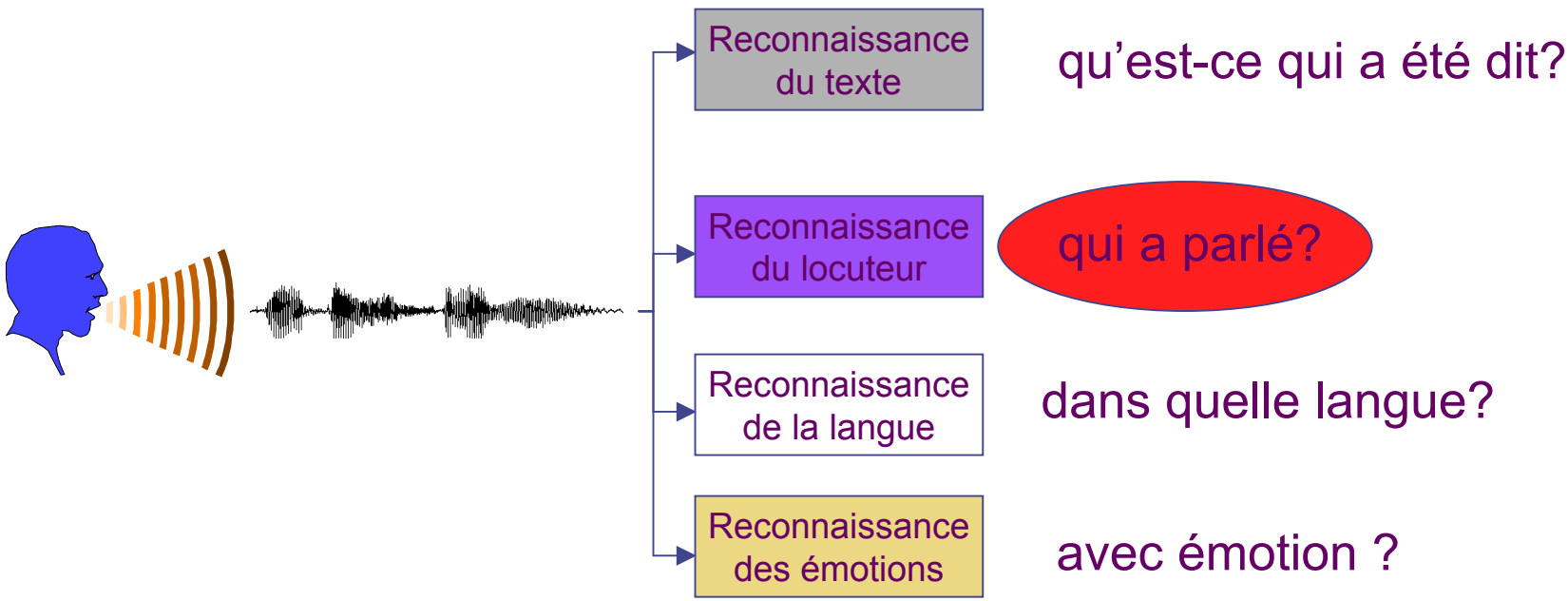
L'authentification biométrique vocale

Jean-François Bonastre

jean-francois.bonastre@lia.univ-avignon.fr
www.lia.univ-avignon.fr
17 Mars 2005

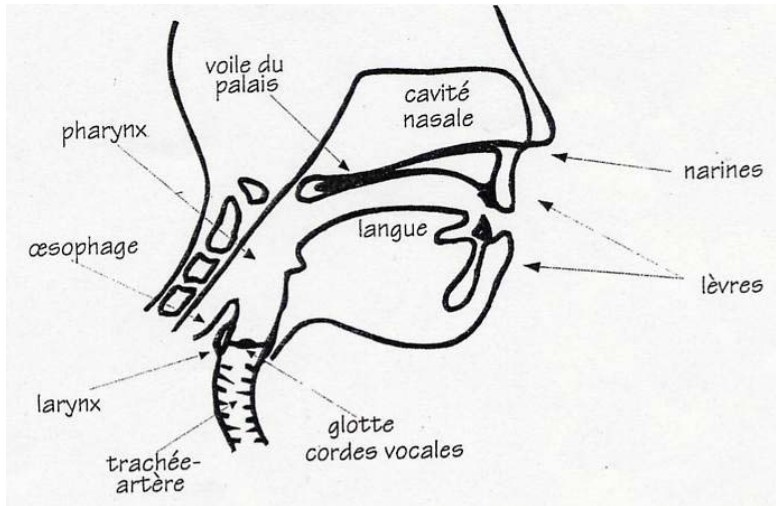
Contexte

A partir d'un signal de parole, des informations de natures très différentes peuvent être extraites:



Production de la parole

- Appareil vocal
  - Poumons et trachée-artère
    - production d'un souffle d'air
  - Larynx
    - vibration des cordes vocales
      Conduit vocal
    - pharynx, cavité buccale,
      cavité nasale
    - organes articulateurs
      - mâchoire, lèvres, langue

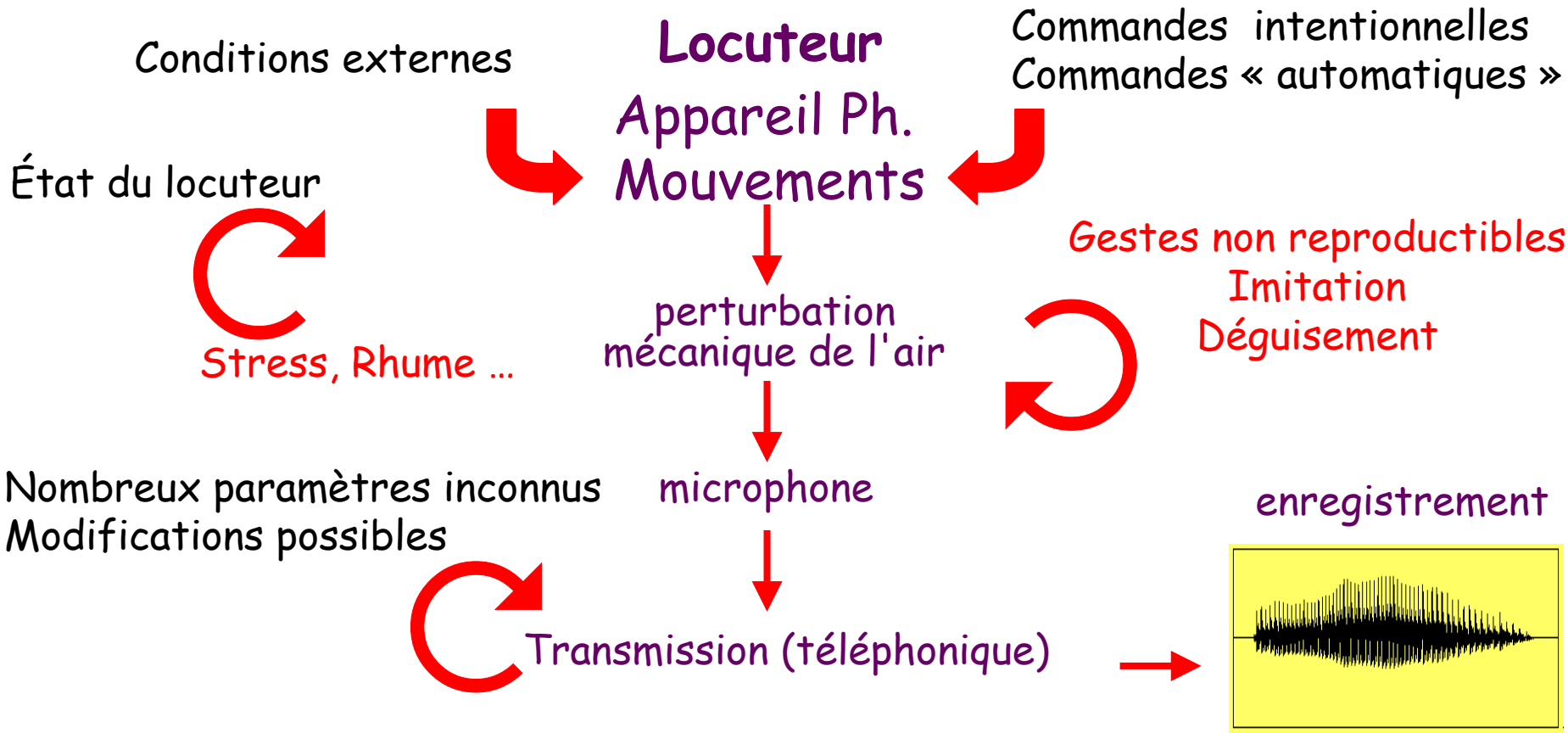


- Sources sonores résonant dans le conduit vocal
  - Vibrations quasi-périodiques des cordes vocales
  - Bruits d'écoulement d'air
  - Occlusions rapides

Caractéristiques du locuteur

- Les humains utilisent différentes sources d'information
- Pas de caractères exclusifs pour l'identité d'un locuteur
- Types d'informations (avec recouvrement)
  - Anatomie de l'appareil phonatoire
  - Prosodie : rythme, vitesse, intonation, volume, modulation
  - Phonétique : cibles phonémiques
  - Accents régionaux
  - Linguistique : syntaxe, grammaire, sémantique
  - Diction, prononciation
  - Emotionnelle, pathologique

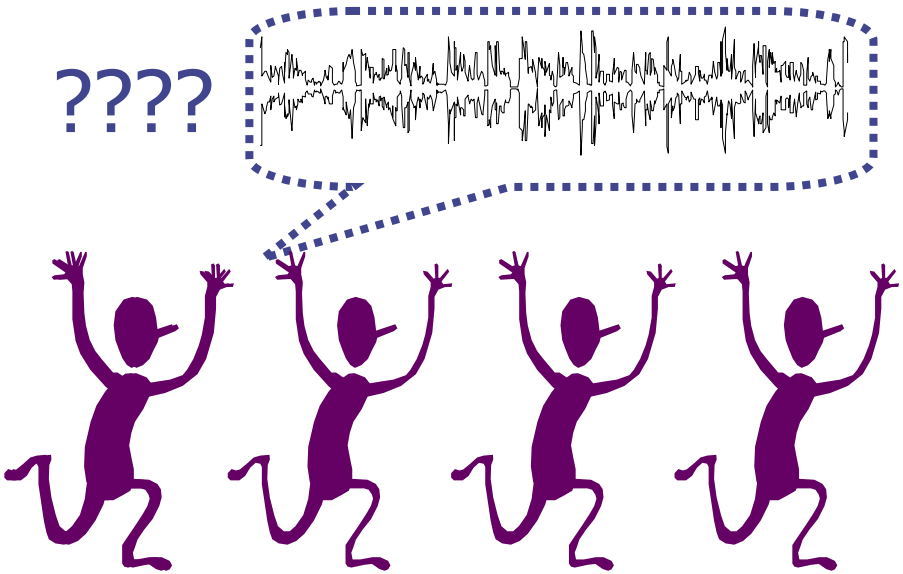
Information captée



Biométrie vocale ?

- La parole est classée
  - Dans la biométrie physique
    - La forme de l'appareil phonatoire est caractéristique de l'individu
    - Cette forme influe sur la production de la parole
    - L'individu ne peut contrôler ces facteurs
  - Dans la biométrie comportementale
    - Car une grande partie de la production de la parole est apprise et non innée (vocabulaire, accent, défauts…)
    - Est-ce contrôlable ?
- Biométrie basée sur des éléments pouvant être modifiés (imitation, déguisement) ou transformés

Les différentes tâches en RAL
Identification et Vérification



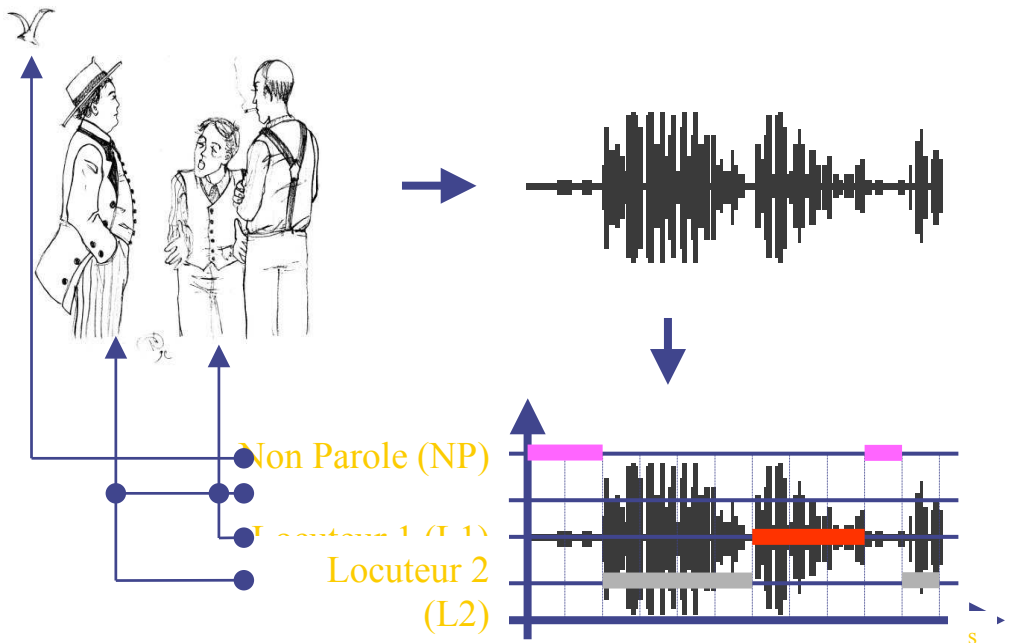
Un message vocal

Identification
« Qui parmi ces personnes
a prononcé le message ? »

- Les locuteurs (client) doivent être connus du système
- Les locuteurs sont coopératifs
- La base peut être fermée (id. = 1 parmi n) ou ouverte
- Le contenu du message peut être contrôlé ou non

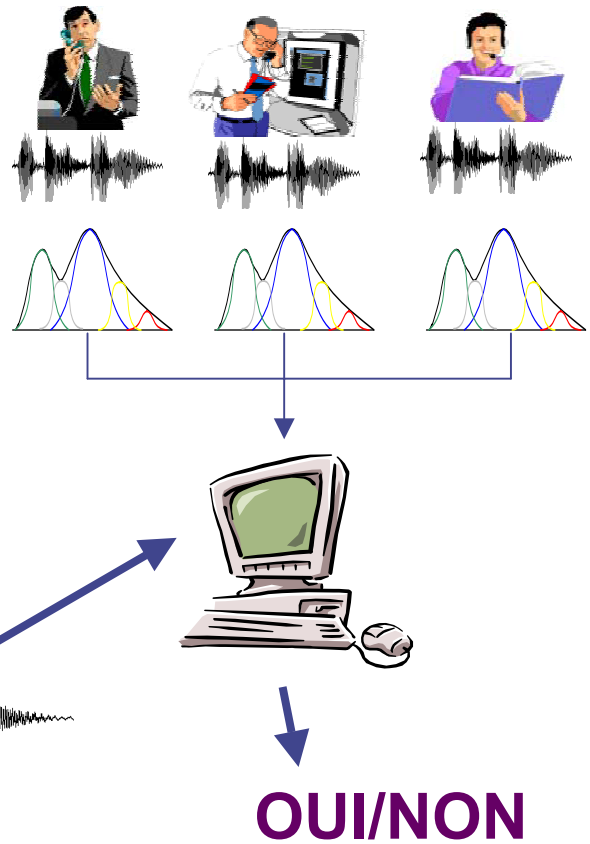
Les différentes tâches en RAL
Segmentation et suivi

- « Qui parle et quand ?»
- Segmentation
  - Pas d'information a priori sur les locuteurs
  - Trouver le nombre de locuteurs
  - Trouver les tours de parole associés à chacun
- Suivi
  - Locuteurs connus a priori
  - Trouver leur tours de parole



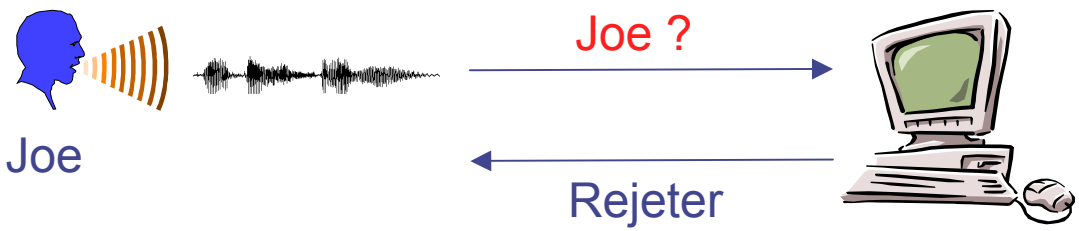
Cadre applicatif

- Généralement Identification + vérification
  - Les clients du système sont connus
  - Mais des imposteurs peuvent usurper une identité

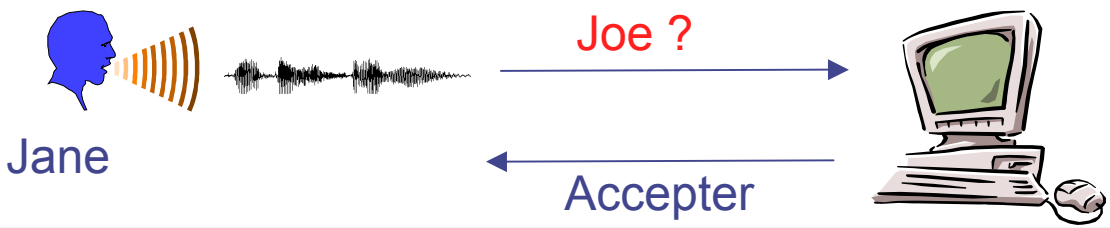


Types d'erreur en vérification
2 types d'erreurs commises par les systèmes

- Le client est rejeté alors que l'identité proposée est la sienne
  - Joe prétend être Joe mais le système d'authentification le rejette
    = **Faux Rejet (FR)** **ou Miss probability**

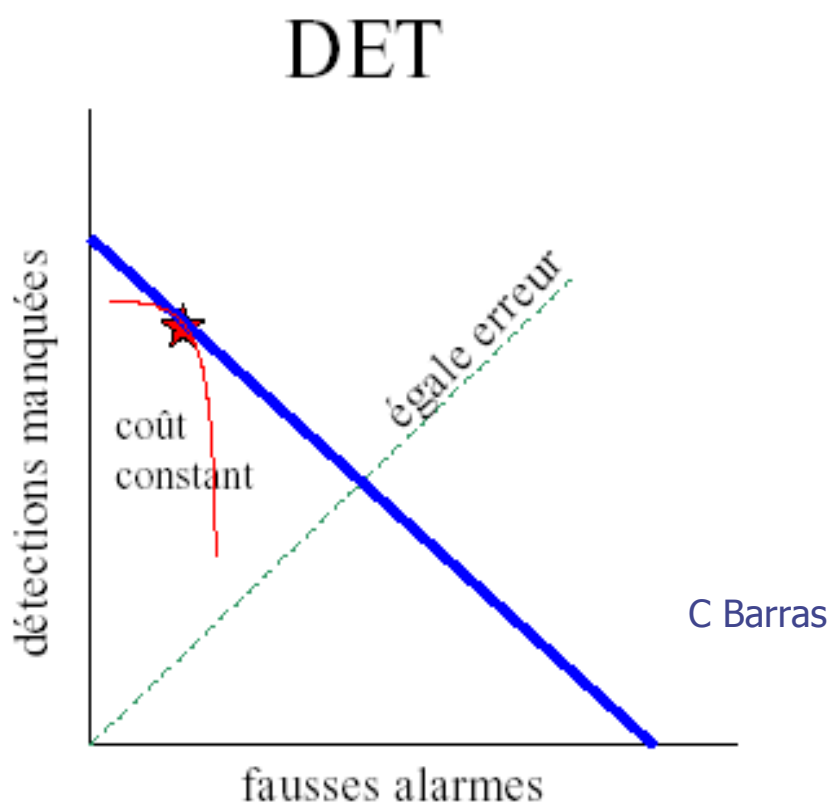


- Le client est accepté alors que l'identité proposée n'est pas la sienne
  - Jane prétend être Joe mais le système d'authentification l'accepte
    = **Fausse Acceptation (FA)**



Représentation des performances

- Identification = % de tests réussis
- Vérification
  - Courbe DET (Martin) = ROC + échelle en déviation de loi normale
  - EER -> FA=FR (ou miss)
  - CDF -> un coût de fonctionnement donné, défini par une fonction de FA et FR



C Barras

Dépendance aux messages

- Systèmes dépendants du texte
  - Messages fixés (mots de passe, uniques ou personnalisés)
  - Messages promptés

  → Meilleures performances
  (dues maj. à la diminution de la variabilité)
- Systèmes indépendants du texte
- Dépendance à la durée des messages
  - *Lors de :*
    - *l'enrôlement des clients (facteur principal)*
    - *des tests*

Dépendance à l'environnement

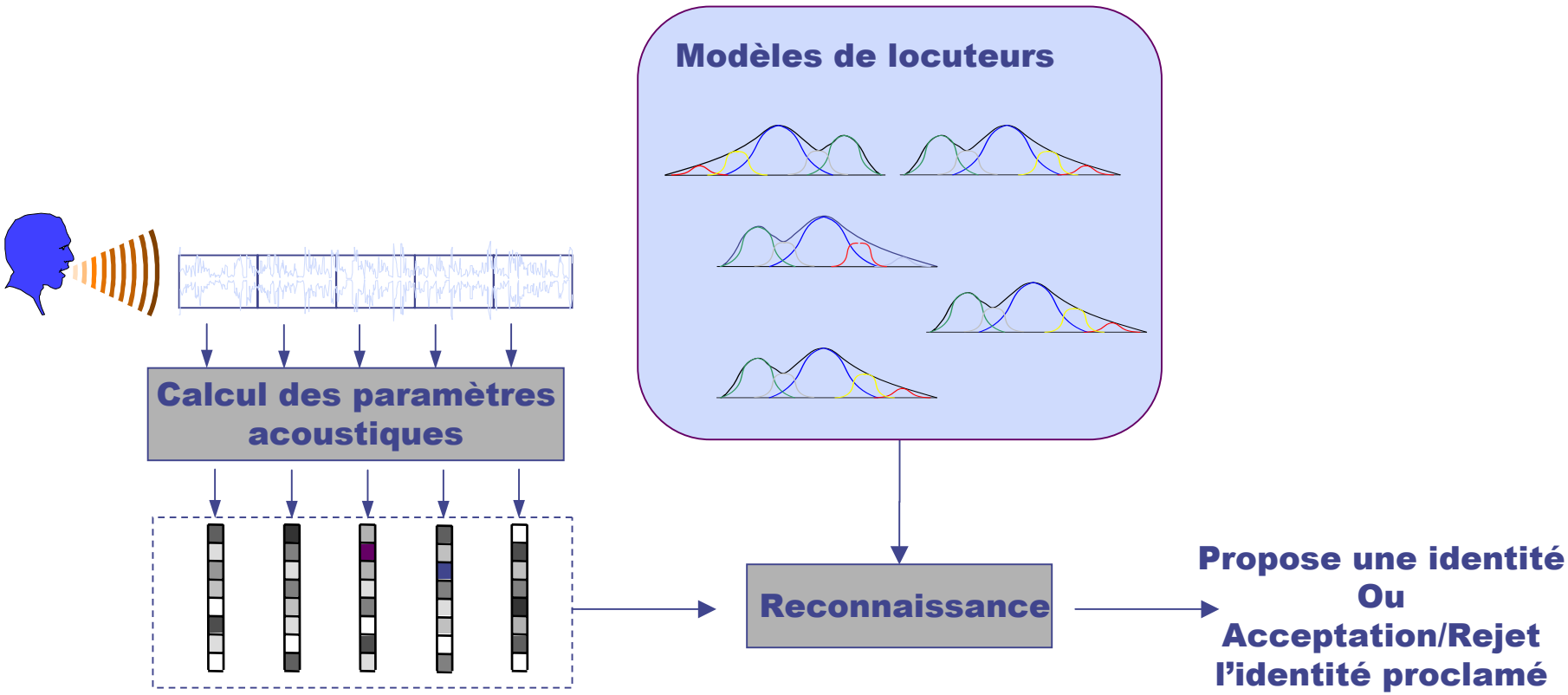
- ☺ Environnement contrôlé :
  - Un local
  - Un micro/un canal
  - Pas de stress

Bonne qualité
et
CONSTANCE

- ☹ Environnement « libre »
  - Multiples lieux
  - Multiples « micros »/canaux
  - Bruits, stress, multiples locuteurs

Qualité pouvant
être mauvaise
et
VARIABLE

Technique d'authentification



Le rapport d'hypothèse Bayésien (1)

- Etant donné un signal de test *X*, et une identité *I*
  - H_0 : *X* a été généré par le locuteur d 'identité *I (OK)*
  - H_1 : *X* a été généré par un locuteur imposteur *(Imposture)*

- Test statistique d 'hypothèses (classique) qui consiste à comparer

$$P_0(X)=P(H_0|X) \text{ et } P_1(X)=P(H_1|X)$$

- En fait, on va comparer les log.

$$\log P_0(X) - \log P_1(X) \begin{array}{l} > \\ < \end{array} \begin{array}{l} accept \\ reject \end{array} Seuil$$

Le rapport d'hypothèse Bayésien (2)

- En log, après Bayes, et en intégrant les apriori dans le seuil

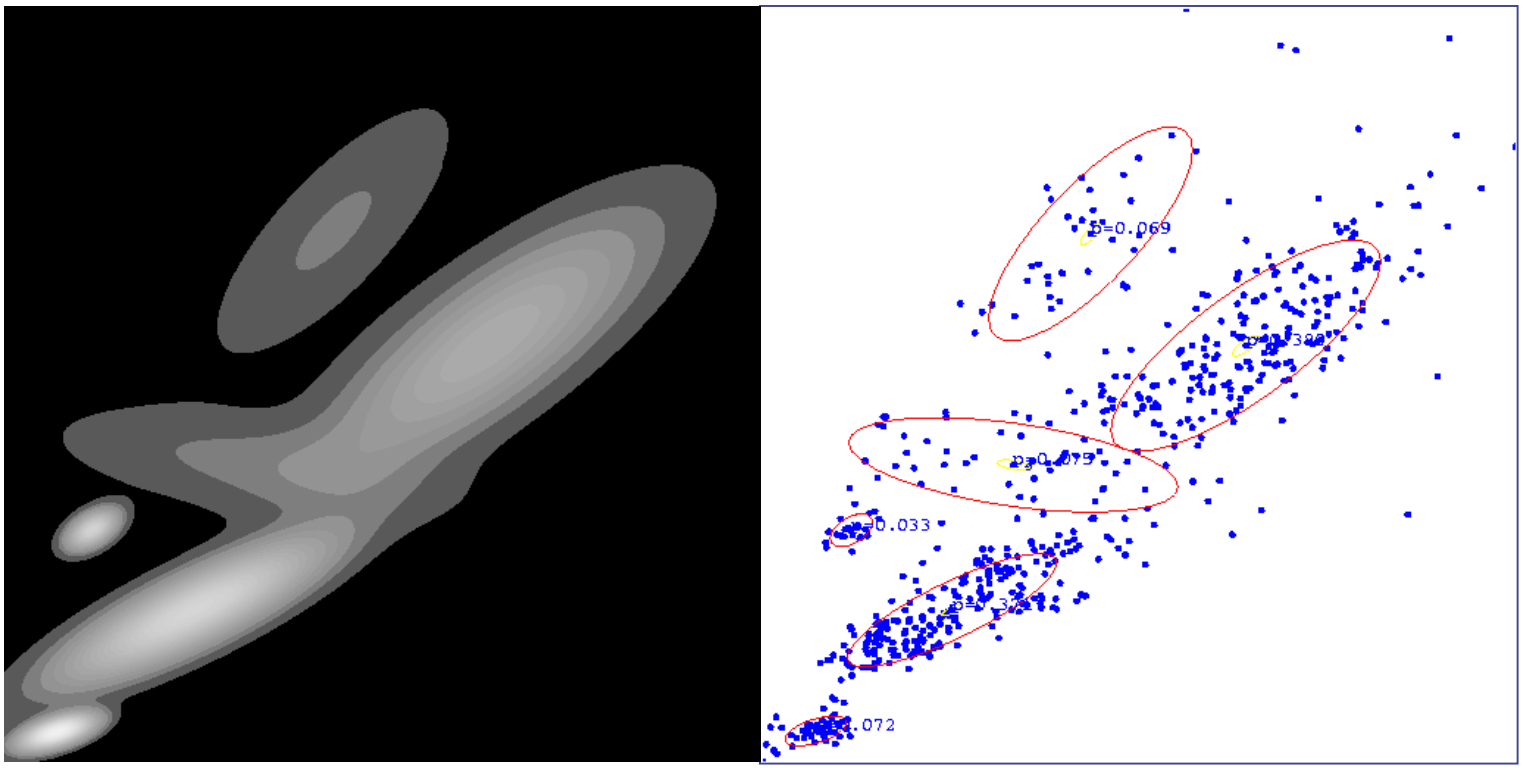
$$\log\left(p(X|H_0)\right) - \log\left(p(X|H_1)\right) \geq Seuil$$

- Modélisation acoustique d'un locuteur client
  - On a des données « représentatives » du client
  - Approche statistique : modèle génératif
- Modélisation de l 'imposture
  - Cohorte : on modélise l'imposture par un ensemble de locuteurs
  - modèle « du monde » : on modélise l'imposture par un modèle génératif appris à partir de données provenant de n locuteurs, en général n'incluant pas les utilisateurs potentiels

Modélisation d'une classe acoustique

- Créer des modèles de H_0 et H_1 à partir de données les représentant
  - Données manquantes -> modèle paramétrique
- Calcul de l'appartenance des données à une classe par le principe de la vraisemblance
- Méthode majoritaire
  - formalisme Markovien
  - Mixture de Gaussiennes (GMM) = 1 état
    - Estimateur de densité de probabilité par une moyenne pondérée de lois gaussiennes

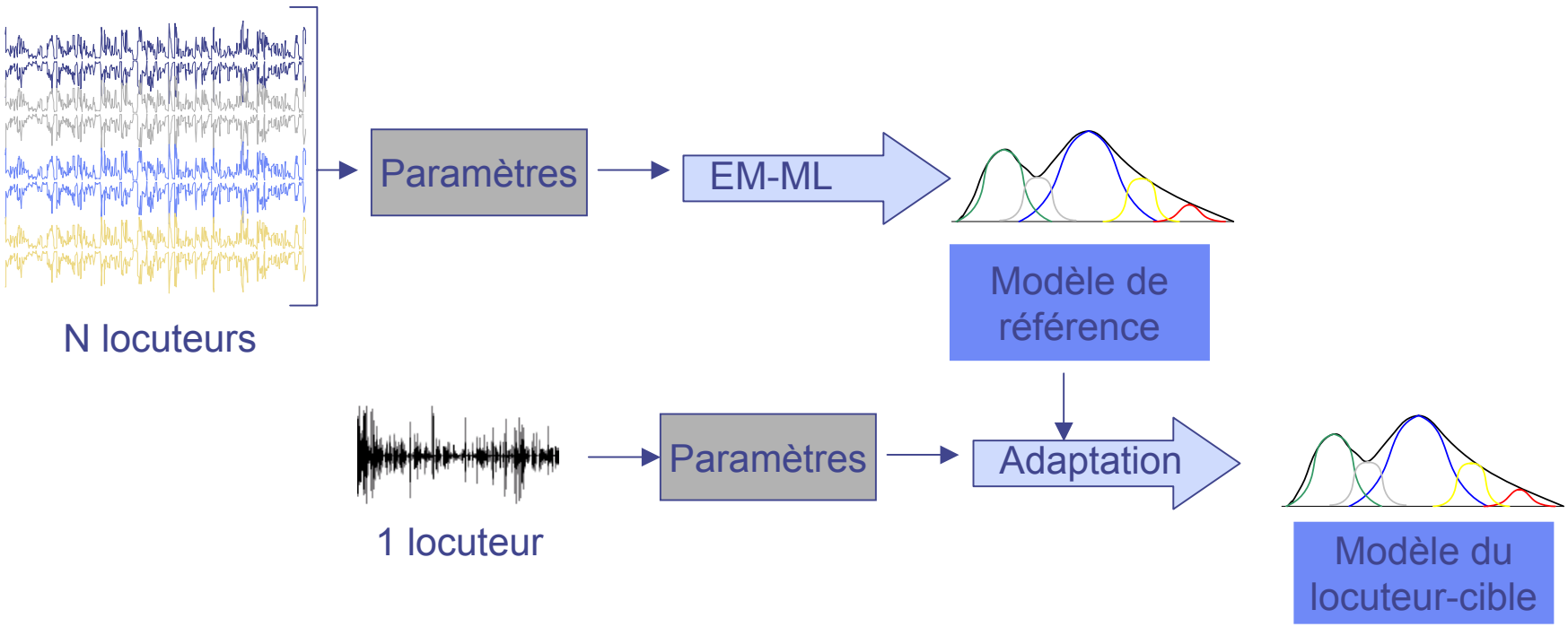
Les mixtures de gaussiennes exemple



Les GMM en Reconnaissance du Locuteur

- Introduit par Reynolds
- Etat de l'art (indépendant du texte)
- Modèles de clients dérivés par MAP
  - D'un modèle générique (world)
  - Uniquement les moyennes
- Pas mal de recettes peu expliquées ou explicables
  - Matrices de covariances diagonales
  - Un grand nombre de composantes (~2000)
  - Initialisation
  - Contrôle de la variance
- Modèle du monde joue un rôle très important !!

Technique d'enrôlement GMM/UBM



Les performances

- Une « référence » actuellement : évaluations NIST
- Bases de données disponibles :
  - Locuteurs américains (hommes & femmes)
  - Conversations téléphoniques réelles
    - coopération implicite
  - Diverses sources de variabilités
  - Tests d'impostures simulés : tests croisés
- Protocoles standards pour la comparaison des systèmes

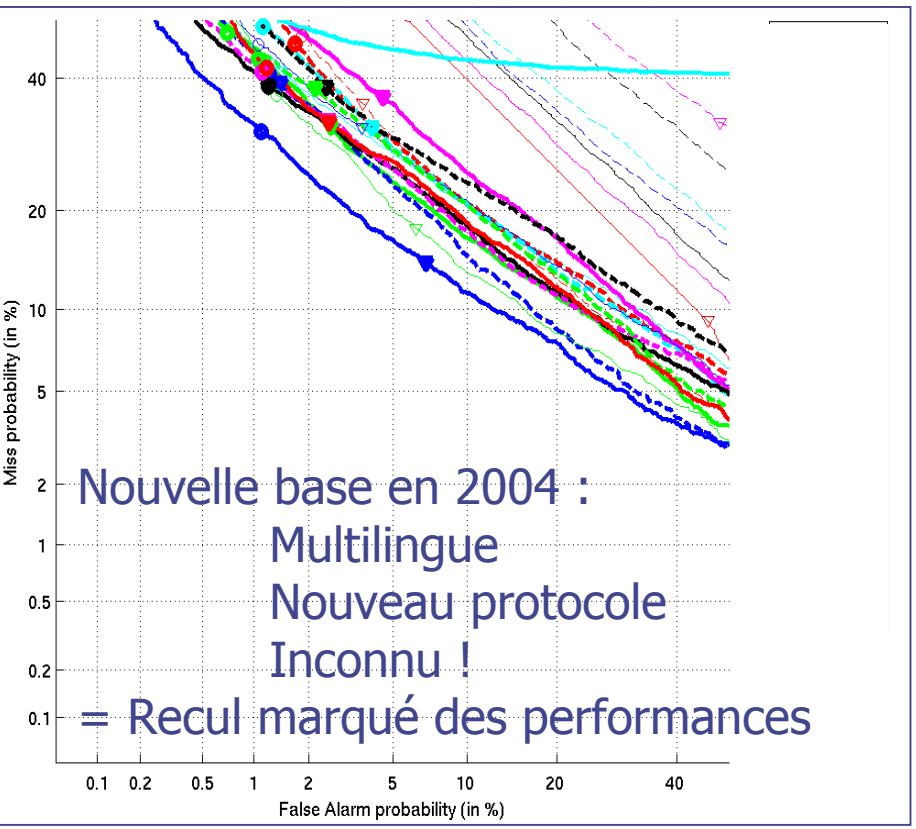
Les performances

- Identification sans rejet :
  - moins de 1% d'erreurs en parole préparée « studio », parmi 630 locuteurs (6s enrol., 3s test)
  - 40% d'erreurs en parole téléphonique spontanée
- Vérification/détection, pourcentage d'égale erreur
  - 0,1% parole propre, prompt fixé
  - 1% parole téléphonique, prompt fixé
  - 10% parole téléphonique spontanée
    - -> 5% NIST 2005
    - Moins de 1% avec env. 30 minutes
  - 25% parole radio bruitée spontanée
- Importance de la durée d'apprentissage et de test

Les évaluations NIST

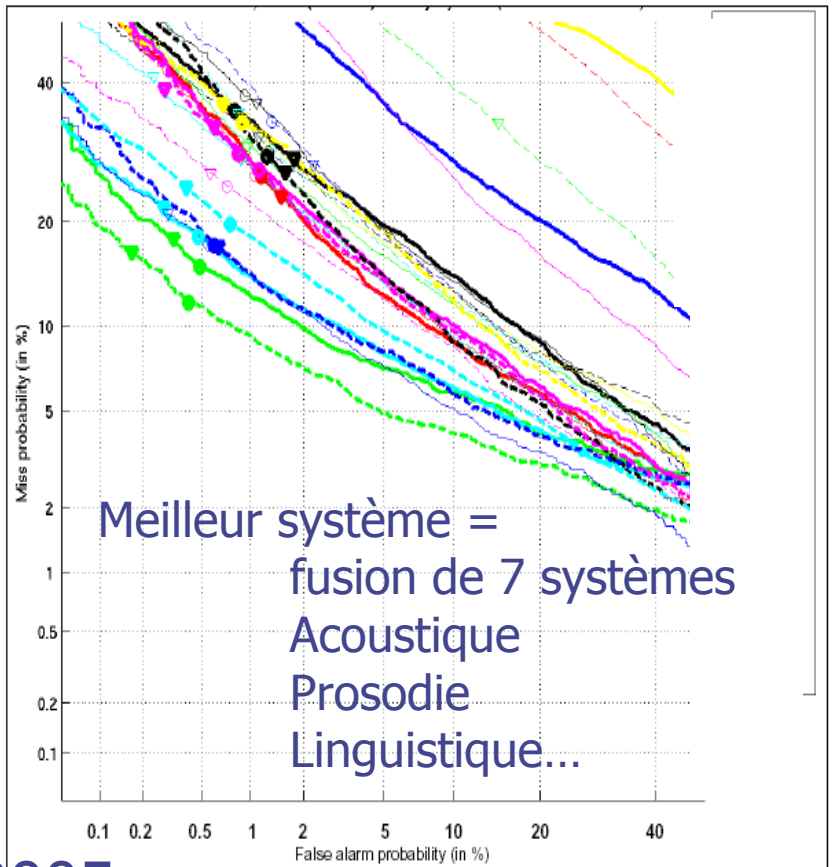
- Depuis 1996
- Evaluation sur de la parole téléphonique
- Tâches orientées par le sponsor….
- Inscription gratuite
  - Donne accès aux données
- Un protocole solide
  - Détection de locuteur
  - Règles précises
  - Evaluation “en aveugle”
- Grand nombre de test
  - Env 5000 clients
  - Env 45000 “imposteurs”
- Tests
  - Sans connaissance des autres clients
  - Sans normaliation inter-tests

NIST04 et NIST05 (one side - one side)



Nouvelle base en 2004 :
Multilingue
Nouveau protocole
Inconnu !
= Recul marqué des performances

Nist 2004 results.v6



Meilleur système =
fusion de 7 systèmes
Acoustique
Prosodie
Linguistique…

2005

NIST_SRE05-a/pdf/nist_nist05.pdf

Interprétation des performances

- Performances d'un système de RAL fortement dépendantes :
  - des locuteurs de la base et de leur nombre
  - des conditions d'enregistrement & canaux de transmission

Résultats peu transposables d'une application à l'autre

- Validité statistique
  « Règle des 30 » = 30 exemples d'une erreur !
  1% EER -> 3 000 tests « client » (par client ?)
  0.001% EER -> 3 000 000 tests « client » !

Erreur d'interprétation des performances

- Exemple : locuteurs cibles Femme + locuteurs imposteur hommes

  → Amélioration des performances !

- Exemple : Base enregistrée avec un combiné/canal & ajout de tests imposteurs réalisés avec un autre combiné/canal

  → Amélioration des performances !

Applications

- Sécurité
  - contrôle d'accès (en complément d'un code, d'un badge)
    - banques, voitures, entreprises…
    - consultation de compte bancaire par téléphone…
- Police criminelle (identification de suspects) ?
  - filtrage de voix suspectes (avec validation humaine)
  - …pas assez fiable pour utiliser comme preuve !
    - Position de l'AFCP
- Transcription automatique
  - adaptation des modèles acoustiques à la voix du locuteur
- Indexation multimédia
  - indexation par locuteur

Utilisation en conditions réelles

- Au niveau théorique :
  - On ne sait pas modéliser correctement le modèle de rejet en vérification (connaissance *a priori* des imposteurs)
  - Approche statistique : que modélise-t-on réellement ?
- Evaluations réalisées :
  - Locuteurs coopératifs ou neutre
  - Résistance aux vraies impostures inconnue
  - Pas encore d'évaluation avec des imitateurs
  - Conditions environnementales connues
- Non transposition des résultats

Autres approches…

- HMM = extension des GMM
  - Peu d'avantages, à part reconnaître le texte
- Classifieurs classiques (Réseaux de neurones, polynomiaux)
- DTW pour des systèmes (très) dépendant du texte
- Support Vector Machines
  - Approche discriminante
  - Complexité dans le noyau (noyau polynomial de Campbell)
  - Assez surprenant (1 vs 1000)
  - Proche des GMM à partir de 2 minutes de test, meilleur ensuite
- En fait, les GMM avec adaptation MAP des moyennes sont discriminants (voir Mariethoz)

Problèmes majeurs

- variabilité due au locuteur
  - émotion, fatigue, stress
- conditions d'enregistrement variables
  - microphone, bruit ambiant
- conditions de transmission variables
  - canal téléphonique
- nouveaux problèmes
  - GSM : codage, bruit évoluant au cours du temps

Le cadre « criminalistique »

Position AFCP & SPLC

J.F. Bonastre, F. Bimbot, L.J. Boe, J. P. Campbell, D. A. Reynolds, I. Magrin-Chagnolleau, *Person Authentication by Voice: A Need for Caution*, 2003 Eurospeech 2003, Genova

J.F. Bonastre, F. Bimbot, L.J. Boe, J. P. Campbell, D. A. Reynolds, I. Magrin-Chagnolleau, Authentification des personnes par leur voix : Un necessaire devoir de précaution, 2004 Journées d'Etude de la Parole, Fèz (Maroc)

UN BESOIN DE PRECAUTION

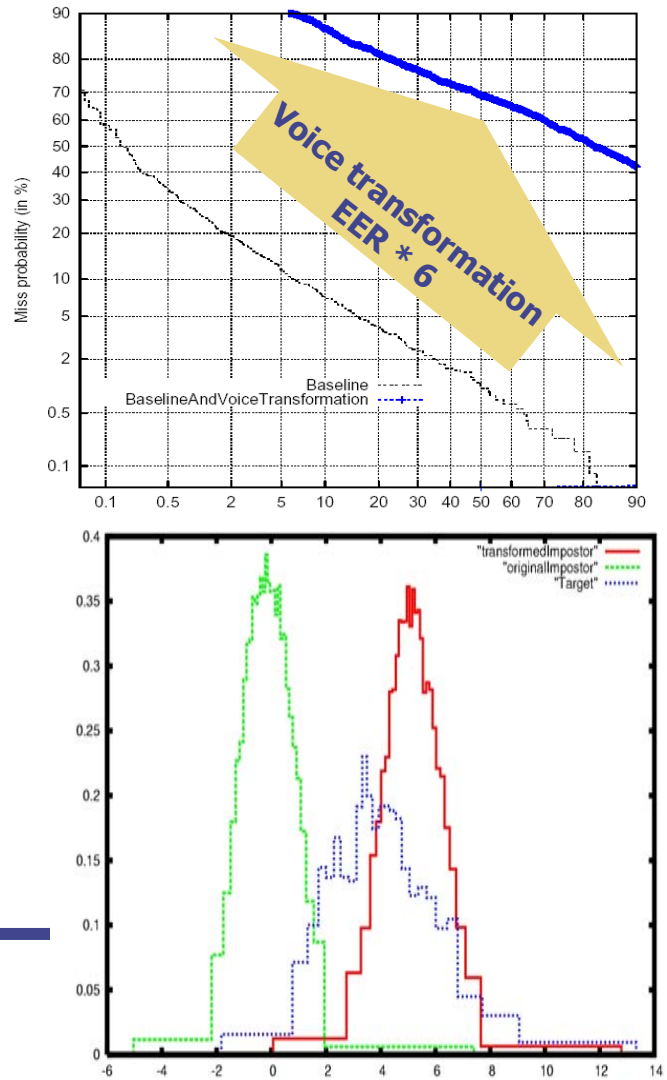
Précaution et réflexion doivent être appliquées lorsqu'un procédé de reconnaissance du locuteur - basé sur une expertise humaine ou automatique - est utilisé, pour tenir compte des facteurs incontrôlables

individu à partir de sa voix

Exemple : Transformation de la voix

- Que se passe t-il si
  - Le système est connu
  - Le client est connu
- Test d'une transformation pour rapprocher les imposteurs des clients AU SENS DU SYSTEME
- Contrainte : La transformation doit rester inaudible

  (COST 275, ICASSP06)



Et après…

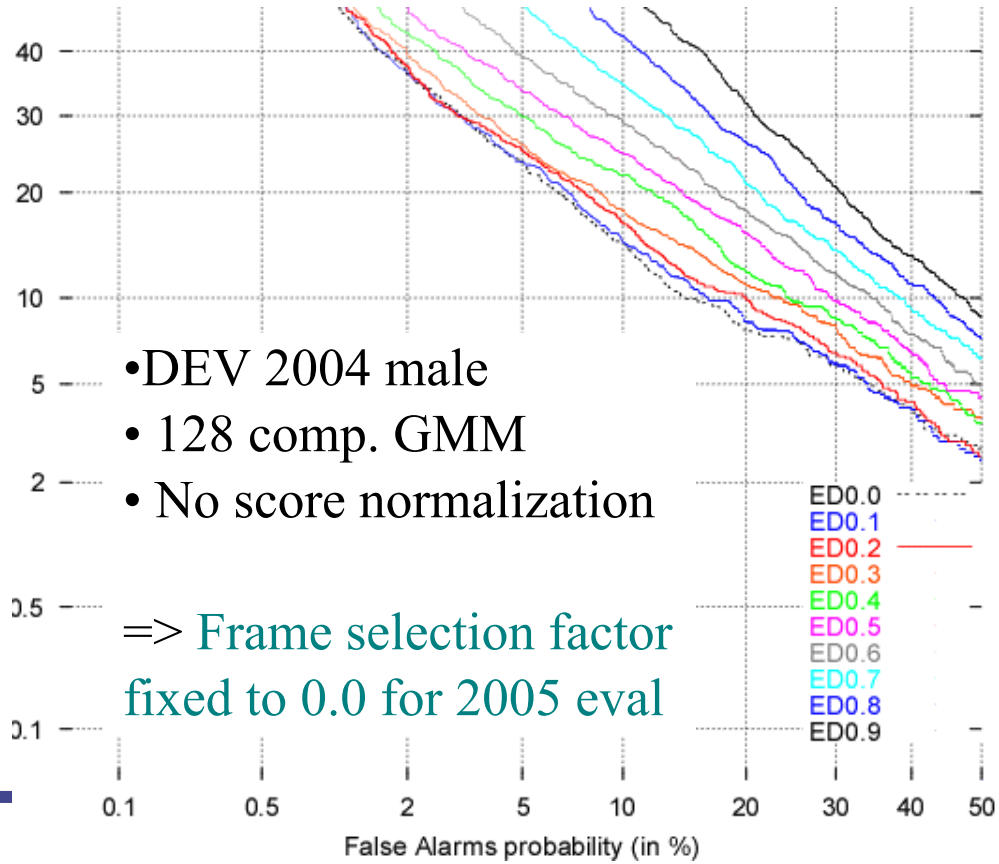
- Autres informations
  - Qualité de la voix (pathologie)
  - Styles de voix (types de parole et émotions)

Conclusion

- Domaine de recherche actif
  - Soutien gouvernemental américain : évaluations NIST, Soutiens à la recherche depuis 1996, Projets…
  - Soutien gouvernemental français :
    - Evaluations ESTER, Projet Technolangue ALIZE
    - Thèses DGA, DGA…
  - Soutien européen, Projets PICASSO/CAVE/BANCA, NOE BIOSECURE, IP BIOSEC

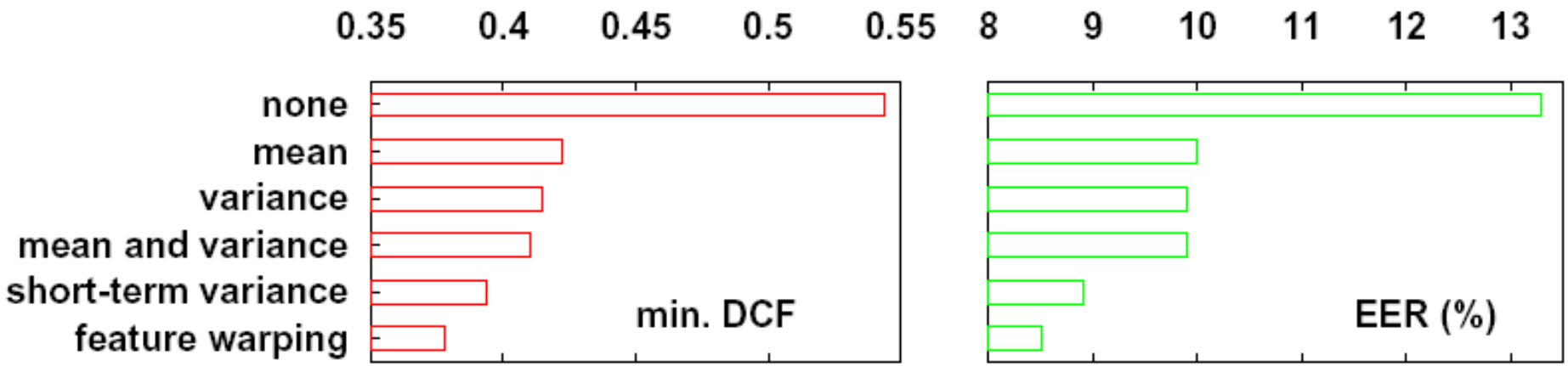
Exemple de points clés : Sélection des données

- Détection basée sur l'énergie- GMM 3 composentes
- Un paramètre de réglage



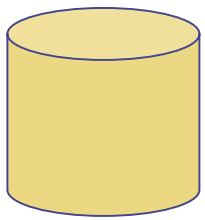
Exemple de points clés : La normalisation des paramètres

- Exemple, LIMSI 2003 (C Barras) – base NIST 2003

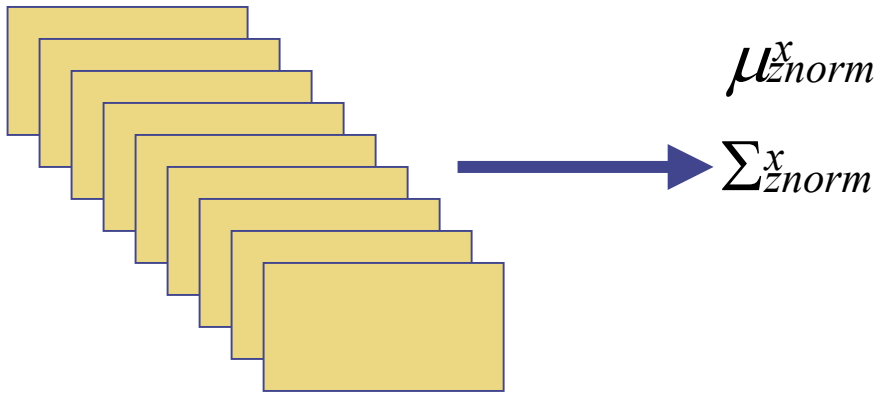


Exemple de points clés : Normalisation des scores

- Znorm



Modèle du client (x)



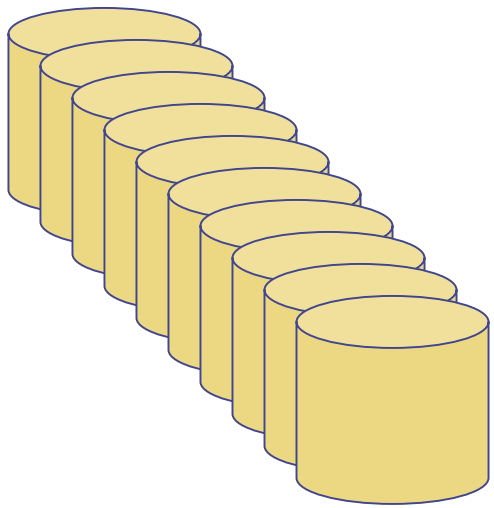
Un ens. de tests « imposteurs » (diff. de x)

$$Score_{znorm} = \frac{Score - \mu^x_{znorm}}{\sum^x_{znorm}}$$

Exemple de points clés : Normalisation des scores

- Tnorm

Un test (y)



μ^{y}_{Tnorm}

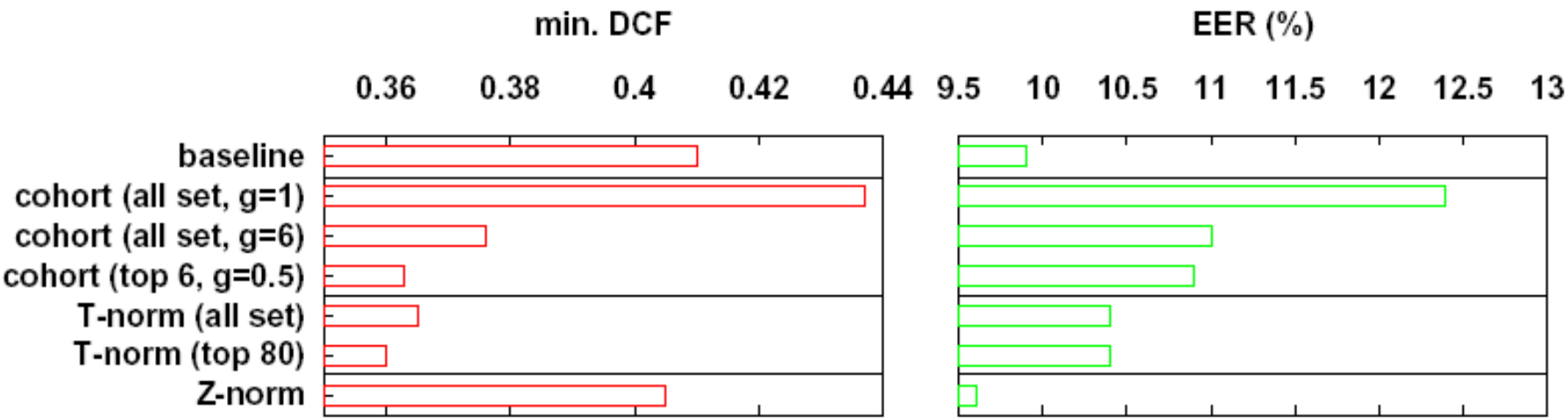
Σ^{y}_{Tnorm}

Un ensemble de modèle « d'imposteurs » (diff. de y)

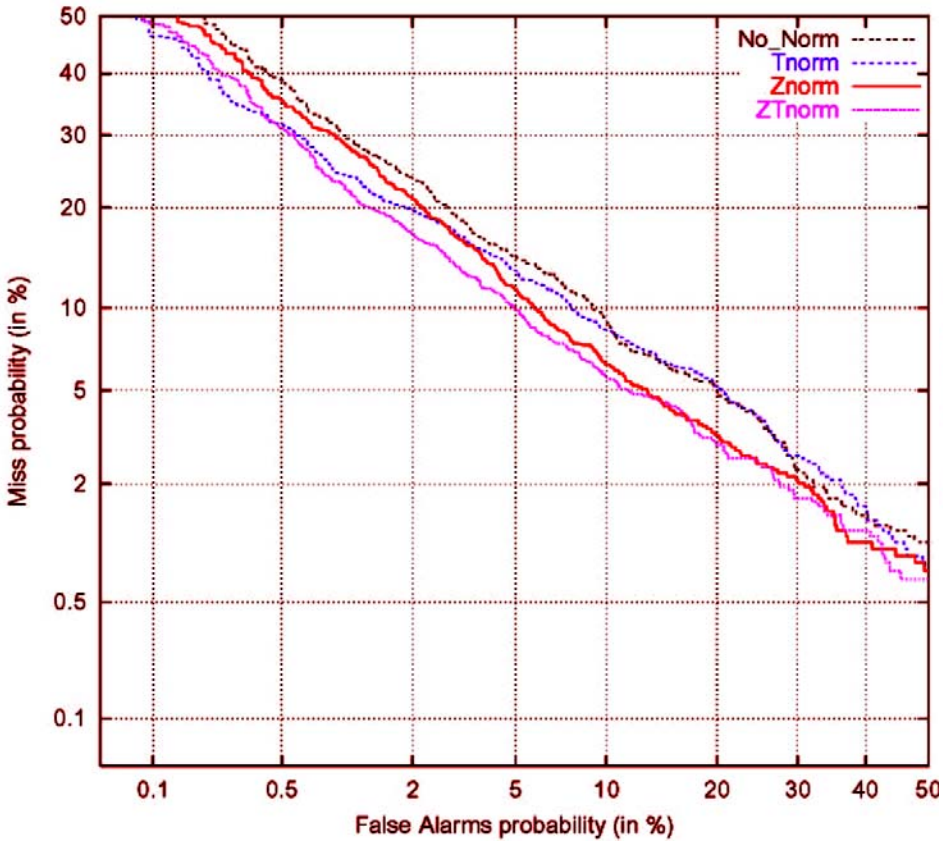
$$Score_{Tnorm} = \frac{Score - \mu^{y}_{Tnorm}}{\Sigma^{y}_{Tnorm}}$$

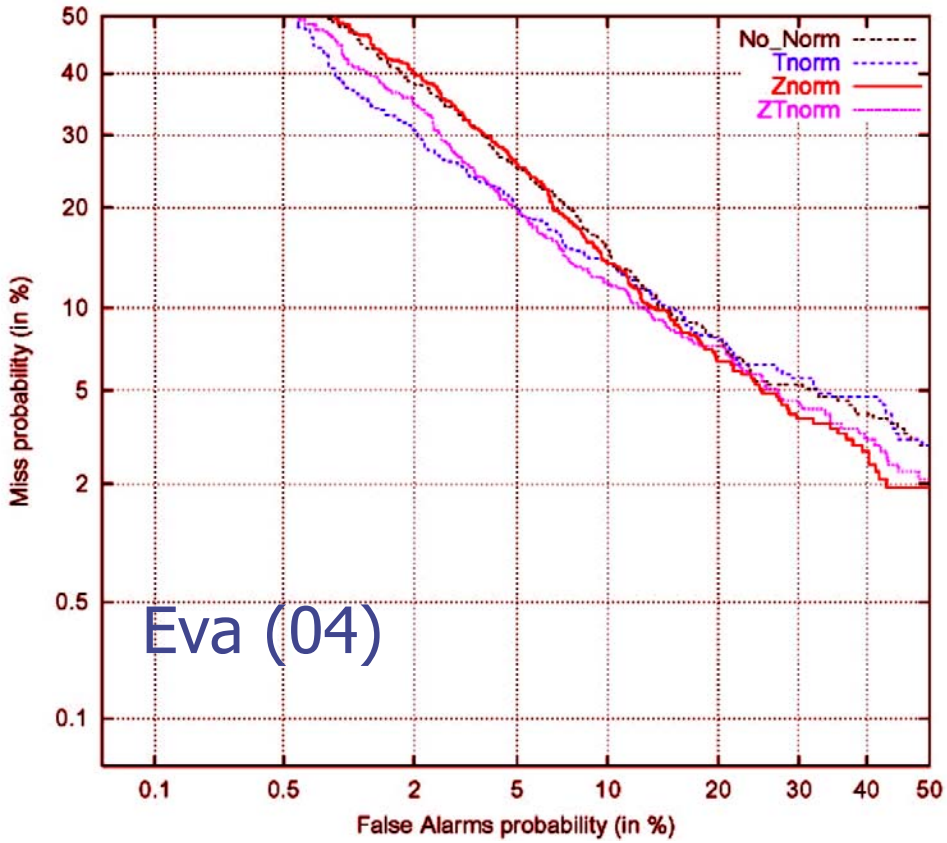
Exemple de points clés : Normalisation des scores

- LIMSI 2003 (C Barras)



Exemple de points clés : Normalisation des scores





Un système complet ALIZE/LIA_SpkDet

- Disponible en ligne

- Tout le logiciel présenté est en « libre »
  - Documenté
  - Maintenu
- Basé sur ALIZE (LGPL), un toolkit intégrant les fonctionnalités « bas niveau »
  - LLK
  - EM
  - Viterbi
- http://www.lia.univ-avignon.fr/heberges/ALIZE/

Sources

- ***Authentification et sécurité**, Sylvain Meignier, cours DESS IVDI, IUP d'informatique Avignon*
- **Traitement de la parole : Reconnaissance de la langue et du locuteur**, Claude Barras, LIMSI-CNRS
- JF Bonastre et LJ Boe : l'expertise vocale en question
- Extraits des cours de C. Barras, M. Adda et JL. Gauvain, TLP, LIMSI
- **Automatic Speaker Recognition, Acoustics and Beyond,** Douglas Reynolds, Senior Member of Technical Staff, MIT Lincoln Laboratory, JHU CLSP, 10 July 2002
- Illustrations extraites des livres **Calliope** et de R. Boité
- **Reconnaissance du locuteur,** Sylvain Meignier, Claude Barras
- **Reconnaissance Automatique du Locuteur, performances et Limites , Corinne Fredouille**, Laboratoire Informatique d'Avignon (LIA), *6ième Congrès d'Acoustique – Session spéciale, Les expertises vocales: l'identification en question, 10 Avril 2002 – Lille*