TRAITEMENT DE LA PAROLE

RECONNAISSANCE DU LOCUTEUR

Le Mans Université

Anthony Larcher - Le Mans Université - 2019

INTERVENANT: ANTHONY LARCHER

- Ingénieur électricien (traitement du signal)
- Master en traitement du signal et des images (détection d'explosif par rayon-X)
- PhD. sur la reconnaissance du locuteur (RAL) dépendante du texte
- Projets:
  - Biométrie bi-modale embarquée (audio-vidéo)
  - RAL pour la domotique (accès et commandes vocales)
  - Authentification vocale via un canal VHF (Port de Singapour)
  - Authentification vocale pour services bancaires
  - Déverrouiller un téléphone portable grâce à un mot de passe personnalisé
  - Systèmes intelligents pour apprentissage non-supervisé
- Directeur de l'institut informatique Claude Chappe
- Membre du CA de l'Association Francophone de la Communication Parlée
- Secrétaire du Groupe d'intérêt scientifique SpLC (Speaker and Language Characterization)

TRAITEMENT DE LA PAROLE

RECONNAISSANCE DU LOCUTEUR

Le Mans Université

Anthony Larcher - Le Mans Université - 2019

LA PAROLE: UN SIGNAL RICHE ET COMPLEXE

Oh, cet **accent du sud**, c'est **MAX** qui me dit « **BONJOUR** » en **Français**.
Il a l'air **fatigué** ce matin.

Un humain est capable de détecter:

- La langue parlée
- L'identité de la personne (si elle est connue)
- Le texte prononcé par cette personne (si la langue est connue)
- Les phonèmes prononcés
- L'accent ou les caractéristiques régionale du locuteur
- Les émotions ou l'état du locuteur
- L'environnement ambiant
- Le canal de transmission

OBJECTIF DU COURS: MODÉLISATION ACOUSTIQUE

- (Détection de la parole)
- Reconnaissance du locuteur
- Identification de la langue

PLAN DU COURS

- Introduction aux tâches visées
- Le paradigme Gaussien en modélisation acoustique
- Introduction au *Factor Analysis*
- Les approches neuronales

COURS 1

- Définition les tâches de classification (locuteur, langue)
- Introduction à la modélisation acoustique:
  - variabilités acoustiques
  - visualisation des données
  - méthodes de classification
- Aperçu d'un système complet
- Sorties d'un système

TRAITEMENT DE LA PAROLE

DÉTECTION DE LA PAROLE

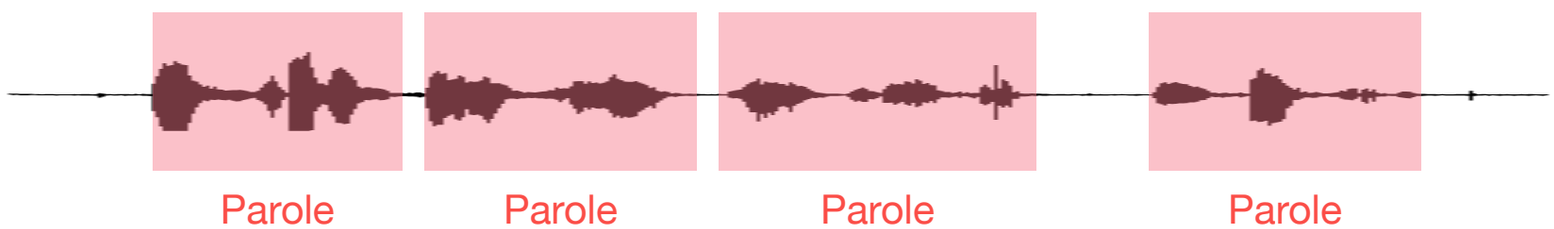
Le Mans Université

Anthony Larcher - Le Mans Université - 2019

DÉTECTION DE LA PAROLE: OBJECTIF

Voice Activity Detection (VAD)
ou Speech Activity Detection (SAD)

Tâche qui consiste à déterminer dans un **flux audio** les **segments** qui contiennent un **signal de parole**.



DÉTECTION DE LA PAROLE: UTILISATIONS

Cas d'usage:

- détecter le début d'une interaction humain/machine (éviter le « ok google » ou « dis Siri »
- Limiter le calcul en ne traitant que la parole
- Limiter l'espace de stockage en ne stockant que la parole

TRAITEMENT DE LA PAROLE

RECONNAISSANCE DU LOCUTEUR

Le Mans Université

Anthony Larcher - Le Mans Université - 2019

RECONNAISSANCE DU LOCUTEUR: OBJECTIF

À **qui** appartient l'échantillon de voix collecté?

Hypothèses:

- le système automatique connaît des locuteurs
- le système est capable de comparer les locuteurs connus à l'échantillon collecté

RECONNAISSANCE DU LOCUTEUR: UTILISATIONS

- authentification à des fins commerciales
- authentification à des fins sécuritaire (détention à domicile)
- détection: écoutes à la recherche d'une personne
- regroupement par locuteur (indexation de bases de données)
- regroupement par locuteur (annotation de réunions)

TRAITEMENT DE LA PAROLE

IDENTIFICATION DE LA LANGUE

Le Mans Université

Anthony Larcher - Le Mans Université - 2019

IDENTIFICATION DE LA LANGUE: OBJECTIF

Dans quelle langue cette personne parle-t'elle?

Hypothèses:

- le système automatique connaît des langues
- le système est capable de comparer les langues connues à l'échantillon collecté

IDENTIFICATION DE LA LANGUE: UTILISATIONS

Étape préliminaire pour:

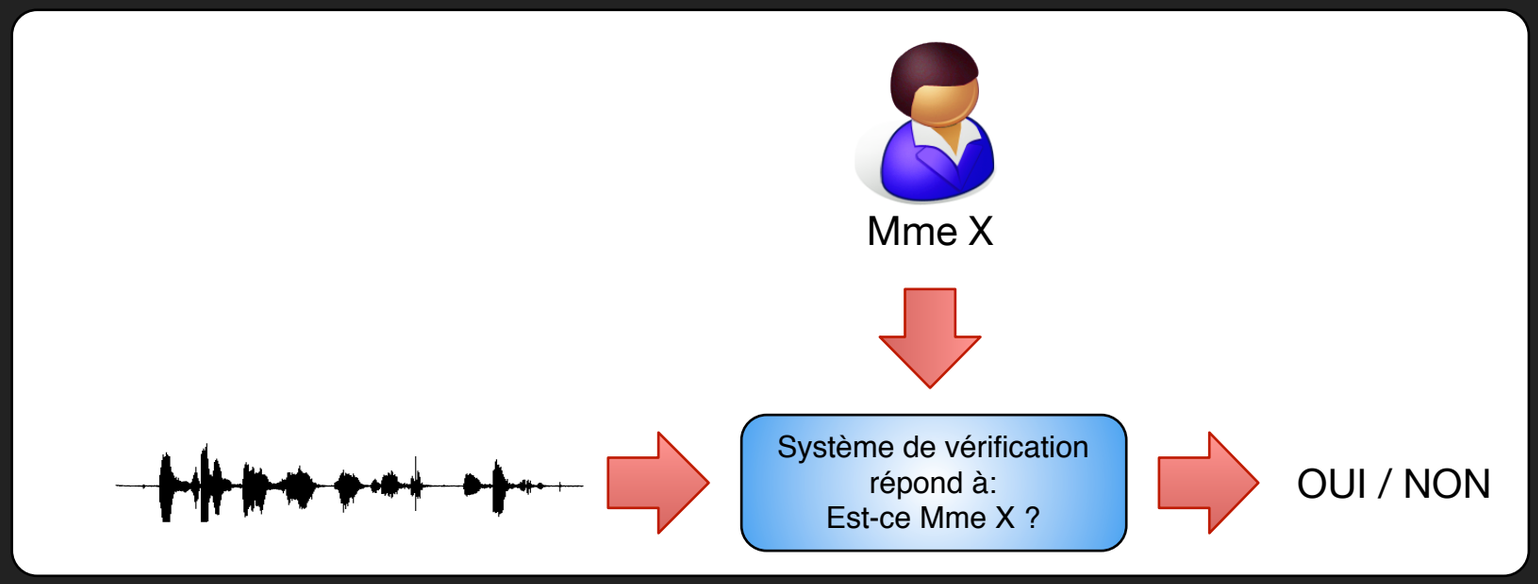
- la transcription automatique
- la traduction automatique
- plus généralement: toute interface vocale
- la reconnaissance du locuteur (variabilité intra-locuteur)
- aiguiller sur un opérateur parlant la langue

TÂCHES DE CLASSIFICATION

- Locuteur, langue, détection de parole... = classification
- Il faut bien différencier:
  - vérification
  - identification en milieu fermé
  - identification en milieu ouvert

TÂCHES DE CLASSIFICATION: VÉRIFICATION

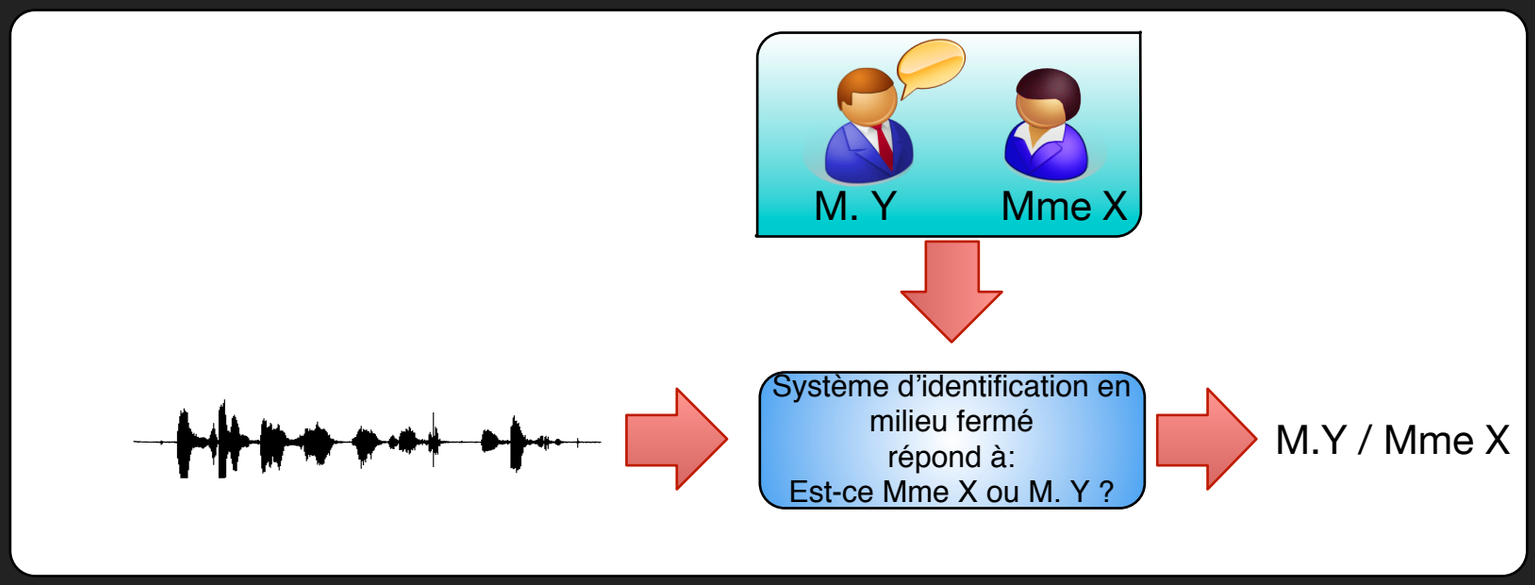
- tâche de classification binaire: réponse oui / non



- entrées: 1 identité + 1 segment audio
- sortie: oui / non

TÂCHES DE CLASSIFICATION: IDENTIFICATION EN MILIEU FERMÉ

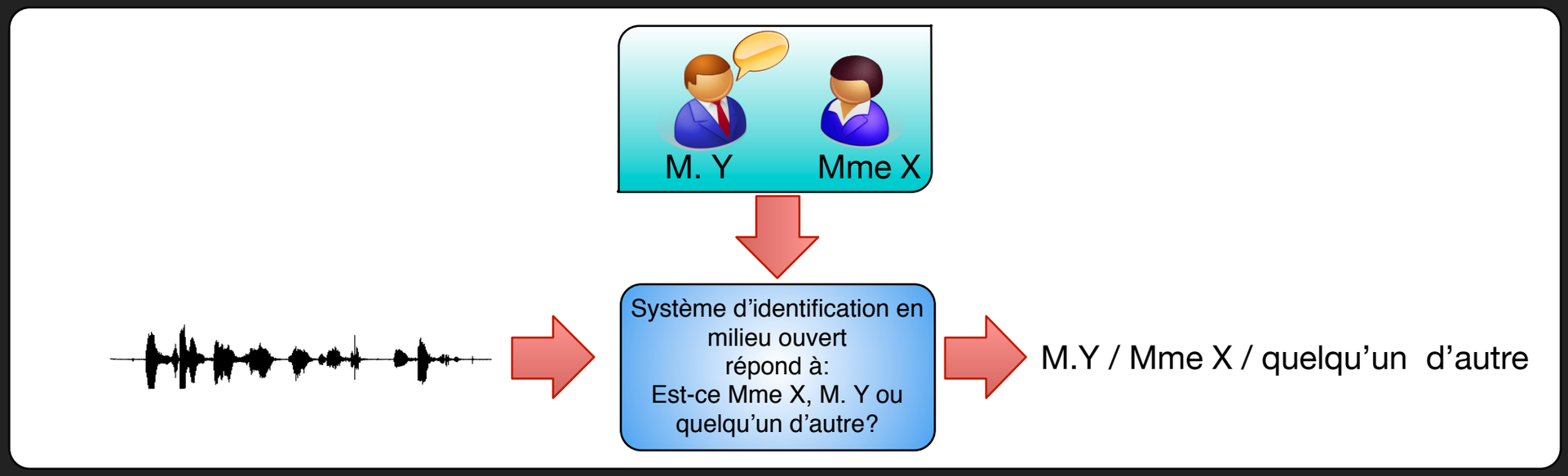
- tâche de classification 1 parmi N



- entrées: 1 segment audio + 1 base de N locuteurs
- sortie: 1 des N locuteurs

TÂCHES DE CLASSIFICATION: IDENTIFICATION EN MILIEU OUVERT

- tâche de classification 1 parmi N + 1



- entrées: 1 segment audio + 1 base de N locuteurs
- sortie: 1 des N locuteurs / aucun des N locuteurs

TÂCHES DE CLASSIFICATION: IDENTIFICATION EN MILIEU OUVERT

- Quelle tâche permet de répondre à toutes les questions?

TÂCHES DE CLASSIFICATION: IDENTIFICATION EN MILIEU OUVERT

- Quelle tâche permet de répondre à toutes les questions?

La vérification

Par la suite nous ne nous intéresserons presque qu'à la vérification

ARCHITECTURE D'UN SYSTÈME DE VÉRIFICATION DU LOCUTEUR

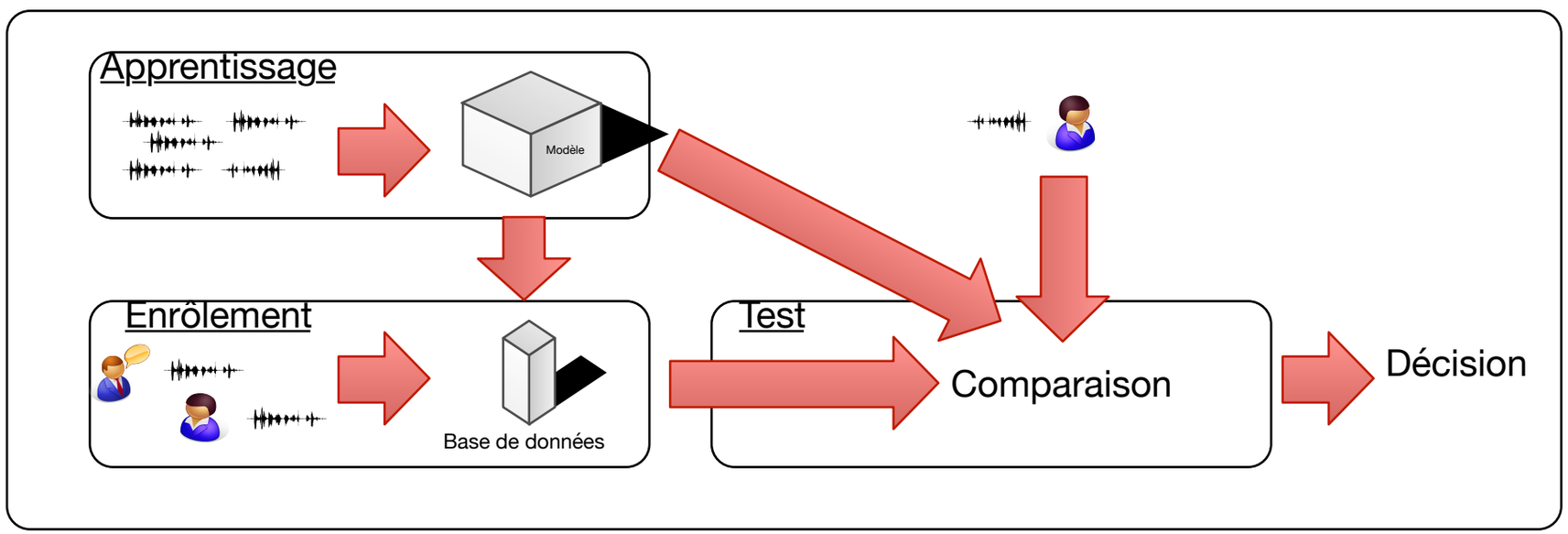
- **Apprentissage**
  apprendre une connaissance a priori sur la voix humaine, les caractéristiques des individus (pour optimiser la comparaison)

- **Enrôlement**
  acquisition des caractéristiques propres à un locuteur à partir d'un ou plusieurs échantillons

- **Test**
  comparaison des caractéristiques de locuteurs connus avec les caractéristiques extraites d'un échantillon test

ARCHITECTURE D'UN SYSTÈME DE VÉRIFICATION DU LOCUTEUR

- Apprentissage
- Enrôlement
- Test



INTRODUCTION À LA MODÉLISATION ACOUSTIQUE

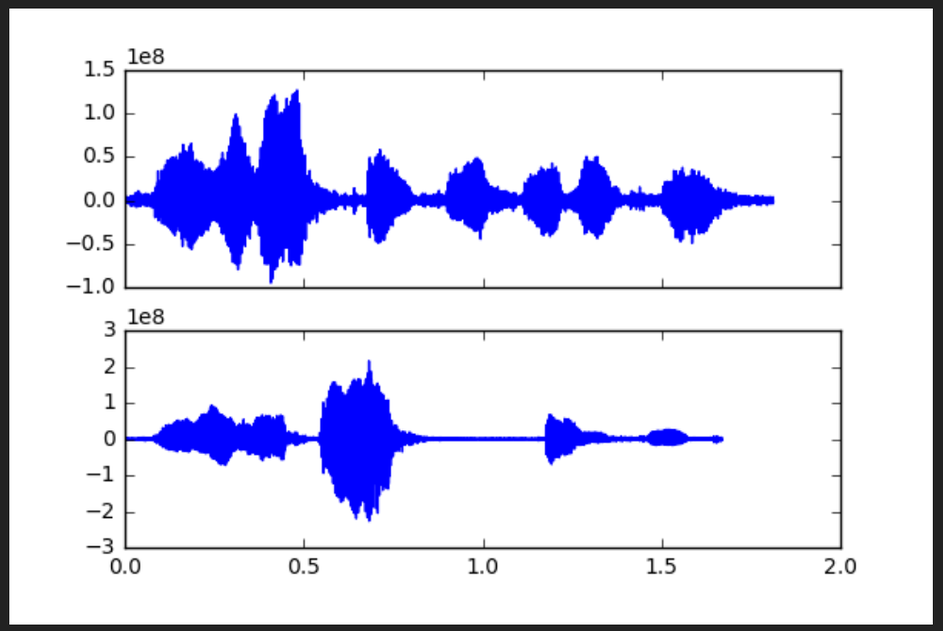
On ne peut pas comparer deux signaux directement

- ils ont une longueur variable
- le signal contient trop de bruit
- la représentation temporelle n'est pas adaptée (redondante, volumineuse)

INTRODUCTION À LA MODÉLISATION ACOUSTIQUE

On ne peut pas comparer deux signaux directement

- ils ont une longueur variable

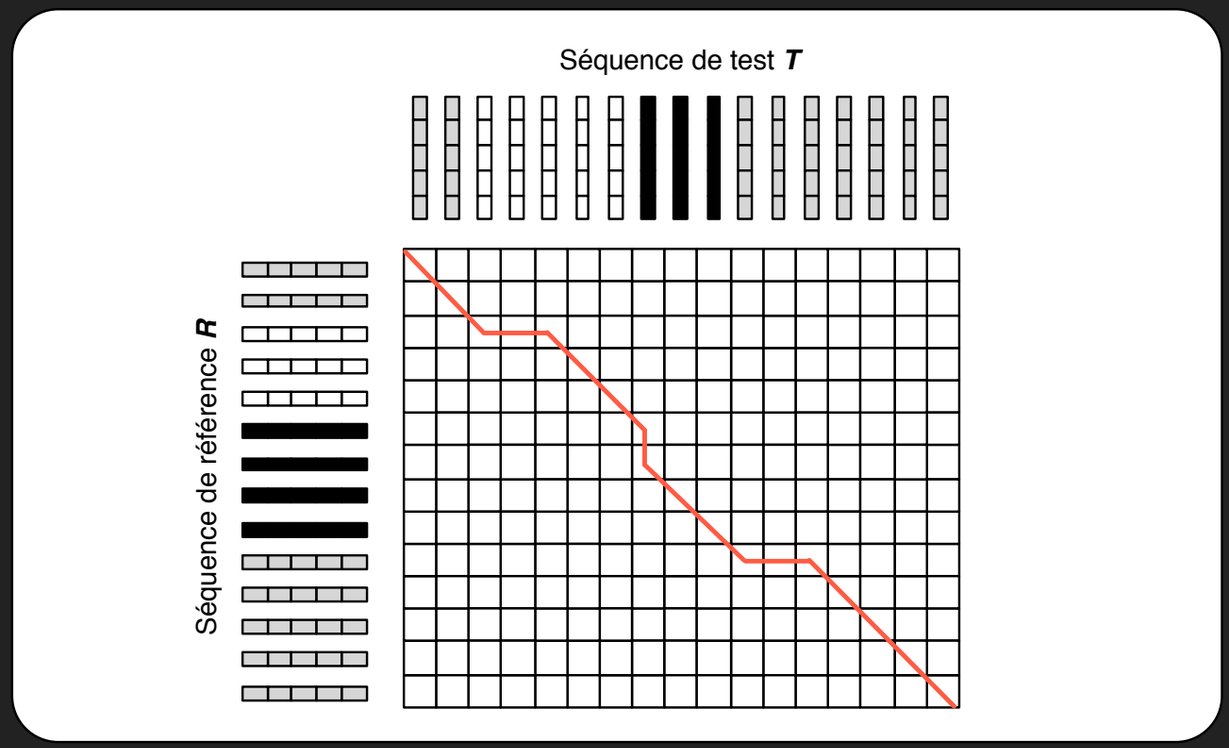


2 phrases du RSR2015: même locuteur, même phrase
quantité de parole différente

INTRODUCTION À LA MODÉLISATION ACOUSTIQUE

On ne peut pas comparer deux signaux directement

- ils ont une longueur variable

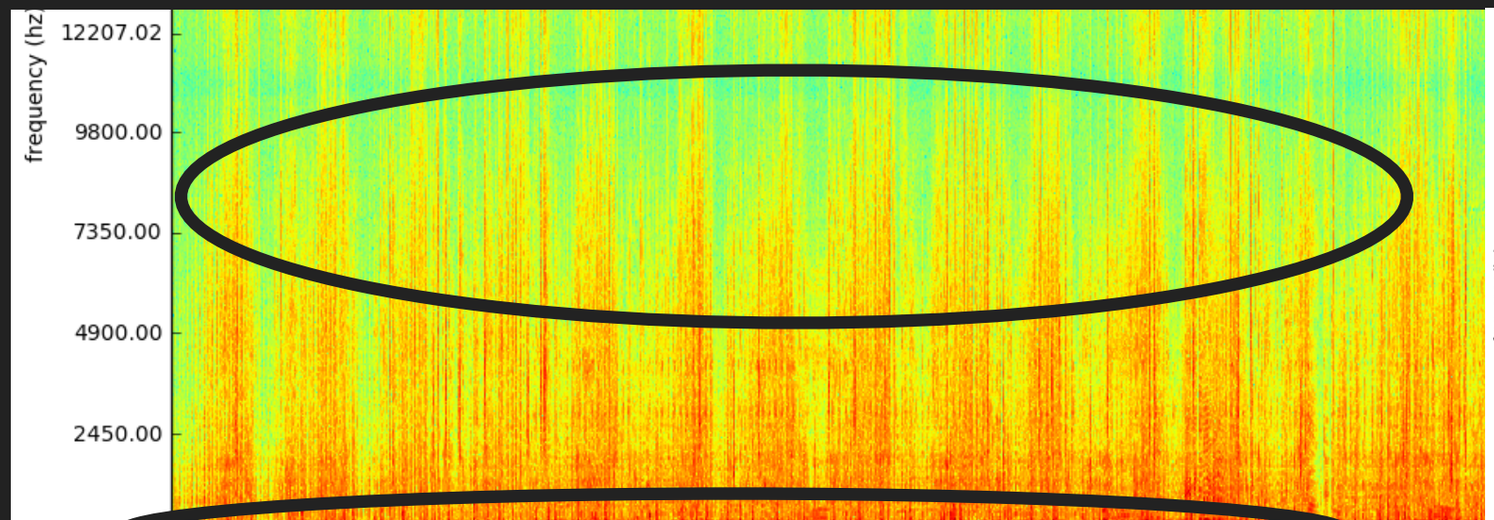


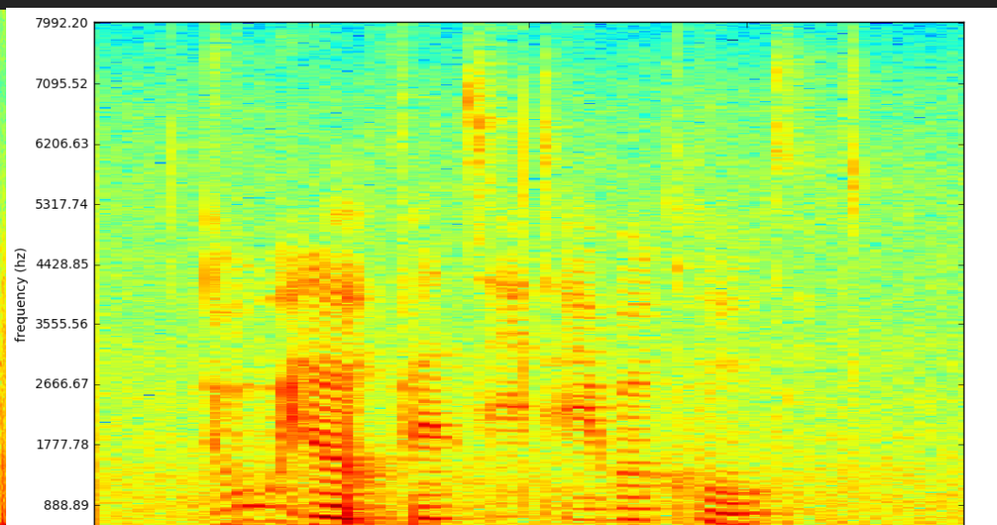
Dynamic Time Warping: comparaison à la trame: peu robuste.

INTRODUCTION À LA MODÉLISATION ACOUSTIQUE

On ne peut pas comparer deux signaux directement

- le signal contient trop de bruit





La représentation temporelle ne permet pas de voir certaines choses, ici certaines bandes de fréquences occupées par la musique (à gauche) sont absentes dans le signal de parole (à droite).

INTRODUCTION À LA MODÉLISATION ACOUSTIQUE

On ne peut pas comparer deux signaux directement

- la représentation temporelle n'est pas adaptée (redondante, volumineuse)
- Comme tous les signaux de communication, la voix contient une forte redondance pour permettre au destinataire de recevoir toute l'information

INTRODUCTION À LA MODÉLISATION ACOUSTIQUE

Pour la reconnaissance du locuteur ou de la langue

- on ne peut pas directement classifier des segments de longueur variable
- on recherche une représentation unique du locuteur ou de la langue (on classifie dans un espace discret!)
- Hypothèse: les représentations de 2 locuteurs ou 2 langues ont des tailles similaires (indépendante de la durée de l'échantillon)

INTRODUCTION À LA MODÉLISATION ACOUSTIQUE

La modélisation acoustique suit deux paradigmes:

- paradigme modèle / segment
- paradigme segment / segment

INTRODUCTION À LA MODÉLISATION ACOUSTIQUE

Paradigme modèle / segment:

- Phase d'enrôlement:
  - on infère un modèle de référence
  - à partir d'un ou plusieurs segments
- Phase de test:
  - comparaison d'un échantillon avec le modèle
  - comparaison asymétrique (modèle / segment)

Note: l'asymétrie du score peut être gênante lorsqu'on normalise les scores

INTRODUCTION À LA MODÉLISATION ACOUSTIQUE

Paradigme segment / segment

- Phase d'enrôlement:
  - une représentation est obtenue pour le segment d'enrôlement
- Phase de test:
  - une représentation est obtenue pour le segment de test
  - les représentations d'enrôlement et de test sont comparés (distance, similarité…)

Note: il faut gérer le cas de multiples segments d'enrôlement (différentes options)

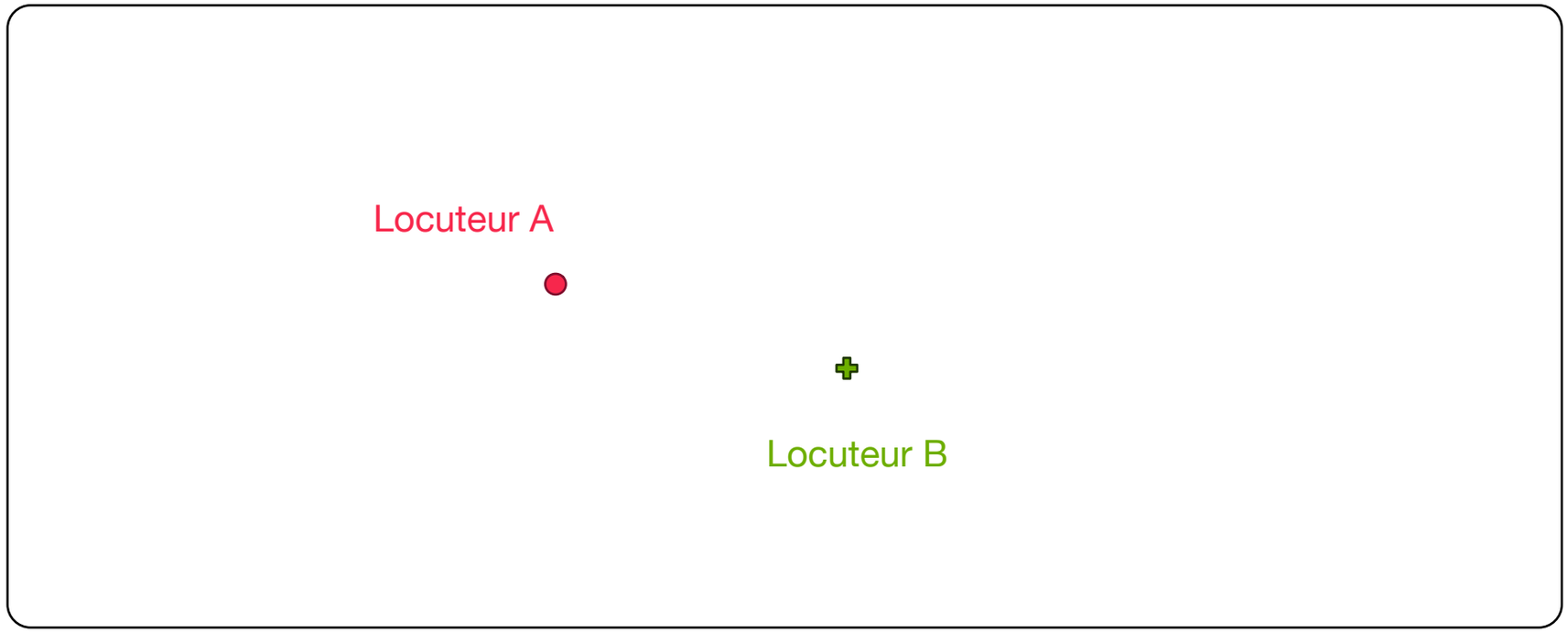
L'ESPACE ACOUSTIQUE VU COMME UN ESPACE EUCLIDIEN

Quel que soit le paradigme, on utilise une représentation vectorielle d'un segment acoustique:

- trame:
  - segment de ~25ms avec recouvrement
  - 20 < dimensions < 100
- super-vecteur:
  - segment de 1s à plusieurs minutes
  - 10 000 < dimensions < 50 000
- i-vecteur ou x-vecteur
  - segment de 1s à plusieurs minutes
  - 100 < dimensions < 1000

L'ESPACE ACOUSTIQUE VU COMME UN ESPACE EUCLIDIEN

On peut visualiser la représentation d'un locuteur dans un espace euclidien.



L'ESPACE ACOUSTIQUE VU COMME UN ESPACE EUCLIDIEN

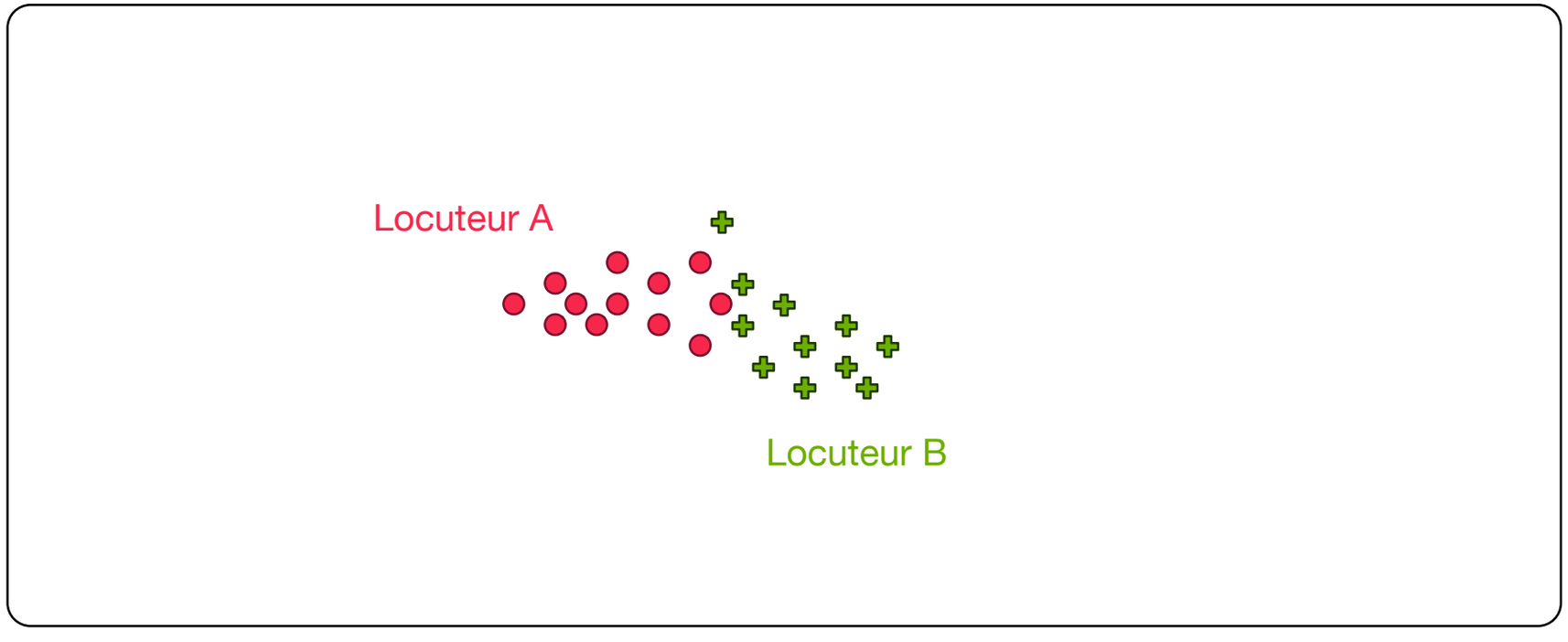
INTRODUCTION AUX VARIABILITÉS ACOUSTIQUES

Les échantillons d'une même classe diffèrent pour de nombreuses raisons

- durée
- locuteur
- texte prononcé
- accent
- bruit ambiant
- réverbération
- canal (microphone, transmission, compression)
- langue

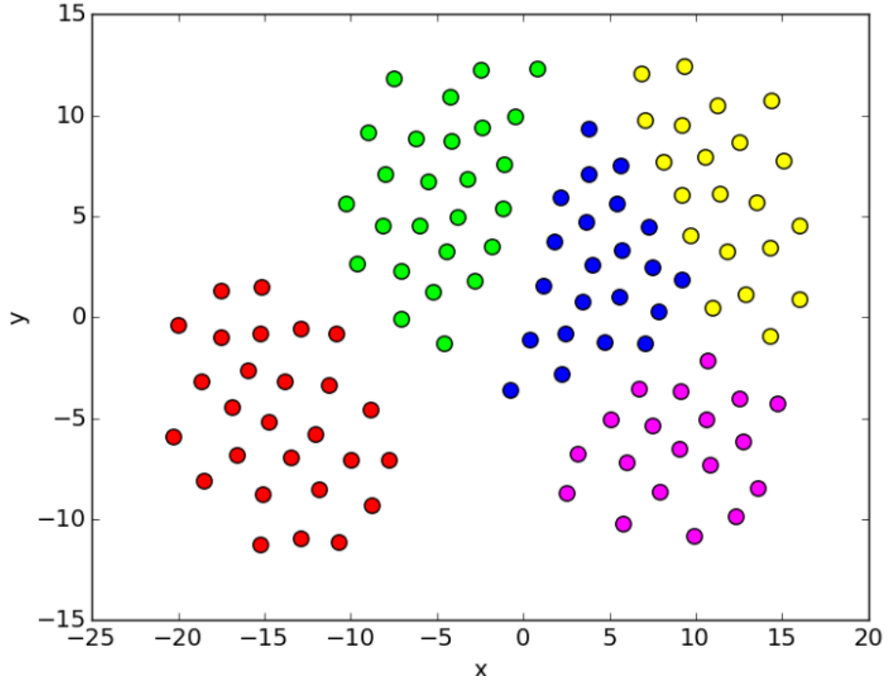
L'ESPACE ACOUSTIQUE VU COMME UN ESPACE EUCLIDIEN

Chaque échantillon obtenu d'un même locuteur fournit une représentation différente de ce locuteur: « variabilité »



L'ESPACE ACOUSTIQUE: UN ESPACE EUCLIDIEN

- En très grande dimension, il est impossible de visualiser on peut utiliser t-sne (t-Distributed Stochastic Neighbor Embedding)



LUKIC, Yanick, VOGT, Carlo, DÜRR, Oliver, *et al.* Speaker identification and clustering using convolutional neural networks. In : *Machine Learning for Signal Processing (MLSP), 2016 IEEE 26th International Workshop on*. IEEE, 2016. p. 1-6.

CLASSIFIER

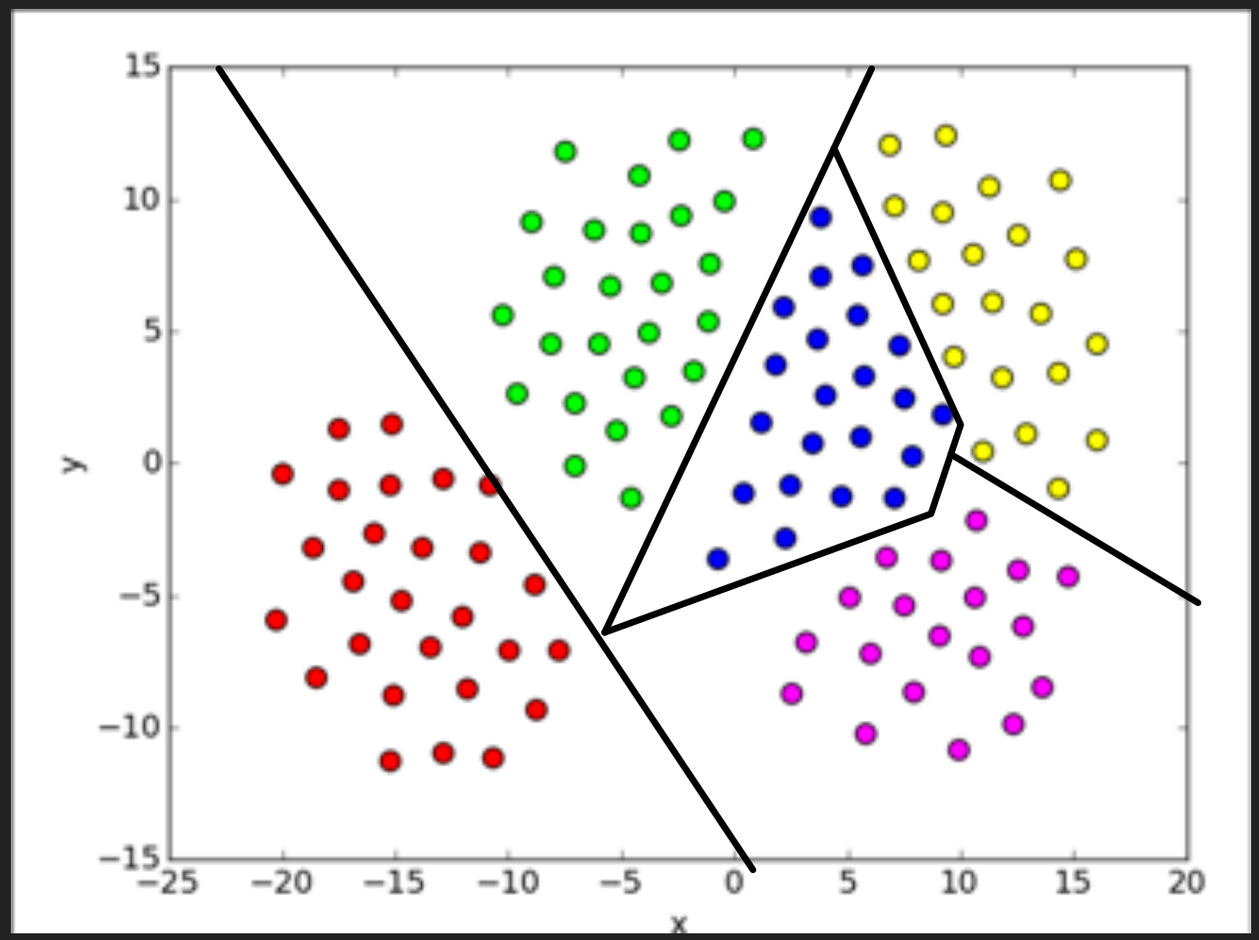
PAR APPROCHES DISCRIMINANTES OU GÉNÉRATIVES

Le Mans Université

Anthony Larcher - Le Mans Université - 2019

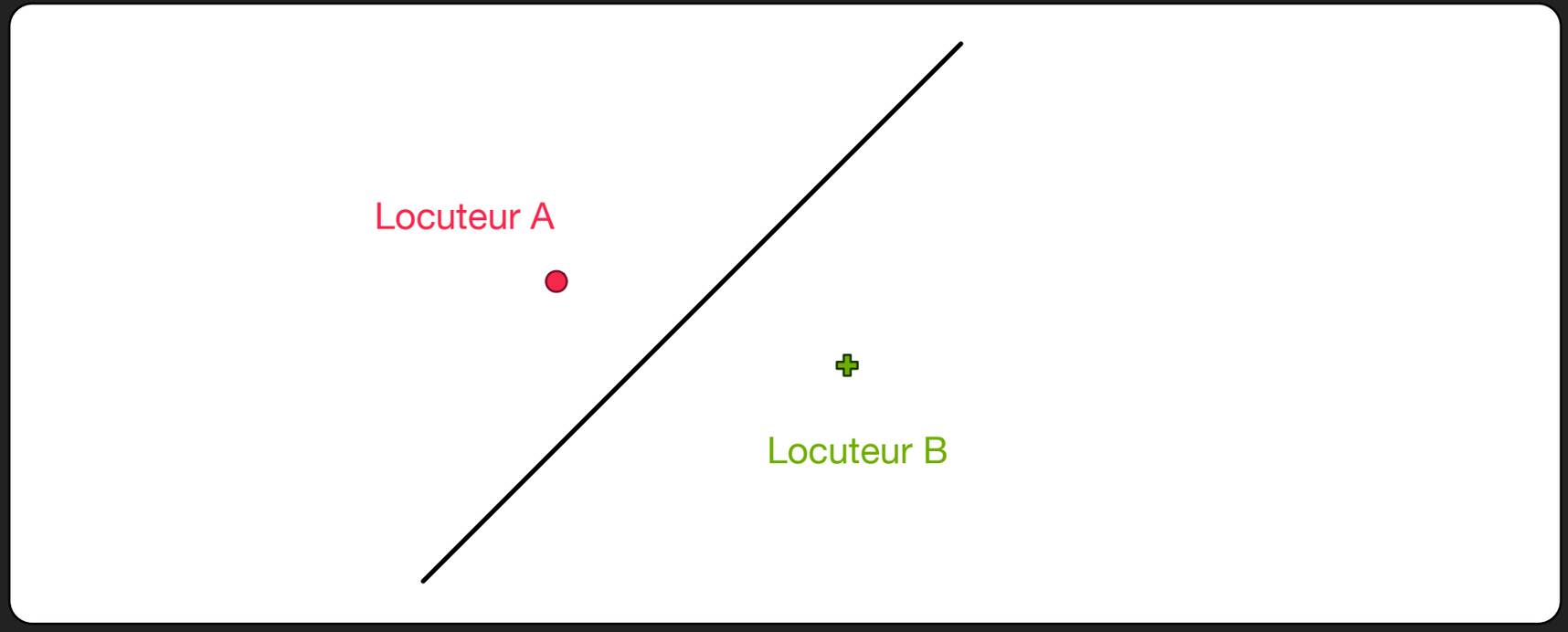
CLASSIFICATION PAR APPROCHES DISCRIMINANTES

- ce qu'on veut faire:



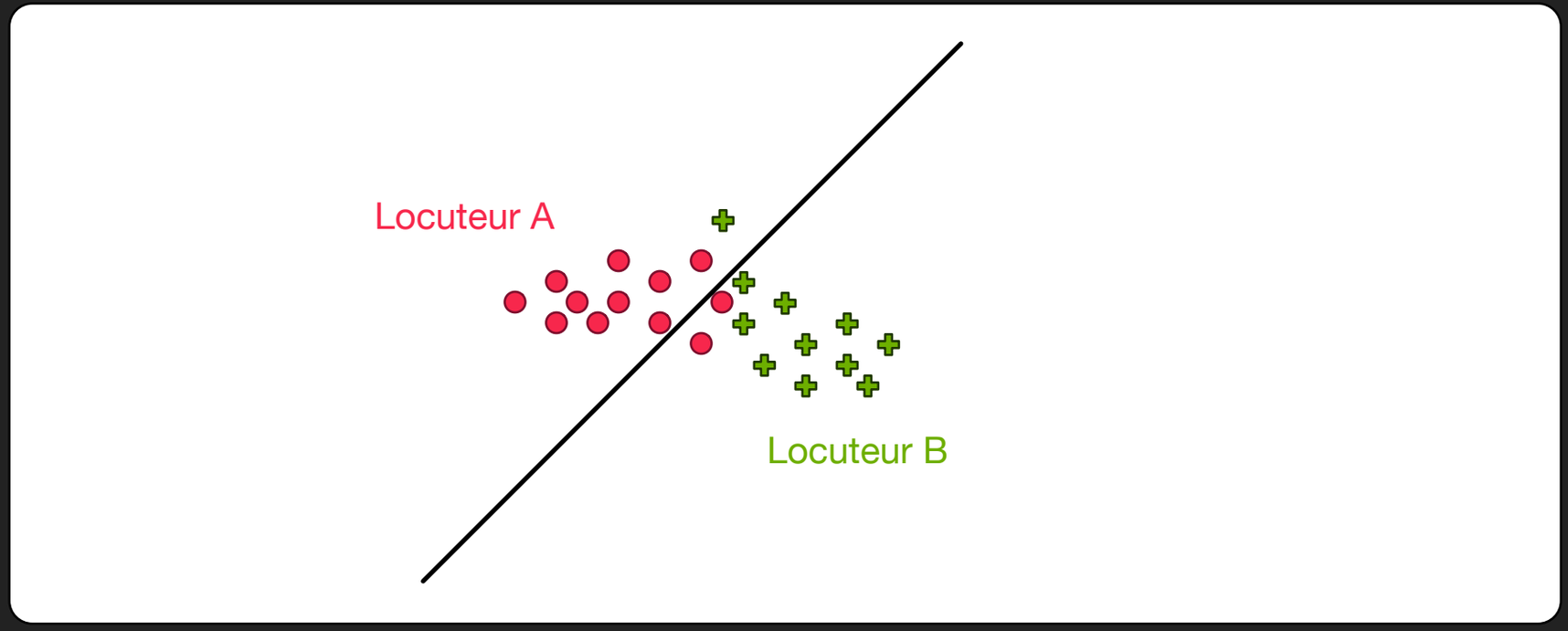
CLASSIFICATION PAR APPROCHES DISCRIMINANTES

- On apprend une limite entre 2 classes…



CLASSIFICATION PAR APPROCHES DISCRIMINANTES

- et en pratique…
  on n'a pas observé tous les échantillons possibles



CLASSIFICATION PAR APPROCHES DISCRIMINANTES

De nombreuses approches discriminantes visent à trouver des frontières entre classes dans des espaces plus ou moins complexes

- Support Vector Machines (SVM)
- K-means
- réseaux de neurones (et réseaux profonds)
- ...

CLASSIFICATION PAR APPROCHES DISCRIMINANTES

Critiques principales

- manque de généralisation
- on ne voit pas tous les cas possibles
- que se passe-t'il dans les zones que l'on n'a pas vu?

En pratique les approches actuelles (DNN) peuvent apprendre avec un grand nombre d'exemples et comblent peu à peu ce manque.
à suivre...

CLASSIFICATION PAR APPROCHES GÉNÉRATIVES

Paradigme modèle/segment ou segment/segment

- On considère que les échantillons observés donnent des renseignements sur leur voisinage
- On infère un modèle de référence par classe
- Objectif: modéliser la distribution de probabilité P(X) des échantillons de la classe

CLASSIFICATION PAR APPROCHES GÉNÉRATIVES

Exemple de la modélisation Gaussienne [1]

- On fait l'hypothèse que les échantillons observés donnent des renseignements sur leur voisinage
- À partir de quelques échantillons, on peut estimer la probabilité d'une nouvelle observation d'appartenir à la classe cible

[1] F. Bimbot, I. Magrin-Chagnolleau et L. Mathan, Second-order statistical measures for text-independent speaker identification, in Speech Communication, 1995, vol. 17, no 1-2, p 177-192

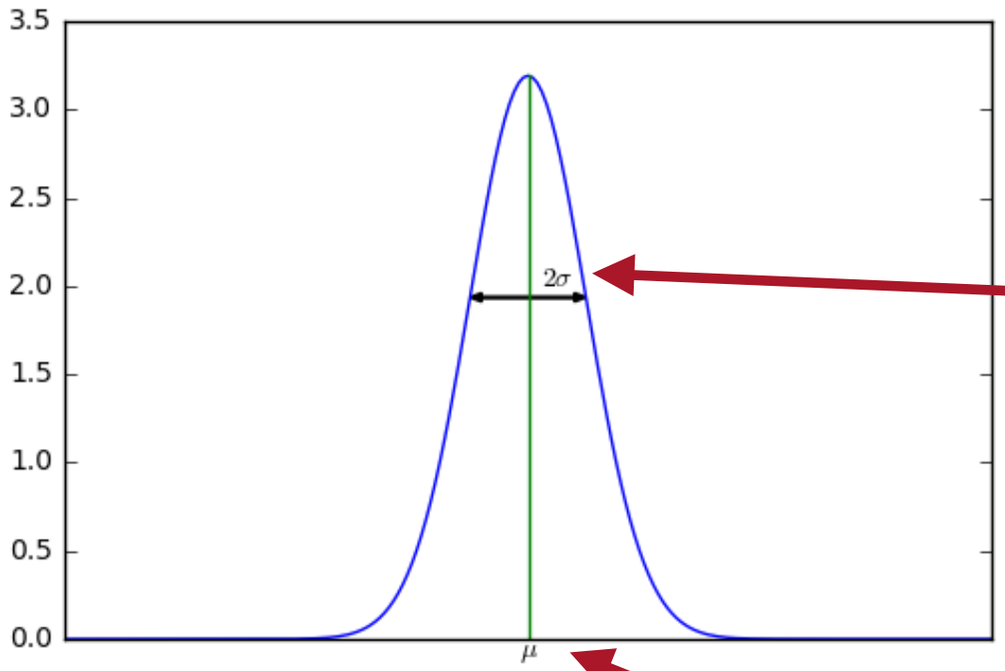
CLASSIFICATION PAR APPROCHES GÉNÉRATIVES

Pourquoi la modélisation Gaussienne?

- mathématiquement « facile »
- nécessite relativement peu de paramètres:
  - moyenne
  - variance
- Théorème Central-limite
  les ***MOYENNES*** d'échantillons indépendants qui suivent une même loi de probabilité tendent vers une distribution normale pour peu qu'elles soient suffisamment nombreuses.

CLASSIFICATION PAR APPROCHES GÉNÉRATIVES

Exemple de la modélisation Gaussienne



écart type
(racine carré de la variance)

moyenne

CLASSIFICATION PAR APPROCHES GÉNÉRATIVES

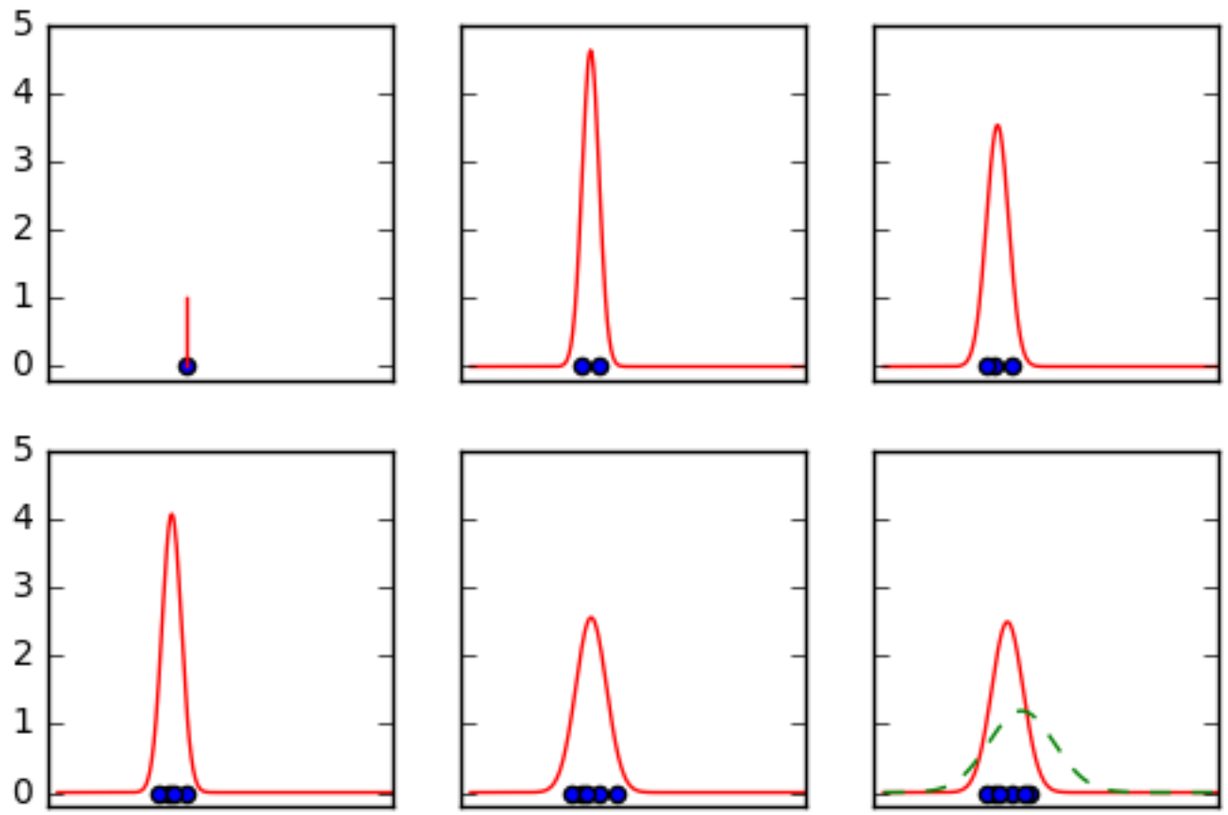
Comment utiliser un modèle Gaussien?

On se place dans l'espace des trames acoustiques

- 1 vecteur de ~50 dimensions
- calculé sur une fenêtre glissante de ~30ms
- toutes les 10ms

CLASSIFICATION PAR APPROCHES GÉNÉRATIVES

Enrôlement d'un locuteur à partir des échantillons observés.



CLASSIFICATION PAR APPROCHES GÉNÉRATIVES

Test: estimer la probabilité qu'une nouvelle observation « o » ait été générée par le locuteur de modèle « X »

On veut calculer P(X|o): la probabilité que ce soit le modèle X qui ait généré l'observation o.

En pratique on calcule: P(o|X) qui d'après le théorème de Bayes est égal à:

$$P(o|X) = P(X|o)\, P(o) / P(X)$$

On fait l'hypothèse que P(o) / P(X) est une constante et donc P(X|o) et P(o|X) sont proportionnels.

CLASSIFICATION PAR APPROCHES GÉNÉRATIVES

Test: estimer la probabilité qu'une nouvelle observation « o » ait été générée par le locuteur de modèle « X »

$$P(o|X) = \frac{1}{\sqrt{2\Pi|\Sigma|}} \exp\left(-\frac{1}{2}(o-\mu)\Sigma^{-1}(o-\mu)^T\right)$$

Où μ est la moyenne de la Gaussienne et Σ la matrice de covariance

CLASSIFICATION PAR APPROCHES GÉNÉRATIVES

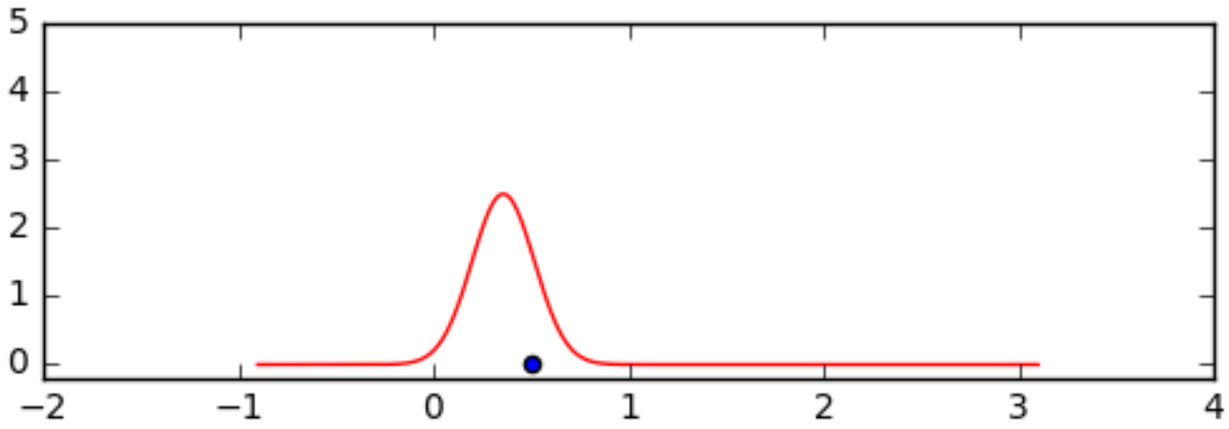
On a calculé:

$$P(o|X) = \frac{1}{\sqrt{2\Pi|\Sigma|}} \exp\left(-\frac{1}{2}(o-\mu)\Sigma^{-1}(o-\mu)^T\right)$$

Alors? Est-ce le bon locuteur?

CLASSIFICATION PAR APPROCHES GÉNÉRATIVES

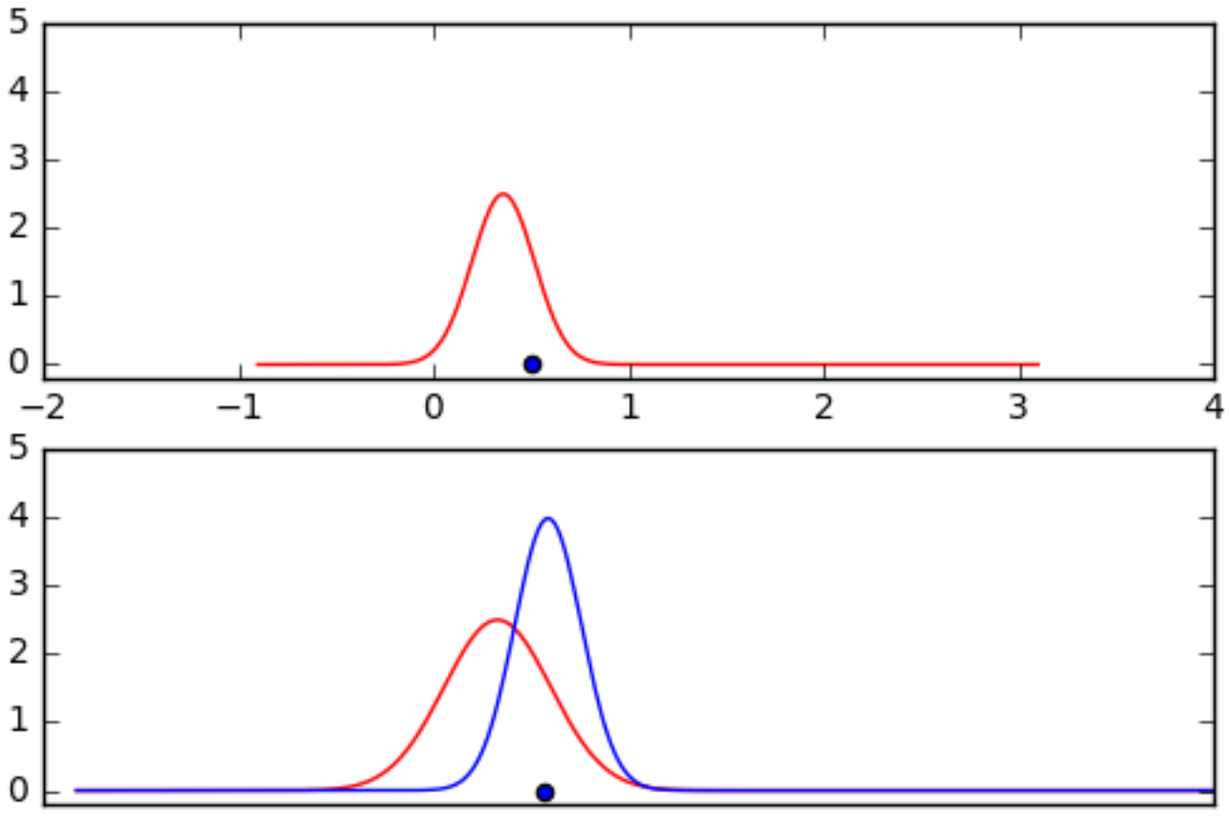
Est-ce le locuteur rouge?



Oui !

CLASSIFICATION PAR APPROCHES GÉNÉRATIVES

Est-ce le locuteur rouge?



Ah non...

CLASSIFICATION PAR APPROCHES GÉNÉRATIVES

- **Problème**:
  contrairement aux approches discriminantes, on a appris le modèle à partir des seules données du locuteur cible.

- Il faut prendre en compte la « spécificité » du locuteur.

- Il faut considérer ce qui est commun à tous les locuteurs et ce qui est spécifique au locuteur cible.

CLASSIFICATION PAR APPROCHES GÉNÉRATIVES

On considère un rapport d'hypothèses: le rapport de vraisemblance:

$$\frac{P(o|H_0)}{P(o|H_1)}$$

Où H0 est l'hypothèse selon laquelle

« o » a été produite par le locuteur cible

et H1 est l'hypothèse selon laquelle

« o » n'a pas été produite par le locuteur cible

CLASSIFICATION PAR APPROCHES GÉNÉRATIVES

On considère un rapport d'hypothèses: le rapport de vraisemblance:

$$\frac{P(o|H_0)}{P(o|H_1)}$$

Modèle d'une personne moyenne

Modèle d'un locuteur particulier
Modélise ce qui diffère de la moyenne

CLASSIFICATION PAR APPROCHES GÉNÉRATIVES

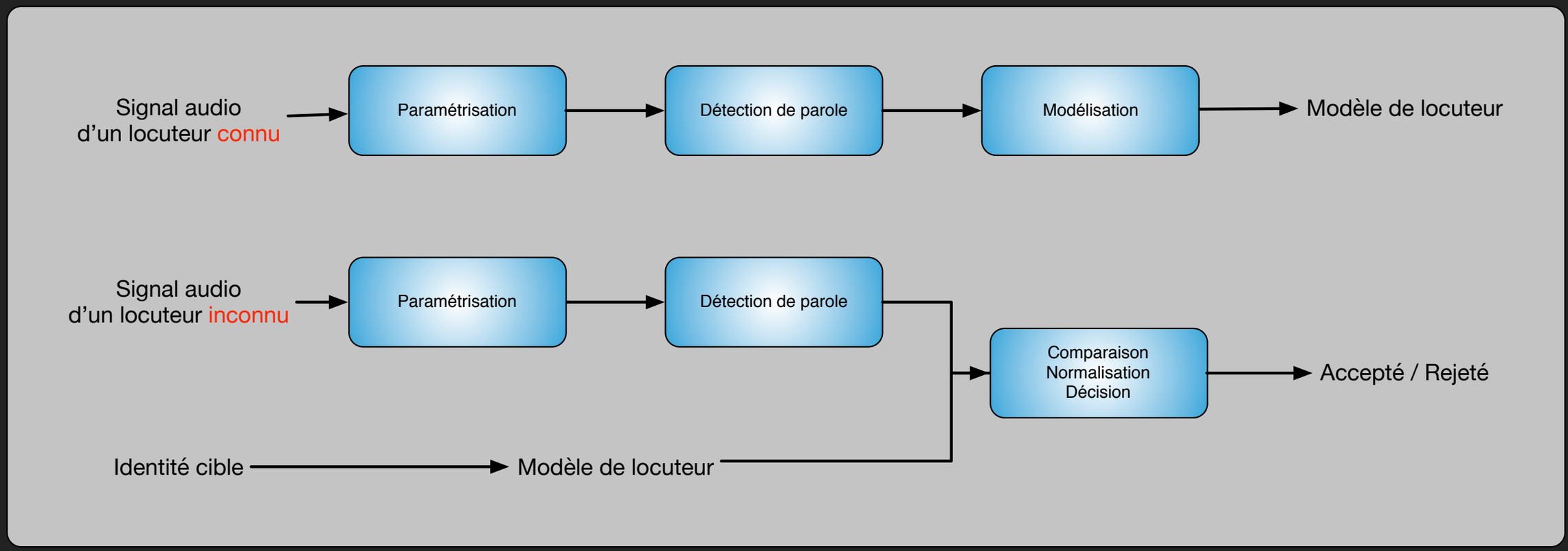
Pour décider si oui ou non c'est le locuteur cible (but de la vérification), on compare ce rapport à un seuil:

$$\frac{P(o|H_0)}{P(o|H_1)} < \Theta$$ Ce n'est pas le locuteur cible

$$\frac{P(o|H_0)}{P(o|H_1)} = \Theta$$ Aucune idée...

$$\frac{P(o|H_0)}{P(o|H_1)} > \Theta$$ C'est le locuteur cible

SYSTÈME DE RECONNAISSANCE DU LOCUTEUR



ANALYSE DES RÉSULTATS

Comment évalue-t'on un système de reconnaissance du locuteur?

Pas d'évaluation absolue

- Est-il possible que deux personnes aient la même voix?
- On ne peut pas connaître tous les locuteurs du monde
- Évaluation statistique (on fait un grand nombre de tests...)

ANALYSE DES RÉSULTATS

Toute évaluation est biaisée par:

- le groupe de locuteur utilisé (apprentissage, développement, calibration)
- le ou les environnement(s) acoustique(s)
- le type d'enregistrement (durée, variabilité…)
- les tests sont ils tous indépendant? (pas si ils font intervenir les mêmes locuteurs)

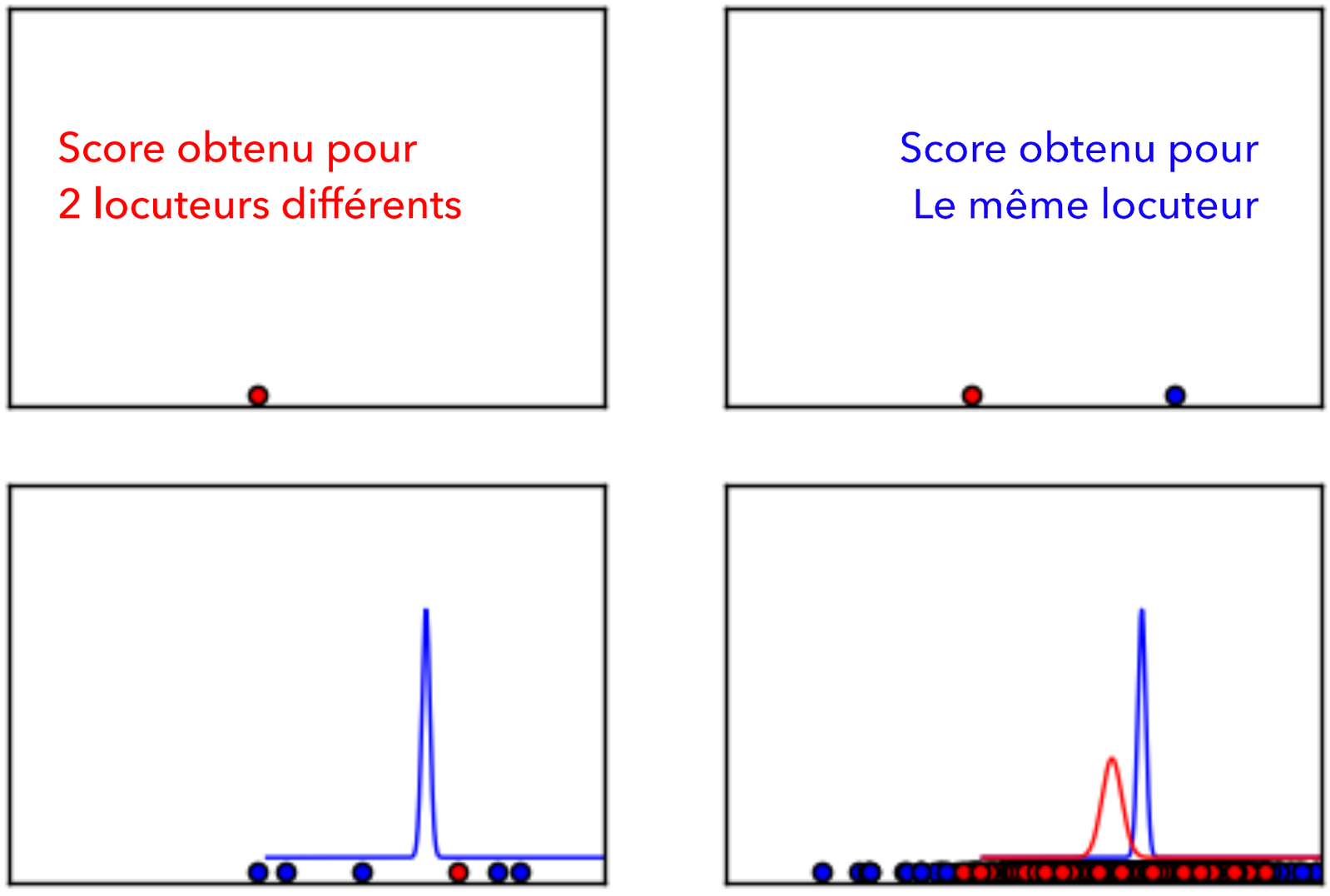
ANALYSE DES RÉSULTATS

Et pourtant...

Il faut évaluer quand même!

ANALYSE DES RÉSULTATS

Pour un système donné, on fait un grand nombre de tests



ANALYSE DES RÉSULTATS

Tâche de vérification = 2 types d'erreurs

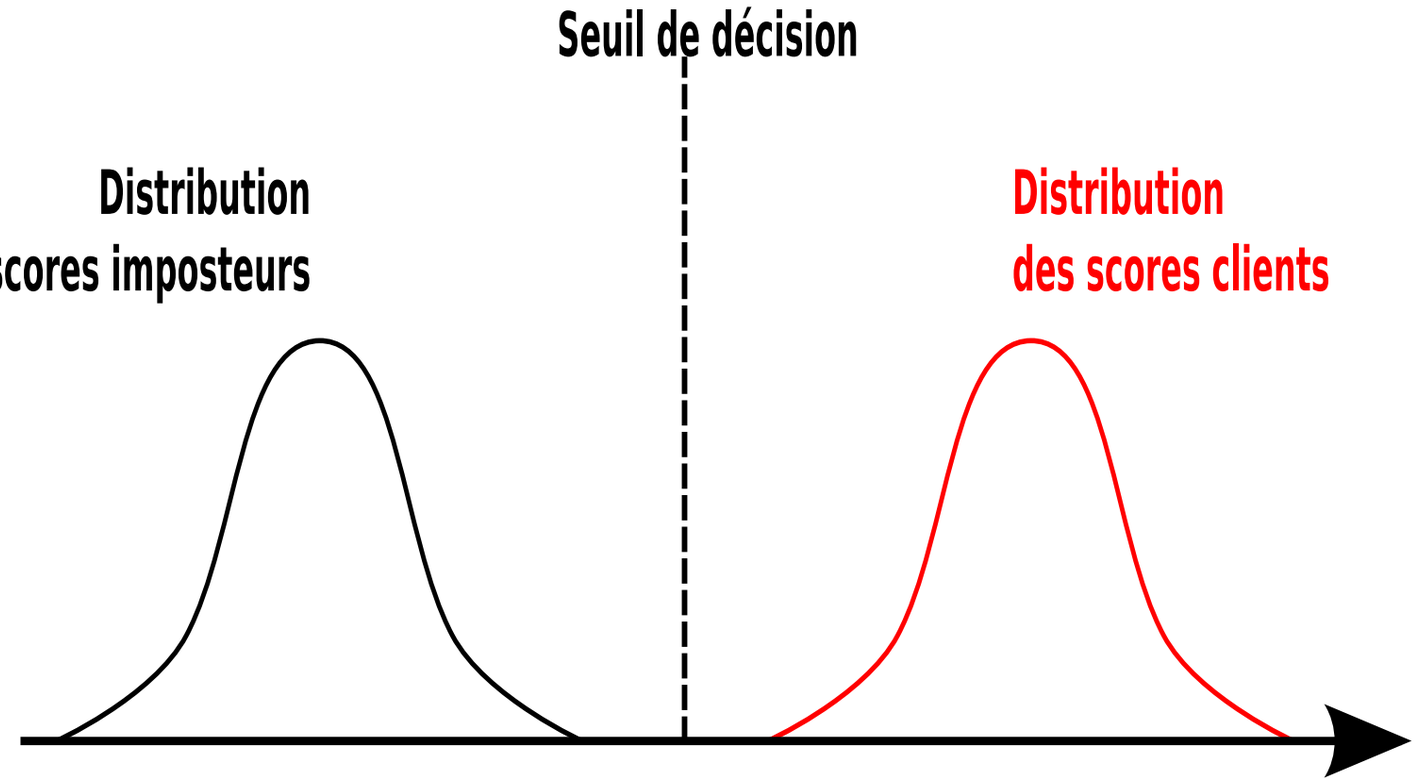
Fausse acceptation (FA: False Acceptance)

Ce n'était pas le locuteur cible mais nous l'avons cru

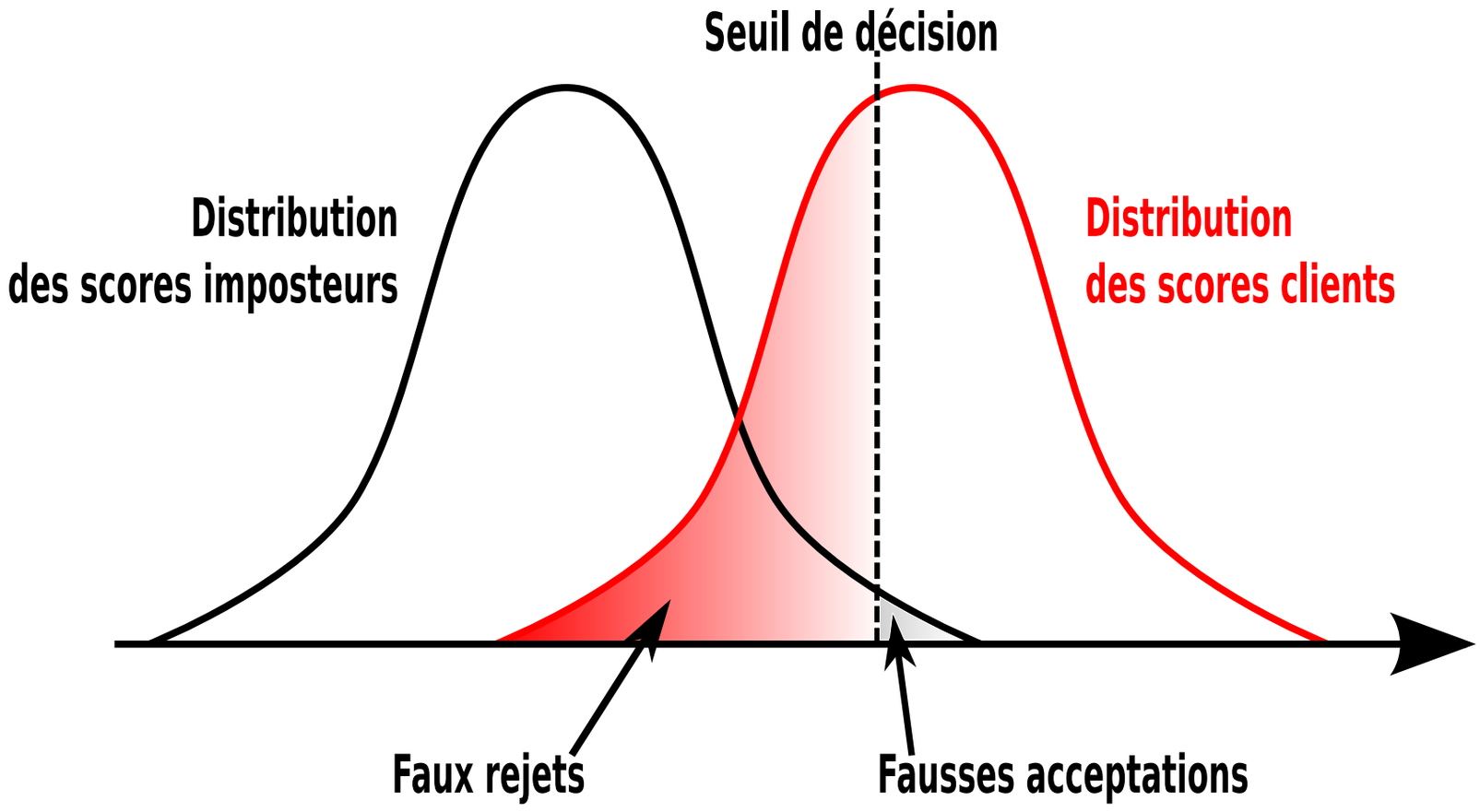
Faux rejet (Miss)

C'était le locuteur cible mais nous l'avons raté

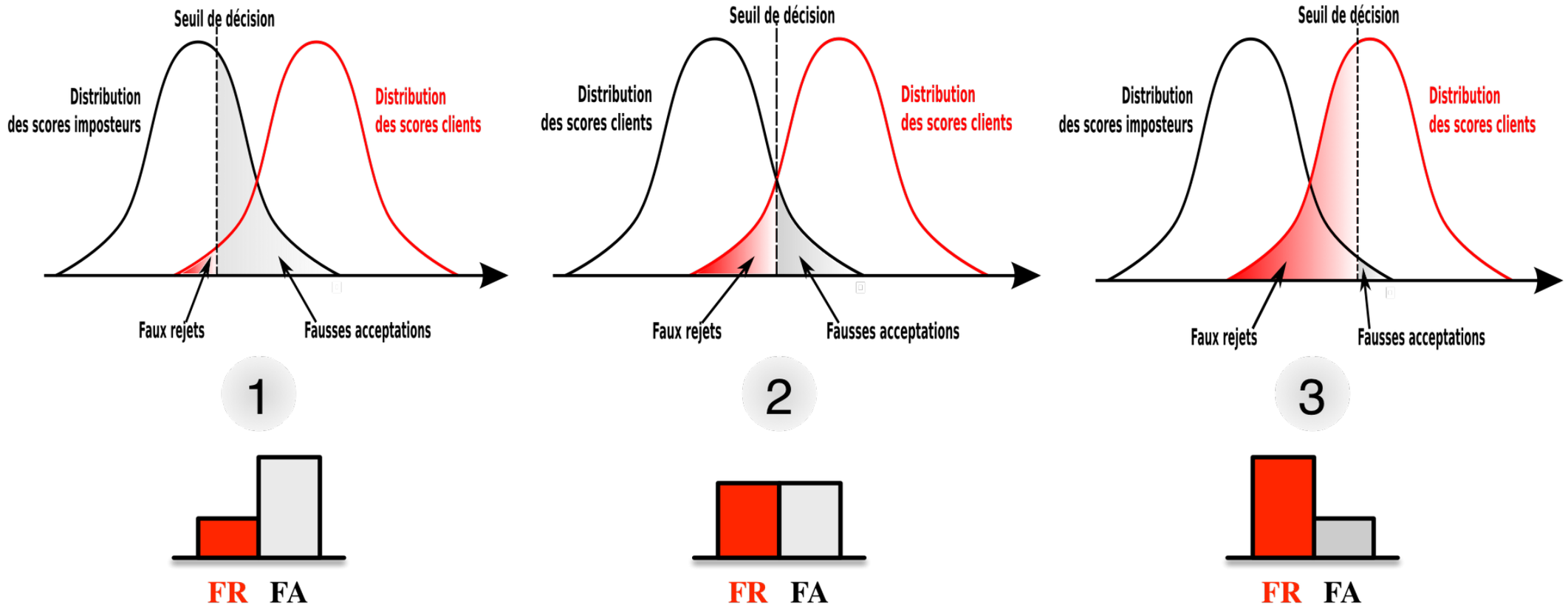
ANALYSE DES RÉSULTATS: CAS IDÉAL



ANALYSE DES RÉSULTATS: CAS RÉEL



ANALYSE DES RÉSULTATS

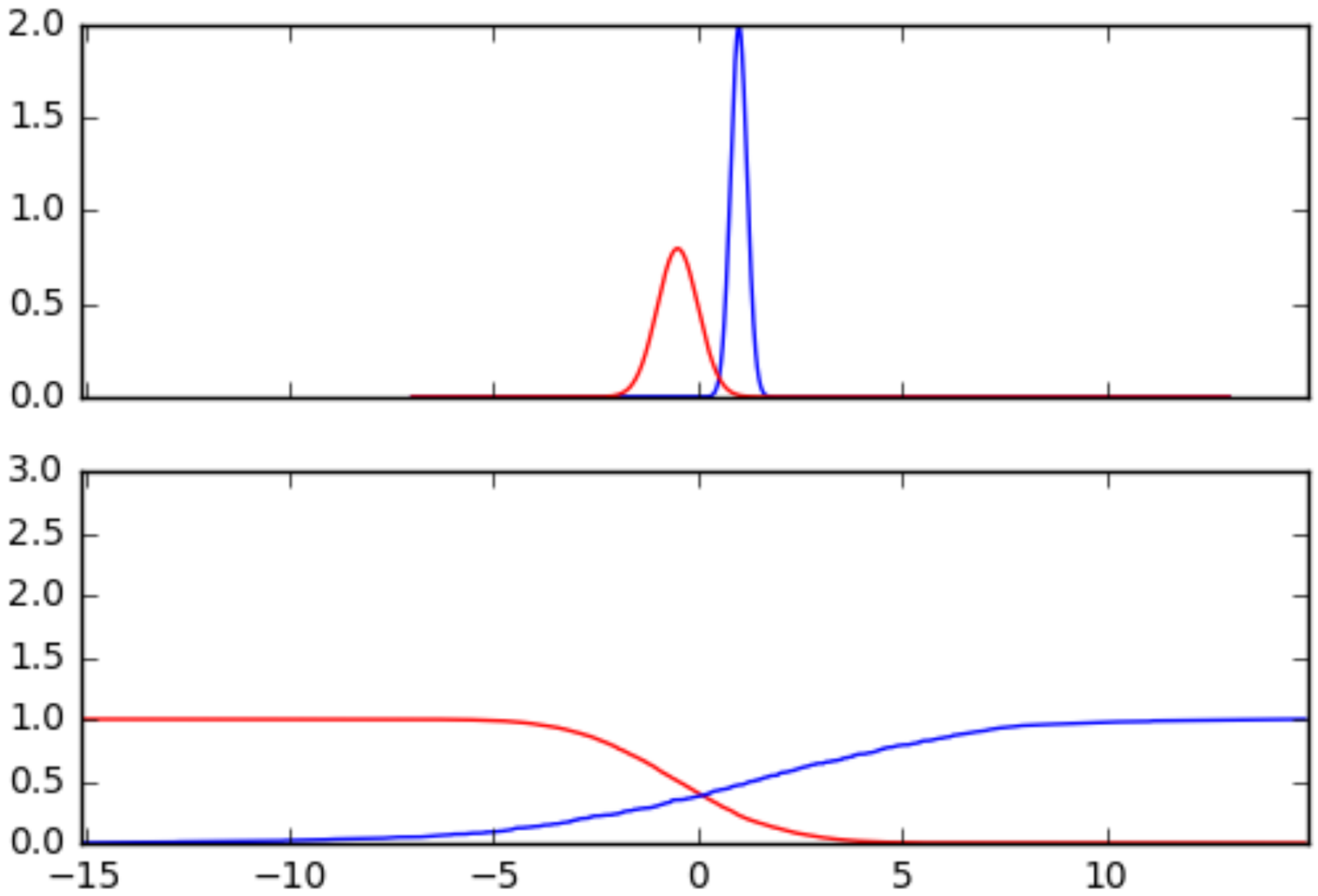


ANALYSE DES RÉSULTATS

Lorsqu'on déplace le seuil de décision, les taux d'erreur évoluent

tests clients

tests imposteurs



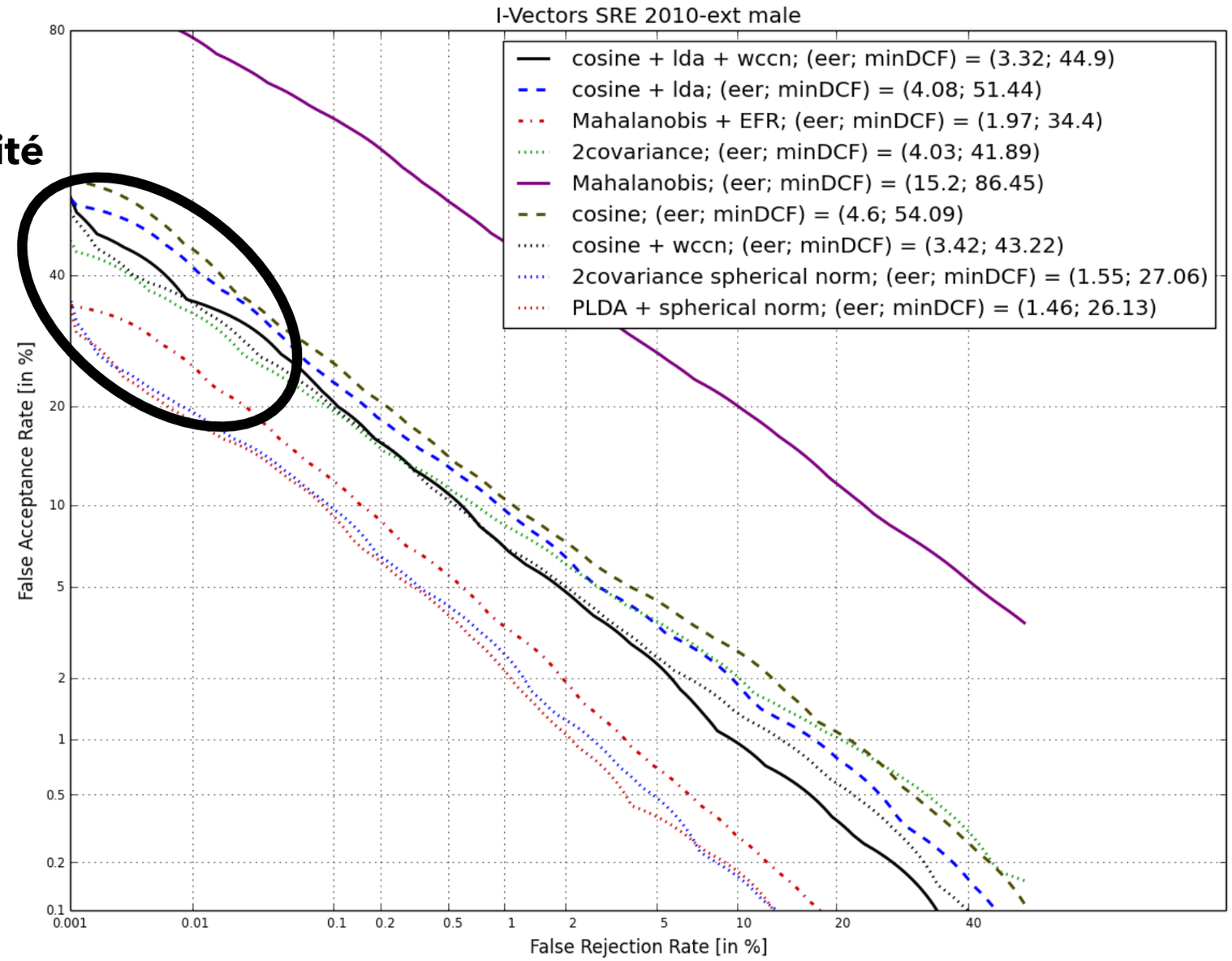
Miss

Fausses acceptations

ANALYSE DES RÉSULTATS

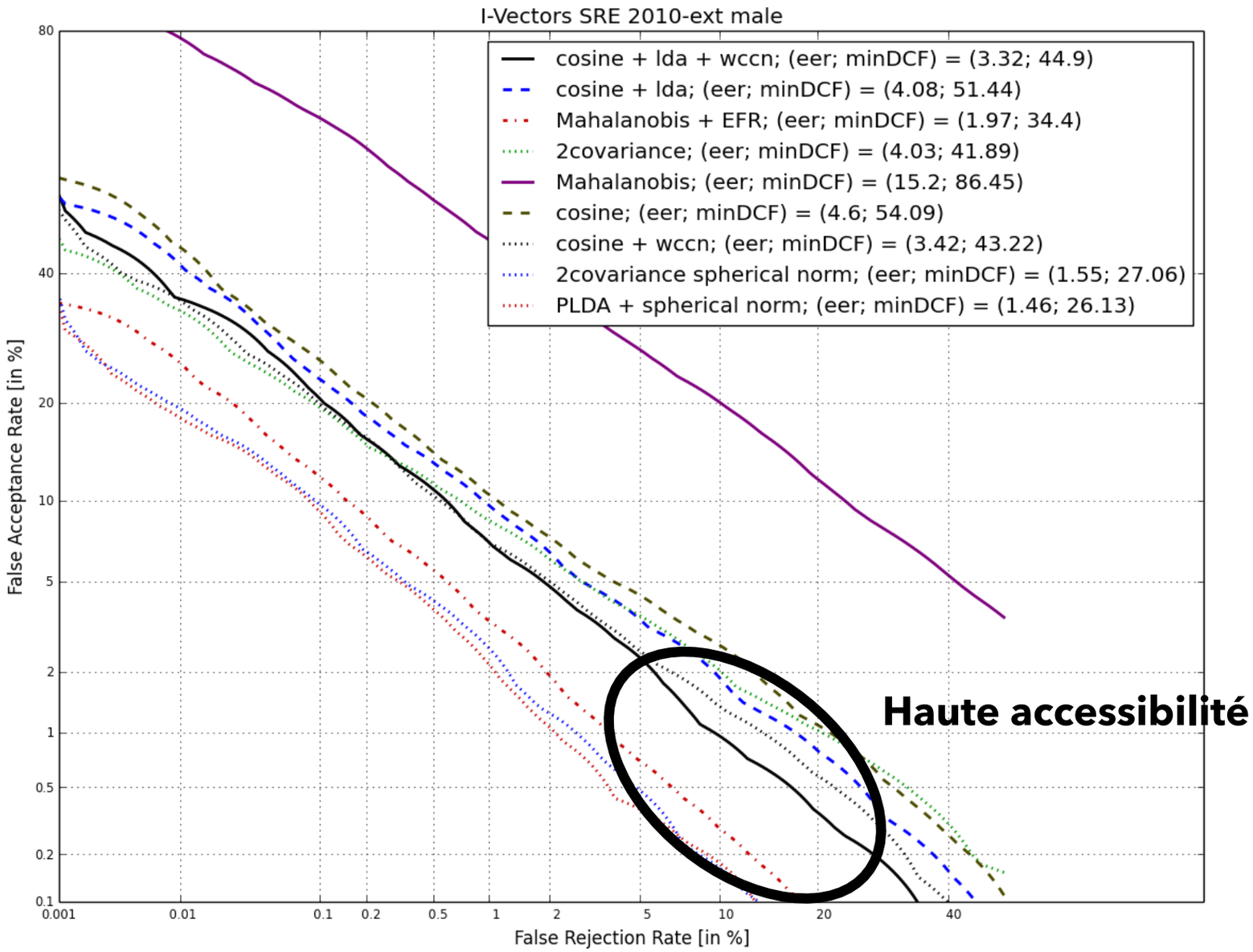
Courbe DET: Detection Error Trade-off

Haute sécurité



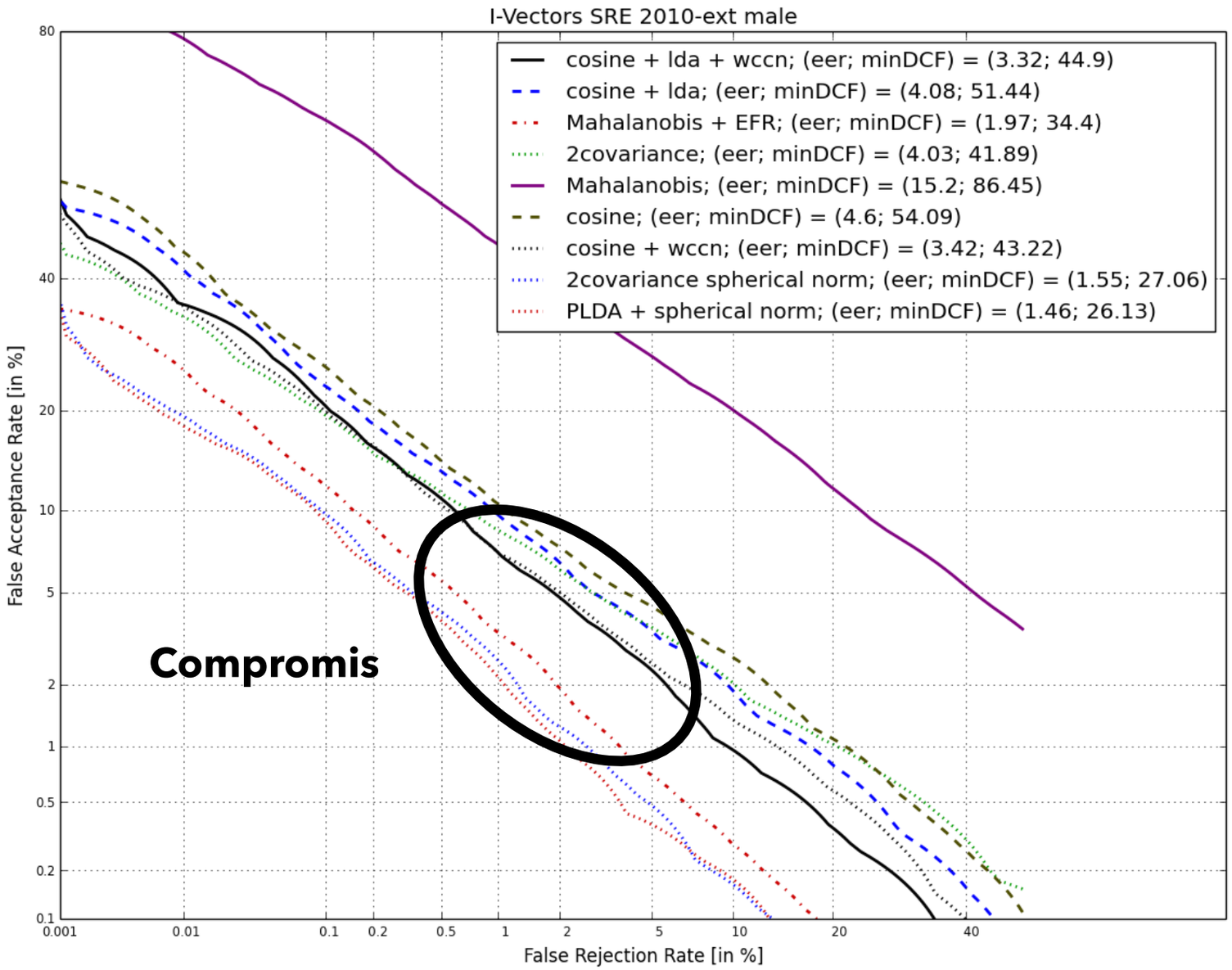
ANALYSE DES RÉSULTATS

Courbe DET: Detection Error Trade-off



ANALYSE DES RÉSULTATS

Courbe DET: Detection Error Trade-off



ANALYSE DES RÉSULTATS

Courbe DET: Detection Error Trade-off