utiliser un algorithme SVM pour effectuer la vérification
- Problème : définir un noyau portant sur des séquences de longueur différentes
- Travail initial par Jaakkola & Haussler (NIPS 1998) : *Fisher score*
- Cet article : propose une généralisation, et se focalise sur le cas des GMM

Plan

– Introduction

1. L'espace des scores
2. Normalisation et définition du noyau
3. Résultats

– Commentaires et discussion

Plan

– Introduction

1. L'espace des scores

2. Normalisation et définition du noyau

3. Résultats

– Commentaires et discussion

Plan

– Introduction

1. L'espace des scores

2. Normalisation et définition du noyau

3. Résultats

– Commentaires et discussion

Plan

– Introduction

1. L'espace des scores

2. Normalisation et définition du noyau

3. Résultats

– Commentaires et discussion

Plan

– Introduction

1. L'espace des scores
2. Normalisation et définition du noyau
3. Résultats

– Commentaires et discussion

Plan

– Introduction

1. L'espace des scores

2. Normalisation et définition du noyau

3. Résultats

– Commentaires et discussion

Construire des scores

- Pour comparer des séquences de longueur différentes, l'idée consiste à utiliser des modèles génératifs
- Etant donné une famille de tels models, on définit

$$\psi_{\mathbf{T},f} \;=\; \mathbf{T}\, f\big(p(\mathbf{X}|M=1,\boldsymbol{\theta}_1),\ldots,p(\mathbf{X}|M=k,\boldsymbol{\theta}_k)\big) \tag{4}$$

où $M = 1, \ldots, k$ est une famille de models, f est une fonction des scores de la famille des modèles génératifs appelée *score argument*, et $\mathbf{T}$ est un opérateur appelé *score operator*

Exemples

1. $M = 1$, $f(u) = \log(u)$ et $\mathbf{T} = \nabla_{\boldsymbol{\theta}}$: Cela définit le score de Fisher. C'est une mesure de l'adéquation des données au modèle, initialement proposé par Jaakkola & Haussler.
2. $M = 1$, $f(u) = \log(u)$ et $\mathbf{T} = \left[\nabla_{\boldsymbol{\theta}} \; ; \; \mathrm{Id}\right]^{\mathsf{T}}$: c'est le score de Fisher augmenté.
3. $M = 2$, $f(u, v) = \log(u/v)$ et $\mathbf{T} = \nabla_{\boldsymbol{\theta}}$: C'est le score de rapport de vraisemblance
4. $M = 2$, $f(u, v) = \log(u/v)$ et $\mathbf{T} = \left[\nabla_{\boldsymbol{\theta}} \; ; \; \mathrm{Id}\right]^{\mathsf{T}}$: C'est le score de rapport de vraisemblance augmenté

Calcul des scores

- Les auteurs dérivent le calcul du score de rapport de vraisemblance, dans le cas où le modèle génératif est un mélange de Gaussiennes
- Cela requiert de dériver la log-vraisemblance par rapport aux poids β_k, aux moyennes $\boldsymbol{\mu}$ et aux covariances $\boldsymbol{\Sigma}$
- Maintenant, les séquences de longueur différentes sont rapportées à des vecteurs score en dimension fixée
- La dimension de l'espace des scores d_S dépend de la dimension des vecteurs $\mathbf{x}_n$ (donc du nombre de MFCCs) et du nombre de Gaussiennes composant le GMM

Plan

– Introduction

1. L'espace des scores

2. Normalisation et définition du noyau

3. Résultats

– Commentaires et discussion

Normalisation

Pour calculer le noyau

- Les auteurs proposent de normaliser les scores
- Deux techniques : Blanchir les scores ou les projeter sur une sphère

Blanchir les scores

– Cela consiste à calculer le noyau

$$K(\mathbf{X}_1, \mathbf{X}_2) \;=\; \psi(\mathbf{X}_1)^{\mathsf{T}} \mathbf{V}^{-1} \psi(\mathbf{X}_2) \;\triangleq\; \langle \psi(\mathbf{X}_1)\,,\, \psi(\mathbf{X}_2) \rangle_{\mathbf{V}} \tag{5}$$

où $\mathbf{V}$ est la matrice de covariance des scores, i.e., la matrice d'information de Fisher

$$\mathbf{V} \;=\; \mathbb{E}\left\{ \big[\psi(X) - \mathbb{E}\big(\psi(X)\big)\big] \big[\psi(X) - \mathbb{E}(\psi(X))\big]^{\mathsf{T}} \right\} \tag{6}$$

– Attention : Dans certains cas l'espace des scores est de grande dimension, ne permettant pas d'estimer correctement $\mathbf{V}$. Les auteurs proposent d'estimer $\mathbf{V}$ diagonale.

Utilisation de ce noyau

- Selon les auteurs, blanchir les scores ne permet pas de s'affranchir du mauvais conditionnement de la matrice noyau, surtout quand le nombre de Gaussiennes dans le GMM est grand (comparé au nombre de données).
- Cela s'explique, selon eux, par le fait que le noyau défini ci-dessus peut avoir des écarts de valeurs importants.
- Idée des auteurs : projeter les données sur une sphère de rayon 1

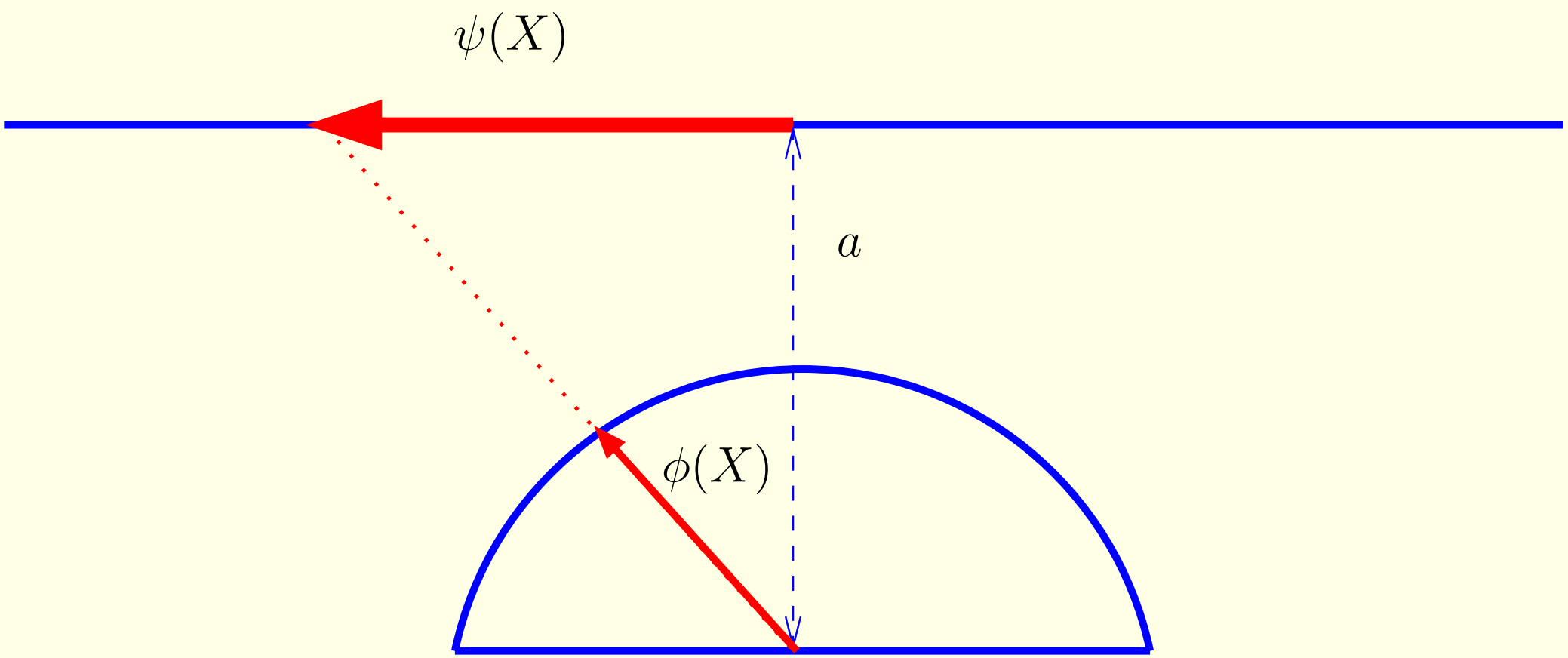
Normalisation

Projeter sur une sphère

- En projetant sur une sphère de rayon 1, on s'assure que $-1 \leq K(\mathbf{X}_1, \mathbf{X}_2) \leq 1$.
- Ici, cela consiste à projeter les vecteurs score sur une sphère de rayon 1, située dans un espace de dimension $d_S + 1$ (projeter sur une sphère en dimension d_S conduit à une perte d'information)
- Les données étant blanchies,

$$K(\mathbf{X}_1, \mathbf{X}_2) = \frac{\langle \psi(\mathbf{X}_1), \psi(\mathbf{X}_2) \rangle_{\mathbf{V}} + a^2}{\sqrt{\left(\|\psi(\mathbf{X}_1)\|_{\mathbf{V}}^2 + a^2\right)\left(\|\psi(\mathbf{X}_2)\|_{\mathbf{V}}^2 + a^2\right)}} \tag{7}$$

Projeter sur une sphère



Plan

– Introduction

1. L'espace des scores

2. Normalisation et définition du noyau

3. Résultats

– Commentaires et discussion

Résultats

Contexte

- Base de données PolyVar (conversation téléphoniques) – 38 clients connus (85 enregistrements par client), 962 enregistrements d'imposteurs
- Le modèle de chaque locuteur connu comporte 200 gaussiennes
- Le modèle du monde comporte 1000 gaussiennes
- Espace des scores de dimension $d_S = 94800$
- Résultats exprimés en termes de EER (equal error rate) (pourcentage d'erreurs lorque le taux de fausses alarmes égale le taux de détections ratées)

Résultats obtenus par les auteurs

GMM	12,07%
GMM-LR (avec modèle du monde	6,12%
Noyau de Fisher	6,87 %
Noyau Rapport de Vraisemblance	4,03 %
idem sans projection sur la sphère	5,55%

Plan

– Introduction

1. L'espace des scores
2. Normalisation et définition du noyau
3. Résultats

– Commentaires et discussion

Commentaires

- Amélioration des résultats de $1/3$ (gain très important)
- Pourtant, les auteurs ne se débarassent pas de l'apprentissage d'une densité
- Non prise en compte de la relation temporelle entre les $\mathbf{x}_n$
- Problématique plus générale : utiliser des modèles génératifs dans un cadre discriminatif (autre exemple entrant dans ce cadre : apprentissage génératif bayésien hierarchique)