

# CHAPITRE-2

## ∞ RECONNAISSANCE AUTOMATIQUE DU LOCUTEUR ∞ (RAL)

### Introduction

Nous allons maintenant nous attarder sur la notion de Reconnaissance Automatique du Locuteur. Nous parlerons brièvement de son historique, avant de commencer à exposer les généralités de cette science. Nous discriminerons les différents aspects qui constituent ce domaine d'actualité, et nous verrons comment en évaluer les performances. Nous présenterons enfin la structure globale d'un système de RAL.

### 1. GENERALITES

La reconnaissance de locuteur est l'une des branches de la reconnaissance vocale, avec la reconnaissance de la parole. La reconnaissance vocale, à son tour, s'inscrit dans l'axe plus vaste du traitement de la parole, avec d'autres spécialités telles que le codage de la parole, l'analyse de la parole et la synthèse de la parole. Le schéma suivant résume cette hiérarchie :

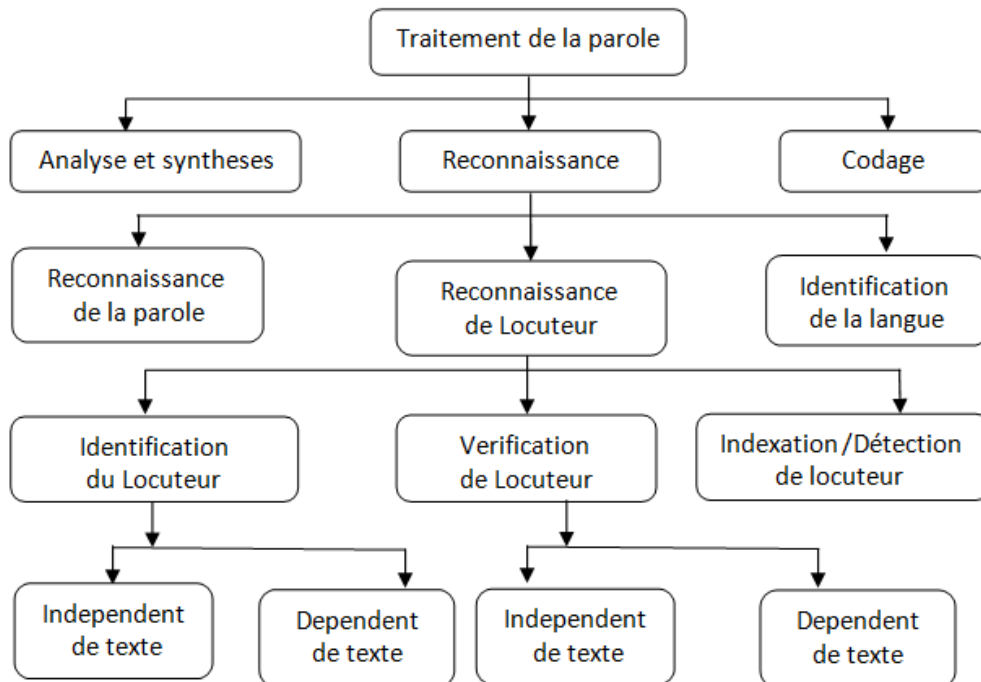


Figure 2.1 Hiérarchisation des domaines du traitement de la parole.

## 1.1 Définition de la RAL

Le terme "reconnaissance" est défini comme étant l'identification de quelque chose, sachant qu'on doit connaître au préalable le modèle de référence de cette chose. La Reconnaissance des formes consiste, alors, à identifier une forme donnée, après avoir déjà conçu le modèle de référence de ces formes. Cette phase s'appelle la phase d'apprentissage.

La reconnaissance automatique d'un individu consiste à utiliser des caractéristiques physiques dans le but de faire une discrimination entre les différents individus. Pour ce faire, plusieurs caractéristiques sont proposées dans la littérature : la photographie du visage, les empreintes digitales, les traits génétiques ou encore le signal de parole.

La caractéristique la plus pratique reste le signal de parole, vues la simplicité de son extraction et la possibilité d'offrir une authentification à distance.

L'authentification par la voix est appelée "Reconnaissance Automatique du Locuteur" (RAL). Cependant, ici nous rencontrons un problème majeur, qui est défini par la difficulté de trouver des caractéristiques pertinentes en discrimination : ce problème implique la nécessité de trouver des caractéristiques possédant une grande variabilité inter locuteur et une faible variabilité intra locuteur [Oua, 09].

## 1.2 Notions sur la Variabilité du Signal Vocal

La Reconnaissance Automatique du Locuteur est basée sur l'étude de la variabilité entre les caractéristiques extraites du signal de parole prononcé par le locuteur. Il existe donc une variabilité entre chaque signal de parole. Cette variabilité prend 3 formes différentes:

### 1.2.1 Variabilité intra-locuteur [Rou, 08]

elle identifie les différences dans le signal produit par une même personne. Cette variation peut résulter de l'état physique ou moral du locuteur. Une maladie des voies respiratoires peut ainsi dégrader la qualité du signal de parole de manière à ce que celui ci devienne totalement incompréhensible, même pour un être humain. L'humeur ou l'émotion du locuteur peut également influencer son rythme d'élocution, son intonation ou sa phraséologie.

### 1.2.2 Variabilité Inter-locuteur [Oua, 09]

La voix de chaque individu possède des qualités qui lui sont propres. L'âge, le sexe, le tempérament du locuteur (et bien d'autres facteurs encore) lui confèrent une identité vocale originale qui est la combinaison de multiples paramètres dont la hauteur (pitch), l'intensité et le timbre de sa voix, la qualité de son articulation, ou encore son accent (national et régional) ;

Le Coefficient de Wolf "CW" est le rapport de la variabilité inter locuteur sur la variabilité intra locuteur [J. Wolf]. Une grande variabilité inter locuteur : indique qu'on peut séparer, facilement, les locuteurs par leurs caractéristiques. Une petite variabilité intra locuteur : indique que chaque locuteur peut être représenté par une référence qui représente très bien ce locuteur.

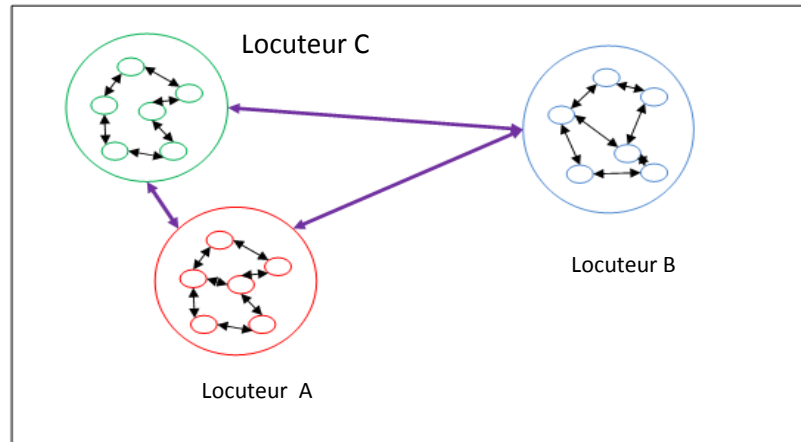
Par conséquent, un grand coefficient de Wolf indique, alors, que le paramètre choisi est pertinent en identification du locuteur et devrait donner un bon score de reconnaissance. Dans ce qui suit, nous définirons les différentes étapes associées à la RAL.

### 1.2.3 La variabilité due au contexte [Dem, 11]

La réalisation acoustique d'un son (phonème, mot,...) dépend de son environnement immédiat :

- les sons qui précèdent et ceux qui suivent influent fortement la production d'un son en raison de l'anticipation ou la rétention du geste articulatoire.
- Il s'en suit que la forme acoustique d'un son, et plus particulièrement ses zones transitoires sont dépendantes des traits articulatoires des sons adjacents.

Le schéma ci-dessous illustre la différence entre variabilité Inter-locuteur et variabilité intra-locuteur.



**Figure 2.2: Variabilités intra-locuteur et Inter-locuteur [Dem, 11].**

Dans ce schéma chaque locuteur est représenté par un cercle. Les petites flèches noires représentent la variabilité intra-locuteur, tandis que les flèches épaisses et violettes représentent la variabilité inter-locuteur.

### 1.3 Niveau de Dépendance au Texte

Une première classification des systèmes de RAL repose sur le niveau de dépendance au texte. En premier lieu, on distingue généralement les systèmes dépendants du texte des systèmes indépendants du texte.

En mode dépendant du texte, la reconnaissance d'une personne est réalisée sur la base d'un message dont le contenu linguistique (mot de passe, phrase...) est connu du système. En mode indépendant du texte, le système de reconnaissance n'a aucune connaissance sur le message linguistique prononcé par la personne.

Concernant le mode dépendant du texte, une terminologie plus fine peut être donnée à un système suivant l'application visée, systèmes à :

- messages fixés : la personne est contrainte de prononcer un message, qu'elle aurait au préalable (mots de passe personnalisés : [Jac, 00] et [Kha, 00]) ou qui sera imposé par le système.
- Messages prompts : un message, différent à chaque nouvelle session de reconnaissance, est imposé par le système sous forme visuelle ou auditive. Ces systèmes ont pour première motivation de se protéger des attaques de personnes

malveillantes (imposteurs) qui disposeraient d'un enregistrement de la voix d'une personne :

- unités segmentales : la personne doit prononcer un message comportant soit une séquence de mots (séquence de chiffres), soit des traits phonétiques (séquence de phonèmes) connus du système.

## 2. DIFFERENTES TACHES DE LA (RAL) ET LEURS APPLICATIONS

L'identification et la Vérification Automatique du Locuteur sont les tâches pionnières du domaine de la RAL, où plusieurs chercheurs sont en train de travailler : plus récemment, les besoins applicatifs ont fait naître de nouvelles tâches comme l'Indexation par Locuteur de flux audio (travaux de Johnson et de Delacourt) [Joh, 99] et [Del, 00] ou le Suivi de Locuteurs ou de nouvelles variantes telles que la détection du locuteur dans une conversation.

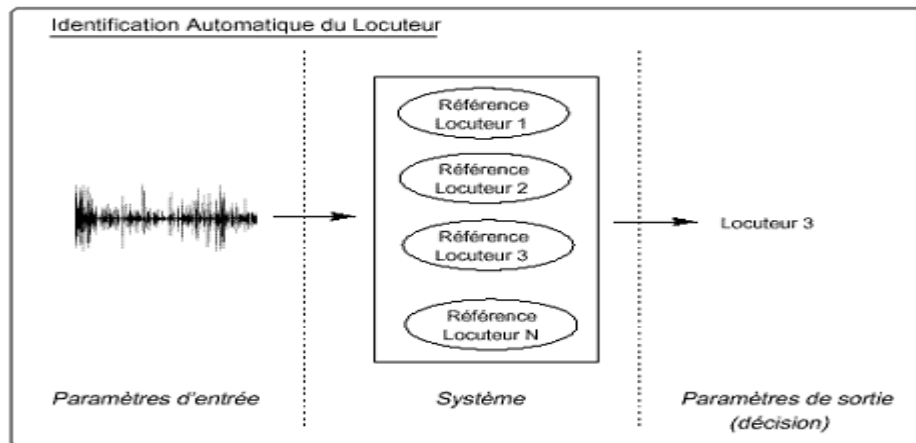
### 2.1 Identification Automatique du Locuteur (IAL)

L'identification Automatique du Locuteur (IAL) est le processus qui consiste à déterminer, parmi une population de locuteurs connus, la personne ayant prononcé un message donné. D'un point de vue schématique (figure 2.3), une séquence de parole est donnée en entrée du système d'IAL. Pour chaque locuteur connu du système, la séquence de parole est comparée à une référence caractéristique du locuteur : identité du locuteur dont la référence est la plus proche de la séquence de parole est donnée en sortie du système d'IAL.

Deux modes sont proposés en IAL, l'identification en ensemble :

- fermé pour lequel on suppose que la séquence de parole est effectivement prononcée par un locuteur connu du système ;
- ouvert pour lequel le locuteur peut ne pas être connu.

En mode "ensemble ouvert", le système d'IAL doit décider de la fiabilité de son jugement en acceptant ou rejetant l'identité qu'il a trouvée. De par son principe déterminer une identité parmi les identités potentielles les performances des systèmes d'IAL se dégradent généralement au fur et à mesure que la population de locuteurs augmente.



**Figure 2.3: Système d'identification de Locuteur [Fre 00]**

Ce schéma illustre la tâche d'identification automatique du locuteur. Le système ne prend qu'un seul paramètre en entrée (la voix de l'utilisateur) et la compare avec l'ensemble des voix stockées dans la base de données du système. Le système doit répondre en disant quel est le locuteur qui a parlé.

Il existe deux types différents d'identification automatique du locuteur

- L'identification en ensemble fermé : Dans laquelle on considère que l'utilisateur est toujours connu dans la base de données du système, c'est-à-dire que le système considère toujours qu'il connaît les caractéristiques vocales de l'utilisateur. Ce type d'identification est mis en œuvre lorsqu'on sait que le système ne sera utilisé que par un groupe restreint d'utilisateurs.
- L'identification en ensemble ouvert : Dans cette identification, n'importe quelle personne peut avoir accès au système. L'utilisateur peut donc ne pas être connu du système. Celui-ci doit alors être en mesure de distinguer les personnes qu'il connaît de celles qu'il ne connaît pas. [Say 03]

### ➤ Applications

En IAL, les applications sont peu nombreuses. On peut retenir, par exemple, l'utilisation d'un système d'IAL en vue de faciliter l'adaptation au locuteur des systèmes de RAP. Par ailleurs, il peut être intéressant pour des applications commerciales d'associer un même mot de passe pour une petite population de locuteurs (membres d'une famille, d'une société). Dans une telle situation, un système d'IAL en ensemble ouvert et dépendant du texte peut être utilisé pour contrôler l'accès à des données sensibles, à un réseau ou à un bâtiment [Ros 98].

## 2.2 Détection de Locuteurs

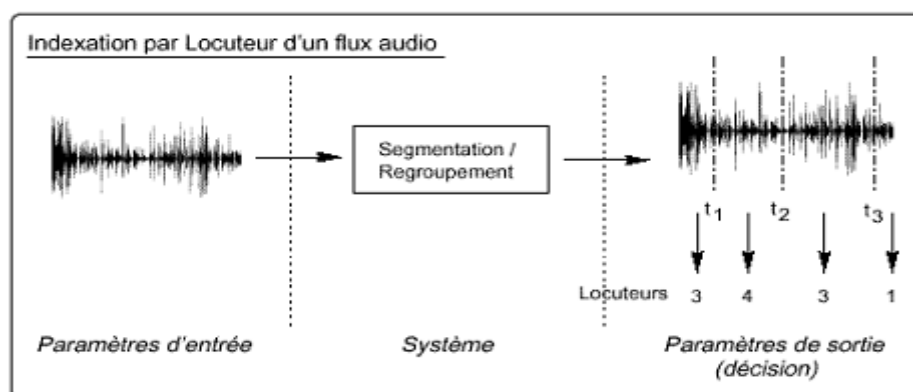
La détection de locuteurs dans un flux audio est une variante de la VAL. Sa particularité est de considérer un flux audio composé de séquences de parole produites par plusieurs locuteurs (conversations, débats, conférences, etc.). Dans ce contexte, la tâche de détection consiste à déterminer si un locuteur donné intervient ou non dans le document audio. Dans le cas d'un flux audio mono-locuteur, la tâche de détection se résume à la tâche de vérification.

### ➤ Applications

- le repérage automatique de messages issus de personnes suspectes dans des conversations téléphoniques (intérêt sécuritaire). Le système de détection adapté à un système de vérification par locuteurs, peut détecter si un locuteur donné intervient ou non dans une collection de documents audio enregistrées (ex : conversations téléphoniques enregistrées).
- la détection de locuteur sur un répondeur téléphonique ou sur une boîte vocale : avec l'explosion de la téléphonie, de plus en plus de services sont proposés. Parmi ces services, figureront peut être bientôt, le filtrage des messages d'un locuteur donné, déposés sur un répondeur téléphonique ou sur une boîte vocale [Oua, 09].
- recherche de journaliste présentateur afin d'extraire à partir de ces paroles des thèmes abordés dans le journal télévisé ;

## 2.3 Indexation par Locuteurs

La tâche d'Indexation Automatique par Locuteurs consiste à cibler les interventions des locuteurs dans un flux audio (figure 2.4). En d'autres termes, indexer un document audio en locuteurs revient à indiquer à quel moment un individu prend la parole et qui est cet individu. La seule entrée d'un système d'indexation est le document audio à indexer. Aucune information n'est donnée au système concernant le nombre de locuteurs présents dans le document ou leur identité.



**Figure 2.4 : Principe de base de la tâche d'Indexation par Locuteurs d'un flux audio**  
 [Fre, 00].

Contrairement aux systèmes d'IAL ou de VAL, les systèmes d'indexation ne détiennent pas de référence pour les locuteurs présents dans un document audio. Leur principe repose généralement sur une phase de segmentation "aveugle" en locuteurs suivie d'une phase de regroupement. Un système d'IAL permet finalement d'identifier les différents locuteurs présents dans le document.

La tâche de suivi de locuteurs peut être considérée comme une version simplifiée de l'Indexation par Locuteurs d'un flux audio (figure 2.5). Le principe reste le même : déterminer les interventions d'un ou plusieurs locuteurs, appelés locuteurs cibles, dans un flux audio. La simplification réside dans le fait que le système de suivi de locuteurs connaît nécessairement les locuteurs présents dans le document à indexer ou, du moins, ceux dont il doit détecter les interventions. Il possède une référence caractéristique pour chacun des locuteurs [Say.04].

Malgré cette simplification, le suivi de locuteurs reste une tâche très complexe. Trois grandes approches sont recensées dans la littérature :

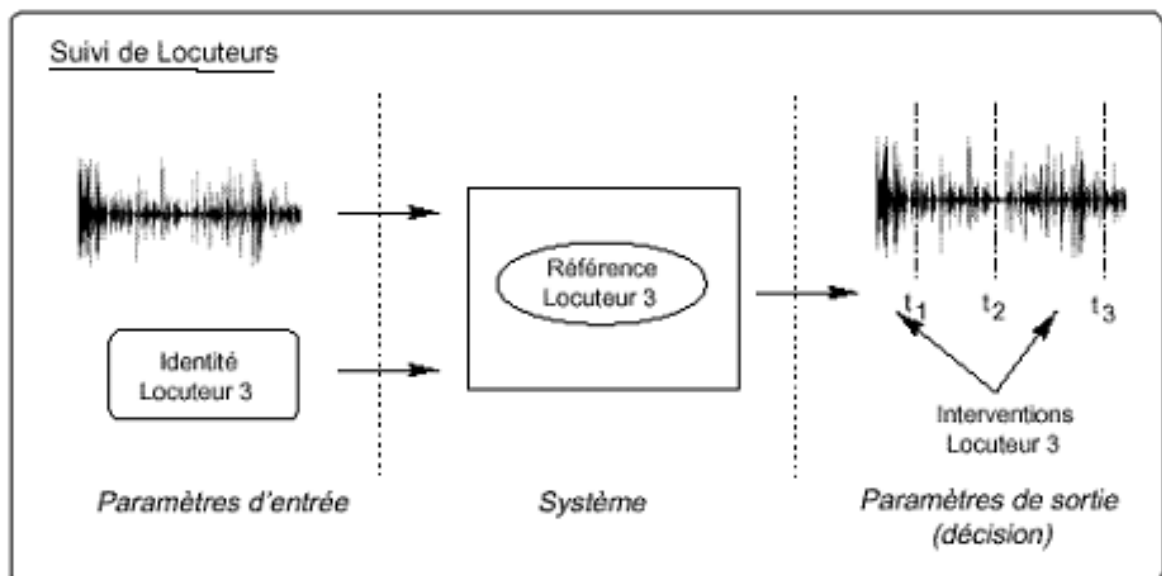
- une segmentation "aveugle" en locuteurs, identique à celle employée pour l'Indexation par Locuteurs d'un flux audio, est appliquée sur le signal de test. Les segments résultant de la segmentation sont soumis à un système de VAL classique afin de déterminer les segments appartenant effectivement au locuteur cible.
- le signal de test est découpé en une suite de blocs de trames, de taille fixe (ce découpage en blocs est entièrement indépendant des événements acoustiques observés sur le signal de parole), sur lesquels sont appliqués un système de VAL. Un processus de décision, à base de



seuils, permet en phase finale d'accepter ou de rejeter les blocs appartenant au locuteur cible [Ros, 98].

- la troisième approche est similaire à la précédente excepté pour le processus de décision. Dans ce cas, la décision repose sur un Modèle de Markov Caché, HMM ergodique composé d'états correspondant au locuteur cible, à un modèle générique de parole et de non parole (silence, bruit, etc.).

Les systèmes d'Indexation Automatique par Locuteurs d'un flux audio sont principalement utilisés pour le traitement de bases de données audio (recherche de séquences d'émissions télévisées ou radiophoniques par le suivi du présentateur, estimation du temps de parole de chaque intervenant lors de débats, etc.). D'autres applications sont envisageables comme la recherche de messages par locuteur sur un répondeur téléphonique ou sur une boîte vocale.



**Figure 2.5 : Principe de base de la tâche de suivi de locuteurs [Fre, 00].**

### ➤ Applications

Les systèmes d'Indexation Automatique par Locuteur d'un flux audio sont principalement utilisés pour le traitement de bases de données audio (recherche de séquences d'émissions télévisées ou radiophoniques par le suivi du présentateur, estimation du temps de parole de chaque intervenant lors de débats, etc.). D'autres applications sont envisageables

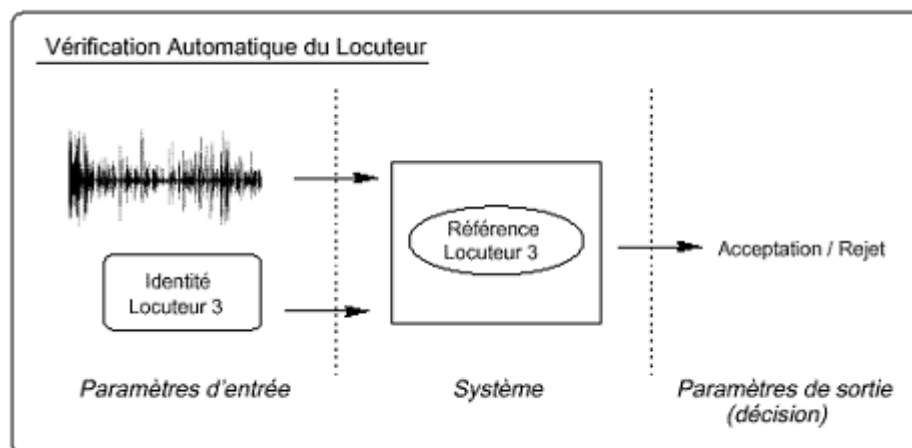
comme la recherche de messages par locuteur sur un répondeur téléphonique ou sur une boîte vocale.

## 2.4 Suivi de locuteurs

Comme nous l'avons évoqué précédemment, le processus de décision repose, pour la tâche de suivi de locuteurs, soit sur une décision classique de VAL (à base de seuils de décision), soit sur un décodage basé sur un modèle HMM.

## 2.5 Vérification Automatique du Locuteur

La Vérification Automatique du Locuteur (VAL) est le processus décisionnel permettant de déterminer, au moyen d'un message vocal, la véracité d'identité revendiquée par un individu (figure 2.6). L'identité ainsi que le message vocal constituent les deux entrées du système de VAL. L'identité, nécessairement connue du système, désigne automatiquement la référence caractéristique d'un locuteur. Une mesure de similarité est calculée entre cette référence et le message vocal puis comparée à un seuil de décision. Dans le cas où la mesure de similarité est supérieure au seuil, l'individu est accepté. Dans le cas contraire, l'individu est considéré comme un imposteur et rejeté.



**Figure 2.6 : Principe de base de la tâche de Vérification Automatique du Locuteur**

[Fre,00].

Sur le schéma ci-dessus on voit qu'un système d'identification automatique de locuteurs prend en entrée le signal de parole et l'identité présumée du locuteur. Le système compare alors la voix de l'utilisateur avec la référence de ce locuteur (caractéristiques stockées en mémoire). La sortie de ce système consiste en une réponse du type « oui ou non ».

### ➤ Applications

Les applications de VAL sont multiples et principalement commerciales [Bov, 98] :

- Serrures vocales pour le contrôle d'accès à des locaux ;
- Authentification pour l'accès à distance à des données sensibles ou à des services spécifiques à travers le réseau téléphonique (consultations ou transactions bancaires, consultations de bases de données à caractère confidentiel, consultations de boîtes vocales, télé-achat, etc.) ;
- Protection de matériel contre le vol (téléphones portables, voitures, etc.) ;
- Incarcération à domicile nécessitant une authentification régulière du prévenu.

### ➤ Décision classique de VAL

Le processus de décision compare la mesure de similarité obtenue sur chaque bloc ou segment à un seuil de décision. Si la mesure de similarité est supérieure au seuil, le bloc ou segment est attribué au locuteur cible. Dans le cas contraire, le bloc ou segment est rejeté.

## 3 EVALUATION DES SYSTEMES RAL

A l'heure actuelle une seule grande campagne d'évaluation est dédiée aux systèmes de RAL. Cette campagne, organisée chaque année par l'institut américain NIST (National Institute of Standards and Technologies) depuis 1996, est accessible aux laboratoires de recherche publics ou privés. Basées initialement sur l'évaluation des systèmes de VAL dans un contexte conversationnel et téléphonique, ces campagnes se sont rapidement étendues aux tâches de détection d'un locuteur dans une conversation, de suivi de locuteurs, et plus récemment d'Indexation Automatique par Locuteur d'un flux audio. Avant chaque nouvelle campagne d'évaluation, un corpus de développement est défini et distribué aux participants, permettant la mise au point des systèmes de RAL.

Pour chacune des tâches de Reconnaissance Automatique du Locuteur, on peut évaluer les performances du système en mesurant différents taux :

### 3.1 Erreurs d'Identification

#### 3.1.1 En ensemble ouvert

On calcule :

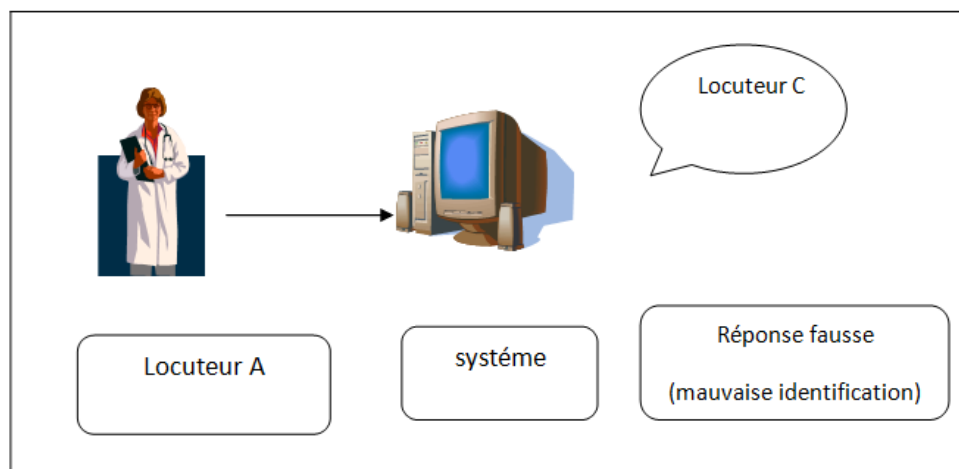
- Taux de mauvaise reconnaissance : Nombre de fois que le système confond un locuteur de la base de données pour un autre.
- Taux de faux rejet : Nombre de fois que le système rejette un utilisateur alors que celui-ci est supposé être reconnu par le système.
- Taux de fausse acceptation : Nombre de fois que le système accepte un utilisateur étranger à la base de données.

#### 3.1.2 En ensemble fermé

On ne mesure que le taux de mauvaise reconnaissance.

### 3.2 Erreurs de Vérification [Dem 11]

On mesure le taux de faux rejet et le taux de fausse acceptation. Les différentes erreurs commises par les systèmes de vérification et d'identification sont résumées par les schémas suivant :



**Figure 2.7: Erreur de mauvaise identification.**

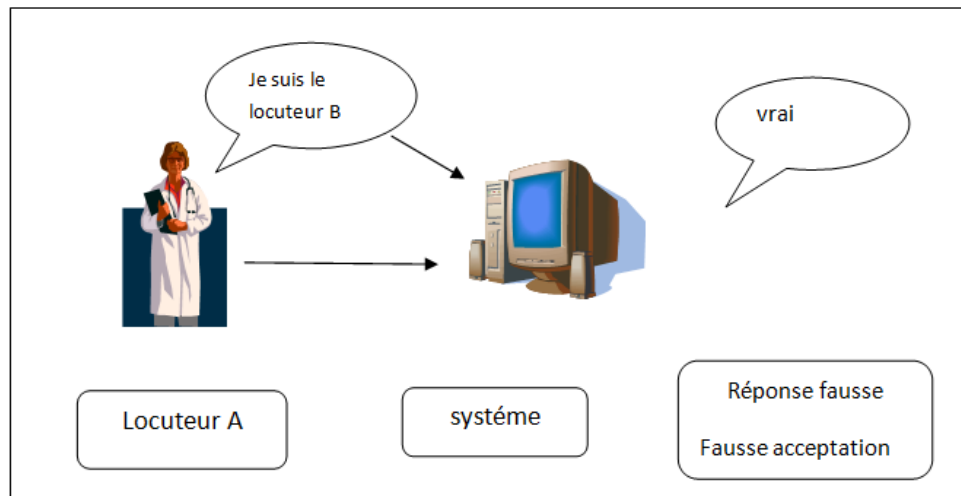


Figure 2.8: Erreur de Fausse acceptation

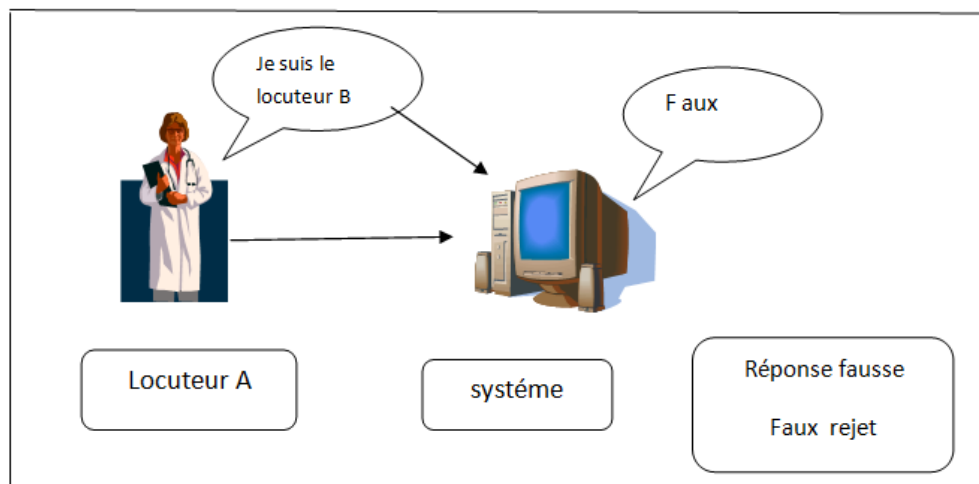


Figure 2.9: Erreur de faux rejet

#### 4 STRUCTURE GENERALE D'UN SYSTEME RAL

La tâche de reconnaissance automatique du locuteur peut se subdiviser en trois étapes: Paramétrisation, classification et décision. Un premier module de traitement du signal réalise l'analyse acoustique du signal de parole. A l'issue de cette étape, le signal est représenté par des vecteurs de coefficients, ce qui permet de réduire l'information en quantité et en redondance. Ces vecteurs sont éventuellement représentés par un modèle mathématique, on parle alors de méthodes paramétriques. Dans la phase de classification, les vecteurs du signal de test (ou leur modèle) sont comparés aux vecteurs des locuteurs de référence (ou à leurs modèles). La phase de décision désigne le locuteur finalement reconnu. Le processus de reconnaissance, par exemple, est différent selon qu'il repose sur une modélisation des

caractéristiques des locuteurs connus du système (modèles clients pour les tâches d'IAL, VAL, DAL et SAL) ou non (Indexation par locuteur d'un flux audio).

Les paramètres de l'analyse spectrale sont généralement répartis suivant une échelle linéaire ou une échelle Mel (permettant une plus grande résolution dans les basses fréquences). Récemment, [Gri, 95] à expérimenté une technique d'analyse spectrale à résolution variable. Cependant, les différentes résolutions spectrales utilisées avec différents classificateurs n'ont pas conduit à une amélioration significative des résultats ; les classificateurs ayant un comportement quasi indépendant des ensembles de paramètres. Dans le même sens, [Auc, 06] propose d'estimer une échelle fréquentielle optimale en égalisant les taux d'erreur sur différentes sous-bandes fréquentielles. Les résultats préliminaires obtenus semblent montrer la sous-optimalité des échelles linéaires et Mel, mais l'estimation d'une échelle optimale reste empirique et fort dépendante des conditions expérimentales.

## **Conclusion**

Dans ce chapitre nous avons étudié en détail les différents aspects de la reconnaissance automatique du locuteur. Dans notre travail nous utiliserons une identification en ensemble fermé (nous savons à l'avance qui est présent dans la conversation) et indépendante du texte. La méthode de discrimination que nous utilisons est basée sur les machines à vecteurs de support, que nous détaillons dans le chapitre suivant.