TABLE DES MATIÈRES

1 INTRODUCTION

1.1 PROBLÉMATIQUE ET BESOINS

Afin de situer la problématique à laquelle l'expert dans le domaine de la reconnaissance des locuteurs en sciences forensiques est confronté, et dans laquelle ce travail s'intègre, voici une situation:

Imaginons une victime d'appels anonymes. Cette personne enregistre ces appels sur un support – enregistreur analogique/numérique – et le remet à la police. Dès lors, un échantillon de la voix de l'auteur est à disposition de la Justice. Suite au dépôt d'une plainte de la part de la personne lésée, une enquête est ouverte. Celle-ci aboutit rapidement à l'interpellation d'un suspect. Des enregistrements de la voix du suspect sont réalisés avec ce dernier pour servir de matériel de comparaison dans le cadre de l'affaire.

Une fois la voix de l'auteur des appels anonymes et celle du suspect à disposition, il s'agit d'entreprendre un processus comparatif afin de déterminer si le suspect appréhendé peut être l'auteur des appels en questions.

Ce processus comparatif, sur lequel nous reviendrons au chapitre 2.1.3, crée certains besoins. En particulier:

- Il est nécessaire de posséder un système de reconnaissance automatique performant capable d'extraire les paramètres caractéristiques des voix en question (auteur et suspect), de créer un modèle de la voix du suspect et de le comparer avec les paramètres caractéristiques extraits de la trace (voix de l'auteur). A cette fin, dans le cadre de ce travail, le système ASPIC (Automatic SPeaker Identification by Computer) a été utilisé. Ce dernier sera présenté dans le chapitre 2.

- En outre, il est nécessaire d'avoir à disposition des bases de données qui permettent non seulement de vérifier la performance du système utilisé, mais également d'appréhender le processus comparatif entre la voix de l'auteur et celle du suspect d'un point de vue forensique.

Il peut-être signalé que le FBI (Federal Bureau of Investigation) est le seul à posséder une base de données suffisamment grande et performante pour permettre une approche forensique complète de la reconnaissance automatique des locuteurs. Nous reviendrons sur cette base de données dans le chapitre 3.

Les besoins énumérés et la remarque faite ci-dessus présentent les conséquences suivantes pour notre travail:

- Au niveau du système automatique: celui-ci était à disposition.
- Au niveau des bases de données: un manque évident s'est fait sentir. Il a donc fallu créer les bases de données adéquates à l'utilisation forensique de la reconnaissance des locuteurs et en particulier en ce qui concerne les conditions d'enregistrements .

1.2 QUELQUES RÉFLEXIONS SUR LA VOIX

La reconnaissance des locuteurs repose sur le postulat de l'unicité de la voix basé lui-même sur l'expérience quotidienne de la reconnaissance de ses proches au son de leur voix, et sur les travaux de plusieurs scientifiques au tournant du XX^{e} siècle.

En effet, c'est à cette époque que le monde scientifique commence à s'intéresser à cette donnée biométrique. En particulier, on notera les travaux de Messieurs Bertillon, Ottolenghi et Locard car ils ont mis en évidence la complexité du signal de parole et de son traitement.
Le premier a senti et affirmé que la voix devait répondre au principe de l'individualité mais que l'oreille humaine ne possédait pas les finesses et les subtilités nécessaires à percevoir dans la voix les caractéristiques répondant à ce principe. Dès lors, l'humain devait être incapable de la décrire correctement. "Le timbre de la voix est l'un des caractères les plus distinctifs de l'individualité […] malheureusement aucun signe n'est plus difficile à noter" (Bertillon, 1881).

Le second, médecin de son art, a compris que la voix était un signal complexe qui, pour être décrit correctement, devait être appréhendé sous plusieurs facettes. "De la voix, il faut considérer le volume ou la force, la hauteur du son ou le ton, la qualité du son ou le timbre, l'agilité, le type et l'intonation" (Ottolenghi, 1910). Il met encore en évidence le fait que la voix varie en fonction de l'âge, du sexe, du message prononcé et de toutes les atteintes physiologiques en relation avec les organes de la production ou du traitement de la parole.

Enfin le troisième précise qu'une victime, qu'un auteur ou qu'un témoin engagé dans une action délictueuse se trouvera dans un état émotionnel tel que tant la production que la perception du signal de parole se trouveront altérées. "Un homme qui menace ou qui frappe, un autre qu'on égorge, ne parle pas sur le ton du commérage quotidien […] d'autre part celui qui écoute est vraisemblablement fort troublé […]. Or l'émotion a pour résultat de troubler la perception et de rendre les souvenirs non seulement imprécis, mais informes et inexacts." et "distinguer une voix est une opération physique complexe qui comporte la perception des trois éléments: hauteur, intensité et timbre, leur comparaison avec des éléments identiques déjà perçus, est l'affirmation de cette identité" (Locard, 1932).
De ce fait, la méthode de reconnaissance des locuteurs par audition doit nécessairement être subjective. Il est ainsi nécessaire de tenir compte des conditions dans lesquelles le message a été non seulement prononcé, mais aussi entendu.

Tout au long du XX^{e} siècle et aujourd'hui encore, les problèmes de base liés à la reconnaissance des locuteurs sont restés les mêmes. Cependant, les techniques pour les appréhender ont connu des évolutions plus ou moins significatives au cours du temps.
Les étude des trois scientifiques cités ci-dessus illustrent merveilleusement la difficulté d'appréhender la voix comme donnée biométrique et bien d'avantage encore la complexité du traitement de la parole dans le cadre de la reconnaissance des locuteurs et à fortiori de l'identification d'une personne sur cette base.
Car, s'il est admis que la voix possède une qualité d'individualité, personne n'a jamais réussi à le démontrer. Les caractéristiques individuelles et immuables devant caractériser toute donnée biométrique devant par essence être unique n'ont, à ce jour, jamais pu être identifiées, isolées et extraites d'un signal de parole.

2 PARTIE THÉORIQUE

2.1 LA VOIX COMME DONNÉE BIOMÉTRIQUE

Au niveau de la production de la voix, deux processus centraux doivent être distingués: la phonation et l'articulation.

Les cordes vocales, au nombre de deux, ne sont rien d'autre qu'un ensemble de muscles et de ligaments recouverts par une muqueuse et attachés aux cartilage thyroïdes et arythénoïdes. La position et la longueur des cordes vocales peuvent être modifiées par des mouvements de pivotement et de déplacement latéral sous l'action des cartilages cités ci-avant. L'espace entre les cordes vocales est la glotte. Elle est une sorte de valve entre la trachée et le conduit vocal.
Lorsque de l'air est expulsé par les poumons, un courant d'air va circuler à travers la trachée jusqu'à la glotte au niveau du larynx. Ce dernier est une structure composée de cartilages, de

ligaments et de muscles qui sont placés juste sous l'os hyoïde et cela de manière plus basse pour l'homme que pour la femme. Ce mouvement de l'air est **la phonation.**
Lorsque la glotte est fermée, une différence de pression existe de part et d'autre de la glotte. Cette différence de pression entraîne l'ouverture de la glotte (déplacement des cordes vocales dont la rigidité n'est pas totale). Par les suite, des phénomènes aérodynamique et biomécanique permettent de resserrer les cordes vocales.

Les tensions variables (résultant en des allongements variables des cordes vocales) pouvant être exercées sur lesdites cordes vont permettre de générer des sons aigus (forte tension) ou graves (faible tension). Ainsi, la tension appliquée influence la fréquence de vibration et donc la fréquence du son périodique généré. L'épaisseur des cordes vocales (généralement plus épaisses chez l'homme que chez la femme) a une influence sur la fréquence fondamentale pouvant être générée. Plus l'épaisseur est importante, plus des fréquences basses pourront être produites.
Cependant, la vibration des cordes vocales n'est pas une constante dans le mécanisme de phonation.
En effet, des lettres comme le 's' ou le 'f' se produisent en l'absence de vibrations. On parle alors de son non voisés par opposition aux sons voisés telles ceux produits par la lettre 'z' ou 'v'.
Dans le premier cas de figure, le son est produit par le frottement de l'air dans le conduit vocal entre la glotte et les lèvres alors que dans le second cas, les vibrations des cordes vocales modulent le flux de l'air issu des poumons. Les oscillations périodiques sinusoïdales des cordes vocales entraînent une onde sonore également périodique sinusoïdale qui se propage dans le conduit vocal jusqu'aux lèvres.
La fréquence d'oscillation se situe entre 150Hz et 200Hz pour les hommes, entre 200Hz et 500Hz pour les femmes et entre 300Hz et 450 Hz pour les enfants.

Le conduit vocale est la cavité, non rigide, remplie d'air, qui s'étend de la glotte aux lèvres. La forme de la cavité peut être modifiée par le mouvement des articulateurs. Le conduit vocal se compose du pharynx, de la cavité orale, de la cavité labiale et de la cavité nasale.

Un grand nombre de muscles sont mis en jeu lors du processus de phonation. Ces muscles exercent le contrôle sur les articulateurs dont les principaux sont la langue, le vélum, les dents, les lèvres et la mâchoire inférieure. Les **articulateurs** permettent de modifier la forme du conduit vocal et ainsi de moduler les sons produits.

Ainsi:

- La phonation désigne le mécanisme de production des sons du langage.
- Un processus expiatoire permet d'amener l'air jusqu'au larynx.
- La vibration des cordes vocales et/ou le frottement de l'air dans le conduit vocal sont responsables des sons émis.
- Le conduit vocal filtre les sons émis au niveau du larynx.
- Les articulateur, par la déformation de la forme du conduit vocal, permettent de moduler les sons produits.

Les éléments présentés dans ce chapitre nous permettent de comprendre que le système phonétique-articulatoire possède une individualité anatomique certaine. Cependant, le reflet de cette individualité, couplé à la dimension comportementale de la production et de la perception de la voix, rendent l'extraction et l'isolement de caractéristiques propres à décrire et à modéliser cet état de fait bien difficiles.

2.2 LA VOIX EN SCIENCES FORENSIQUES

Du point de vue pénal, ce sont principalement les articles 179^{septies} et 179^{octies} du Code pénal qui présentent un intérêt dans le domaine de la reconnaissance des locuteurs en sciences forensiques. Ils répriment respectivement l'utilisation frauduleuse d'une installation de télécommunication et les mesures officielles de surveillance.

Ainsi, c'est la voix transmise par le biais d'un réseau de télécommunication et enregistrée sur un support, qu'il soit analogique ou numérique, qui constitue le plus souvent l'indice.

L'enregistrement direct sur un appareil enregistreur de type dictaphone peut également se rencontrer comme trace dans le domaine mais cela est plutôt rare. En effet, la plupart des affaires soumises à expertise sont des traces obtenues selon ce qui est décrit dans les deux articles cités ci-dessus.

Entre le moment ou la voix est générée par le locuteur et celui ou elle est enregistrée sur un support, un certain nombre de transformations interviennent. Au niveau du combiné du téléphone (de l'appelant et de l'appelé), de la ligne de transmission du signal et du support d'enregistrement.

Le microphone du téléphone transforme l'énergie acoustique de la voix en énergie électrique. Cette transformation est assurée par une membrane souple que l'onde sonore fait vibrer. Les vibrations de la membrane sont converties en une tension électrique alternative par un dispositif qui dépend de la technologie du microphone. Cette tension est alors acheminée vers le système d'amplification ou d'enregistrement auquel le micro est branché. La surface et la tension de la membrane ainsi que la sensibilité de la bobine au champ magnétique de l'aimant déterminent la gamme de fréquence pouvant être captée par le micro. Dans un combiné de téléphone, le microphone est généralement de type électrostatique – fondé sur le principe de condensateur dans lequel les ondes sonores font varier la distance entre les armatures – mais avec la particularité de posséder un matériau à polarisation permanente, l'électret.
(Ribary, 2002) a notamment étudié l'influence du microphone sur le système de reconnaissance automatique ASPIC. Il est apparu que plus la qualité du microphone était faible, plus les scores obtenus avec le systèmes automatiques étaient bas. Dans les expérimentations qu'elle a mené, la moyenne des scores diminuait de manière continu avec la qualité descendante des microphones utilisés.

Une fois l'énergie acoustique de la voix transformée en énergie électrique, celle-ci est convertie en un signal numérique.

Sur le Réseau Téléphonique Public Commuté (RTCP) le signal est numérisé avec une fréquence d'échantillonnage de 8KHz sur 8bits, soit un débit de 64 $Kbits^{-1}$.
Lorsque deux correspondants communiquent, les signaux vocaux envoyés et reçus sont dus à une modulation de l'amplitude du courant continu dans la bande de fréquences 300Hz à 3.4Khz. La communication est bidirectionnelle et le courant sur la ligne est la somme du courant continu et des deux courants variables émis par chacun des postes. L'Union Internationale de Téléphonie (International Telephone Union – Telephony division) a édicté en 1991 une norme (G.728) utilisant l'algorithme Low Delay Excited Linear Prediction Coder (LD-CELP) qui permet de coder le signal avec un débit binaire de 16 $Kbits^{-1}$. Ce système de codage est basé sur une technique alliant une modélisation à une quantification vectorielle. L'avantage de cet algorithme est de posséder un faible taux de reconstruction.

Le réseaux cellulaire européen repose sur la norme Global System for Mobile communication (GSM) depuis 1989.
Le codeur/décodeur de voix intégré au téléphone portable permet la transmission numérique de la voix sur l'interface radio GSM. Le signal analogique du microphone est échantillonné avec une fréquence de 8KHz puis transformé en un signal numérique. Le signal est envoyé au codeur de voix qui le code à un débit de base de 13 $Kbits^{-1}$. L'interface radio transmet le signal en y ajoutant le codage du canal avec un débit binaire brut de 22.8 $Kbits^{-1}$. Le système GSM est exploité dans les bandes de fréquences de 900MHz (GSM900) et 1800MHz (GSM1800). La largeur de bande (35MHz pour GSM900 et 75MHz pour GSM1800) est assurée aussi bien dans la liaison *uplink* (téléphone => station) que *downlink* (station => téléphone).
La technologie actuelle GSM permet des débits de transfert de données pouvant atteindre les 9.6 $Kbits^{-1}$.
Une nouvelle norme de téléphonie cellulaire est en train de faire lentement son apparition sur le marché. Il s'agit de la norme Universal Mobile Telecommunications System (UMTS) développée en Europe. Le débit peut atteindre 2 $Mbits^{-1}$. Le système UMTS utilise une interface radio ou aérienne

(UMTS Terrestrial Radio Access (UTRA)). L'avantage de l'UMTS et de permettre à tous les usagers d'un réseau de travailler sur la même fréquence grâce à la méthode de transmission Code Division Multiple Access (CDMA). Les canaux sont séparés au moyen d'un code. Le signal de donnée de chaque usager est multiplié auprès de l'émetteur et du récepteur par un code unique déterminé par une communication donnée. Dans le système GSM, les usagers sont séparés les uns des autres par plusieurs fréquences ou intervalles de temps.
L'exploitation Frequency Division Duplex (FDD) utilise deux fréquences séparées pour les communication *uplink* et *downlink*. L'exploitation Time Division Duplex (TDD) n'utilise en revanche qu'une seule fréquence.
Les systèmes UMTS travaillent dans la bande de fréquence de 2 GHz. Le téléphone portable émet dans une bande de 1920MHz à 1980MHz alors que la station de base émet, elle, dans une bande de 2110MHz à 2170MHz. Un seul canal dispose d'une bande de larguer de 5MHz. Il existe 4 canaux (20 MHz) dans la bande partielle TDD et 2 * 12 canaux (2*60 MHz, le 2 signifie que deux fréquences différentes sont utilisées pour le up et downlink) dans la bande partielle FDD. Les canaux UMTS sont ainsi bien plus larges que les canaux GSM.

Au final, il apparaît que la qualité du codage du signal de parole dans le réseau cellulaire est inférieur à celle du codage dans le réseau fixe.

(Jayant, 1992) a mis en évidence le fait que dans une transmission téléphonique, les dégradations du signal de parole dépendaient non seulement du type de réseau utilisé (RTPC ou GSM) mais également du système de transmission (analogique, numérique ou combiné) des informations dans le réseau.
D'une manière générale, les algorithmes de codage numériques sont élaborés pour la transmission du signal de parole. En ce sens, la seule contrainte pour les opérateurs est de conserver l'intelligibilité du signal. Ainsi, de nombreuses distorsions non linéaires (bruit de fond) interviennent dans le signal qui ne peuvent être modélisées analytiquement. Ce bruit de fond est plus caractéristique du signal transmis par l'intermédiaire d'un téléphone mobile du fait de son utilisation possible dans tout type d'environnement.
(Meuwly, 2001) a étudié l'influence du canal de transmission sur le système de reconnaissance automatique ASPIC. Il a conclu que "l'utilisation d'un réseau cellulaire pour l'enregistrement du modèle à la place d'un réseau téléphonique commuté altère les performances du système de reconnaissance automatique de locuteurs, que les enregistrements de tests proviennent du réseau commuté ou du réseau cellulaire" et que "les performances sont très nettement diminuées lorsque l'enregistrement du test provient du réseau cellulaire (GSM) et le modèle du réseau téléphonique public commuté (RTCP)".

A la fin de la chaîne, dans le combiné récepteur, le signal numérique est retransformé en signal analogique et l'énergie électrique résultante est retransformée par le haut parleur (à la manière du microphone) en énergie acoustique audible pour la personne réceptrice du signal. Ce signal peut ensuite être enregistré sur un support analogique/numérique.
L'influence du type d'enregistrement sur le système de reconnaissance a notamment été étudiée par (Meuwly, 2001 et Rossy, 2002). Les résultats obtenus montrent que la comparaison de traces enregistrées analogiquement à des modèles enregistrés numériquement (et réciproquement) ne devrait pas être entreprise dans un cas réel.

Nous voyons donc que toute une série de transformations du signal de parole intervient lors de la transmission dudit signal. La conséquence sur la reconnaissance automatique de locuteurs en sciences forensiques en est, dès lors, que seules la maîtrise et la compréhension de l'intervention de chacun de éléments de la transformation peuvent assurer une interprétation correcte de la valeur finale de la force probante de l'indice.

2.3 SYSTÈMES DE RECONNAISSANCE AUTOMATIQUE DE LOCUTEURS

Dans ce chapitre seront abordés brièvement les développements techniques et idéologiques en matière de reconnaissance de locuteurs. Le traitement auquel un échantillon de voix et soumis et la façon dont les comparaisons s'opèrent entre un test et un modèle, dans une procédure de reconnaissance automatique, seront exposés.

2.3.1 BREF APERÇU HISTORIQUE

La reconnaissance des locuteurs est décrite par (Atal, 1976) comme étant "tout processus de décision qui utilise quelques caractéristiques du signal de parole pour déterminer si une personne particulière est l'auteur d'un énoncé donné."

Sa forme la plus ancienne est l'analyse auditive exercée au début uniquement par des profanes (victime ou témoin) et par la suite également par des experts (approche phonétique auditive perceptive et les techniques de phonétiques acoustiques; voir à ce sujet notamment (Künzel, 1987 et Braun, 1995)).
Au début des années 40 a été mis au point au *Bell Telephone Laboratories* le spectrographe sonore, sur la base des travaux de (Steinbreg, 1934), considéré comme l'instrument d'analyse fondamental de la voix. Dans les années 60, une version dotée d'un système de lecture continu des enregistrements a été développée et appliquée au domaine de l'identification (Kersta, 1962 A et B). Cependant, dès le début des années 70, cette technique a subi les critiques et controverses permanentes de nombreux scientifiques et a généré d'incessants débats. Il apparaît aujourd'hui que cette méthode de comparaison visuelle de spectrogrammes vocaux ne peut être considérée comme valide et ne peut être acceptée dans le monde des sciences forensiques (Meuwly, 2001).

Actuellement, la tendance est à l'approche automatique globale, dont les premiers balbutiements remontent aux années 60 ((Pruzanski, 1963) notamment), qui utilise très largement l'outil informatique et l'intelligence artificielle qui l'accompagne. La technique est automatique car toute analyse ou évaluation subjective du signal est réduite à un minimum et elle est globale car le signal est traité comme un phénomène physique i.e. comme une vibration complexe variant dans le temps.
(Bunge, 1991) décrit cette technique comme" l'étude de la capacité de l'outil informatique à procéder à la reconnaissance de personnes à partir d'une donnée biométrique variable, la voix, sur la base de méthodes exploitant la théorie de l'information, la reconnaissance automatique de forme et l'intelligence artificielle perceptive."
Au cours des dernières années de nombreux systèmes usant de la reconnaissance des locuteurs, principalement dans des tâches de vérification comme dans les contrôles d'accès à des lieux sensibles ou à de l'information classifiée ou encore dans des applications liées aux interactions entre l'homme et la machine ont été développés.
Dans le domaine judiciaire, le développement est plus lent et souffre du manque d'intérêt qui caractérise cette application moins commerciale. (Champod & Meuwly, 2000) soulignent que "la méthodologie et le rôle de l'expert dans le domaine de la reconnaissance automatique en criminalistique n'ont reçu de l'attention que très récemment".

Quoi qu'il en soit, il convient de se souvenir, comme le rappelle (Künzel, 1994), que même dans un système automatique, une part de subjectivité demeure puisqu'à un moment ou un autre le facteur humain intervient.

2.3.2 SYSTÈME UTILISÉ (ASPIC)

Le système automatique ASPIC (Automatic SPeaker Identification by Computer) a été utilisé pour ce travail de recherche. Ce système a été développé dans le cadre d'une collaboration entre l'Institut de Traitement des Signaux de l'Ecole Polytechnique Fédérale de Lausanne et l'Ecole des Sciences Criminelles/Institut de Police Scientifique de l'Université de Lausanne. Le fonctionnement est décrit dans les paragraphes suivants.

2.3.2.1. ACQUISITION DU SIGNAL

L'acquisition du signal se fait sur un support enregistreur après que, la plupart du temps, le signal soit passé à travers une ligne de téléphonie comme cela a été présenté dans le chapitre précédent.
Ainsi, le signal est un échantillon de voix codé de manière analogique ou numérique; il est la trace d'intérêt.

2.3.2.2. AUDITION DU SIGNAL

Suite à l'acquisition du signal, intervient une phase d'audition par l'expert. Cette phase est primordiale. En effet, elle permet non seulement de se déterminer sur le degré d'exploitation possible du signal, et donc sur la nécessité d'entreprendre un travail d'expertise, mais également de déterminer quelles sont les bases de données adéquates à sélectionner afin que la congruence entre la trace et la population d'auteurs potentiels soit assurée. Cela du point de vue des conditions d'enregistrement, de la langue, du mode d'élocution ou encore de la quantité de signal à disposition (une dizaine de secondes semble être un minimum). Certaines fois, l'audition permet d'orienter l'enquête par la mise en évidence, avec une certaine probabilité, notamment du sexe de l'auteur de l'énoncé ou encore de son origine ethnique.

2.3.2.3. PRÉTRAITEMENT DU SIGNAL

Le pré traitement du signal consiste à supprimer les zones de silence afin de diminuer la probabilité d'extraire et de modéliser des paramètres qui ne sont pas représentatifs de la voix de l'auteur de l'énoncé de question mais qui sont la conséquence du bruit de fond et notamment du "bruit de la ligne".
Une application (appelée SILREM) basée sur l'algorithme de Murphy permet, au moyen d'un seuil adaptatif de séparer les zones de fortes énergie – considérées comme du signal de parole – des zones de faible énergie – considérées comme des zones de silences (Reynolds, 1992). Les diverses zones sont localisées par un processus itératif d'évaluation, en fonction du temps, des changements du rapport signal sur bruit.
Les signal est découpé en plusieurs fenêtres et l'algorithme estime le niveau d'énergie de la fenêtre suivante afin de se déterminer entre zone de silence et de parole dans la fenêtre qu'il analyse. Enfin, un filtre supprime les zones de silence.

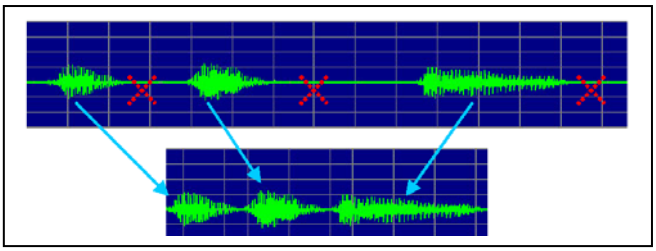


Figure 2.3.2.3.1 : suppression des zones de silence

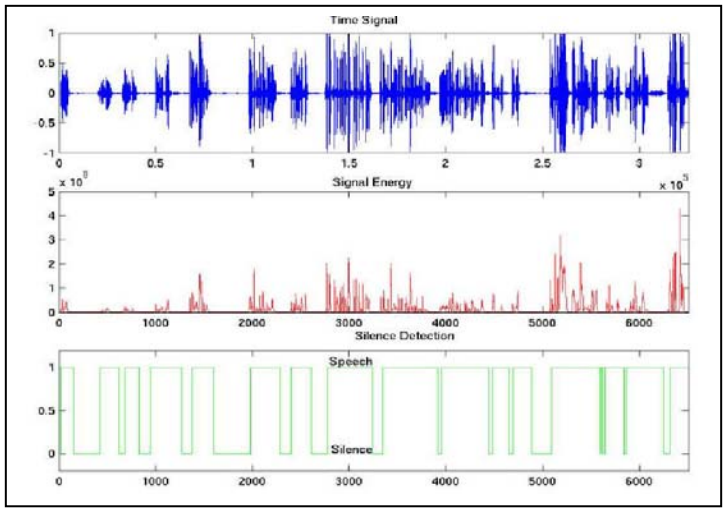


Figure 2.3.2.3.2 : procédure de suppression des silences

2.3.2.4. EXTRACTION DES PARAMÈTRES CARACTÉRISTIQUES (RASTA-PLP)

La méthode utilisée est la méthode de prédiction linéaire perceptuelle (PLP) couplée à la technique spectrale relative (RASTA) (Hermansky et Morgan, 1994). Les paramètres fournis par cette technique sont reconnus comme faisant partie des plus robustes (avec les coefficients cepstraux en échelle de Mel (MFCC)) pour décrire l'enveloppe spectrale de la voix du locuteur.

La méthode de prédiction linéaire perceptive (PLP), étudiée par (Hermansky 1990), consiste en un filtrage en bandes critiques du spectre à court terme suivit d'une correction de l'intensité. L'amplitude du signal est alors compressée et enfin l'analyse par prédiction linéaire intervient. Cette dernière étape est en réalité une technique de compression spectrale qui modifie le spectre de puissance à court terme avant son approximation par un modèle autoregressif.

Les zones du signal qui correspondent à des composantes linguistiques n'ont pas la même évolution temporelle que les zones qui n'y correspondent. La technique spectrale RASTA utilise ce principe en supprime les composantes spectrales dont l'évolution temporelle est plus rapide ou plus lente que celle du tractus vocal..
La fonction de transfert du filtre peut être calculée en couplant les méthodes RASTA et PLP. La méthode est présentée en détail dans (Hermansky et Morgan, 1994).

D'un point de vue pratique, le signal de parole est découpé en tranche de 20ms (framing). Il est à noter que des tranches trop courtes ne permettent pas l'analyse spectrale alors que des tranches trop longues risquent de tomber sur des parties non stationnaires du signal (il est en effet admis que pour des ordres de grandeurs de 20 à 30ms, le signal est stationnaire).
Les tranches découpées sont alors pondérés par des coefficients ou "fenêtrées" afin d'éviter les effets de bords. La fenêtre utilisée est celle de *Hamming*. Un pas, ou recoupement, de 10ms lie deux tranches adjacentes.

La fréquence d'échantillonnage du signal est de 8KHz.

De chacune des fenêtres sont extraits 12 coefficients PLP. Chacun de ces 12 coefficients est la coordonnée d'un vecteur à 12 dimensions. L'ensemble des vecteurs (12x1) forme une matrice (12xY) ou Y dépend de la longueur totale de l'enregistrement traité. Ces matrices contiennent les caractéristiques dépendantes du locuteur.

2.3.2.5. MODÉLISATION STATISTIQUE DES PARAMÈTRES EXTRAITS (GMM)

Afin de pouvoir effectuer des mesures de similarité, les paramètres extraits doivent être modélisés statistiquement. Il existe plusieurs méthodes de modélisation (discrimination par la valeur moyenne, alignement temporel par programmation dynamique, représentation par quantification vectorielle ou encore modélisation par modèles de Markov cachés en sont des exemples) mais c'est la représentation par mélange de fonctions de densité gaussiennes (GMM Gaussian Mixture Models) qui a été retenue pour le système ASPIC.
La méthode GMM est une méthode de classification paramétrique globale. Cela signifie que la forme de la distribution vectorielle est supposée connue et que l'ordre de mesure des vecteurs n'est pas considéré comme significative.
L'hypothèse sous-tendant cette méthode est que la distribution des caractéristiques dépendantes du locuteur peut être modélisée par une fonction de densité de probabilité gaussienne multidimensionnelle par un vecteur de moyennes et une matrice e covariance considérée diagonale (Reynolds DA et Rose RC, 1995).
Le système ASPIC utilise 32 fonctions gaussiennes pour entraîner un modèle.

L'algorithme Expectation Maximisation (EM) est utilisé pour estimer les classes du modèle qui permettent de maximiser la distribution des caractéristiques par un processus itératif de prévision et de maximisation non supervisé.

2.3.2.6. MESURES DE SIMILARITÉ

La mesure de similarité se fait toujours entre des paramètres extraits 'un échantillon de voix et le modèle construit sur la base des paramètres extraits d'un autre échantillon de voix. Ni les paramètres ni les modèles ne peuvent être comparés directement entre eux.
La mesure de similarité entre les caractéristiques et le modèle se fait en calculant la probabilité conditionnelle des vecteurs construits avec les paramètres extraits connaissant le modèle. Le résultat obtenu est donc une valeur. Celle-ci représente la proximité entre les deux enregistrements d'intérêt. Plus la valeur est grande, plus les échantillons de voix comparés sont proches l'un de l'autre.

2.3.3 COMPARAISONS TRACE-MODÈLE DANS UN CADRE BAYESIEN

La Méthode des Scores (Meuwly et Drygajlo, 2000) propose de considérer l'approche bayesienne de la reconnaissance de locuteurs en sciences forensiques de la manière suivante:

Avec le suspect, des enregistrements de comparaison de deux types sont réalisés: des enregistrements de contrôles **(C)** et des enregistrements de références **(R)**.
Les premiers doivent être les plus similaires possible à la trace en termes de conditions de production, de transmission ou encore d'enregistrement du signal, ainsi qu'en terme de langue, de mode d'élocution et de contenu linguistique. Quant aux seconds, ils doivent être les plus similaires possible à la population potentielle **(P)**, selon les mêmes critères que ci-dessus.

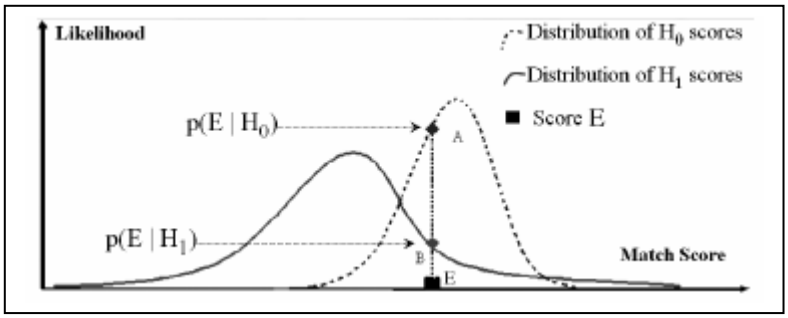
Les enregistrements de références **(R)** servent alors à modéliser la voix de l'auteur, et le modèle produit est comparé avec les paramètres extraits de la trace **(T)**.

> **La valeur de proximité rendue par le système automatique est le score E, qui représente la force probante de l'indice.**

Dans un contrôle d'accès par reconnaissance vocale par exemple, une valeur seuil est fixée pour **E**. Lorsque la valeur rendue par le système automatique, suite à la comparaison de deux échantillons de voix, est supérieure au seuil fixé, le système considère que la voix de la même personne se trouve sur les deux échantillons analysés. De façon opposée, lorsque la valeur de E est en dessous de la valeur du seuil, le système considère que les voix de deux personnes différentes sont présentes sur les échantillon.
Une telle approche binaire d'acceptation ou de rejet ne peut être acceptée en biométrie forensique.
En effet, la notion d'identité n'est jamais qu'une probabilité et qui plus est, conditionnelle à deux hypothèses alternatives.
Ainsi, des enregistrements de contrôles réalisés avec le suspect sont extraits les paramètres de sa voix et ces derniers sont comparés avec son modèle (obtenu avec les enregistrements de référence). Cela permet de déterminer la variation intralocuteur et donc d'évaluer l'indice E sous l'hypothèse **H0 => P(E/H0).** *H0 est l'hypothèse selon laquelle le suspect est à l'origine de l'énoncé de question.*
Enfin, les paramètres extraits de la trace sont comparés avec les modèles de la population d'intérêt afin d'obtenir la variabilité interlocuteur. L'indice E peut ainsi être évalué sous l'hypothèse **H1 => P(E/H1)**. *H1 est l'hypothèse selon laquelle quelqu'un d'autre d'une certaine population donnée est la source de l'énoncé de question.*
Le rapport de vraisemblance est alors la probabilité relative d'observer un score E dans la distribution des scores qui représentent d'un part, la variation des caractéristiques propres au suspect (intravariabilité) et d'autre part, la variation des caractéristiques propres à la population potentielle (intervariabilité), et cela au regard de la trace.

$$LR = \frac{P(E|H_0)}{P(E|H_1)}.$$

Graphiquement, cela se présente de la manière suivante:



LR = A/B

Figure 2.3.3.1 : Distribution des scores sous H_0 et sous H_1

Il est ainsi nécessaire de disposer non seulement de la trace mais également de trois bases de données contenant respectivement des données de références **(R)** propres au suspect, des données de contrôles **(C)** propres au suspect, et finalement des données propres à une population de référence **(P)**. Si en plus les performances du système veulent être mises à l'épreuve, une base de donnée supplémentaire contenant des traces **(T)** doit pouvoir être exploitée.

3 RECHERCHE BIBLIOGRAPHIQUE SUR LE SUJET

Dans ce chapitre seront abordés les trois thèmes suivants: les conditions d'enregistrements, les bases de données et l'interprétation des résultats.

3.1 CONDITIONS D'ENREGISTREMENTS

(Meuwly, 2001) met en évidence le fait que les performances du système automatique de reconnaissance de locuteurs sont nettement altérées lorsque le modèle est construit sur la base d'enregistrements provenant du réseaux cellulaire alors que les enregistrements de tests proviennent du réseau cellulaire ou fixe. Comme le soulève cet auteur, le résultat ne dois pas étonner puisque le réseau publique commuté assure un débit linéaire de $64Kbits^{-1}$ alors que le réseau cellulaire n'assure lui, qu'un débit de $16Kbits^{-S}$. La fidélité du signal de base, transmis par le système de codage cellulaire, est moindre que celui du système de codage du réseau publique. Il ajoute encore que du point de vue forensique, la conséquence immédiate est qu'il est "nécessaire de connaître le type de réseau par lequel a été transmis l'indice, de manière à réaliser l'enregistrement des modèles et les enregistrements de comparaison avec le même type de réseau téléphonique ; dans ce cas seulement la méthode de codage de la parole sera homogène dans tous les enregistrements."
(Meuwly, 2001) a comparé les résultats rendus par le système automatique lorsque l'enregistrement de test provient du réseau cellulaire alors que le modèle provient du réseau public commuté. Il est advenu que les performances du système se trouvaient nettement affaiblies.
Dès lors, l'auteur conclut qu'il est nécessaire d'utiliser des bases de données dont les enregistrements proviennent d'un même réseau pour être en mesure d'évaluer la variabilité interlocuteur mais que ces bases de données sont encore peu nombreuses pour le réseau cellulaire.

Dans (Dessimoz, 2004), les performances du système automatique de reconnaissance ont été testées avec des comparaisons entre des échantillons de voix provenant de conditions similaires et différentes d'enregistrement. Il est apparu que lorsque les conditions sont similaires, les performances sont meilleures. Les expérimentations faites ont montré que lorsque les tests et les

modèles étaient réalisés en condition cellulaire, les résultats étaient meilleurs que ceux obtenus en condition fixe. L'auteur note que ces résultats sont étranges et qu'ils vont à l'encontre de la logique puisque la qualité du signal de parole est inférieur sur le réseau cellulaire. Dessimoz explique les résultats obtenus par le nombre restreint de locuteurs testés. Il souligne qu'il serait intéressant de poursuivre les expérimentations avec un nombre supérieur de locuteurs.

3.2 BASES DE DONNÉES

(Meuwly, 2001), sur la base de (Tippett et al, 1968), soutient qu'une procédure de reconnaissance de locuteur nécessite de posséder deux bases de données. La première doit servir à modéliser la variabilité interlocuteur dans la population potentielle des locuteurs pouvant être à l'origine de l'énoncé de question. La seconde, doit permettre d'évaluer la variabilité intralocuteur de la personne suspectée.

(Alexander et al., 2004) rappellent que la méthodologie bayesienne requière l'utilisation de trois bases de données en plus de la trace. D'une part une base de donnée comportant des enregistrements de références élaborés avec le suspect, d'autre part une base de données d'échantillons de contrôles construits avec le suspect et enfin une base de données de la population potentielle. De plus, lorsque les performances du systèmes veulent être testées, une base de données contenant des traces est également souhaitable.
La méthodologie utilisée pour calculer le rapport de vraisemblance est proposée par (Meuwly et Drygajlo, 2001). Les paramètres caractéristiques de la voix sont extraits par le système automatique de la trace et ils sont comparés avec le modèle de référence du suspect. Puis, la trace est également comparée avec les représentations statistiques de tous les locuteurs de la base de donnée potentielle.
La distribution des *log-likelihood scores* indique l'intervariabilité de la trace avec la population potentielle. La base de données de contrôle permet d'effectuer des comparaisons avec le modèle du suspect et la distribution de *log-likelihood scores* donne l'intravariabilité du suspect. Enfin, le rapport de vraisemblance peut être obtenu par le rapport de la distribution des scores de la variation intralocuteur sur celle de la variation interlocuteur pour une valeur donnée de l'indice (E).

(Botti et al, 2004) ont proposé une méthodologie afin de traiter les cas dans lesquels seul une quantité limitée de données provenant du locuteur est disponible. Dans ces cas, il est impossible d'estimer la variation intralocuteur du suspect. Ces travaux font suite à ceux réalisés en 2003 par les mêmes auteurs (Botti et al, 2003).
La comparaison entre l'enregistrement du suspect et la trace, par le système automatique, donne la valeur (E) de l'indice. Pour évaluer cette valeur, deux bases de données sont nécessaires. La première (SDB pour Speakers Database) contient des enregistrements réalisés avec "pseudo-suspects". Leur modèle est construit. Les conditions d'enregistrement devraient être similaires à celles de l'enregistrement du suspect. La seconde base de données (TDB pour Trace Data Base) contient des "pseudo-traces" réalisées avec les mêmes locuteurs que SDB. Cette base de donnée est utilisée pour tester les modèles des locuteurs et les conditions d'enregistrements doivent être similaires à la trace.
Dès lors il s'agit de simuler deux sortes de cas: ceux ou les enregistrements proviennent de la même source (cas ou H_0 es vérifiée et cas ou H_1 est vérifiée). Les distributions des scores pour chacune des hypothèses vérifiées sont *plottés* dans un graphique. Suite à cela, la trace indiciaire est comparée avec les modèles des "pseudo-suspects".
Deux mesures permettent d'évaluer la valeur de E: le rapport de vraisemblance (LR) ainsi que le taux d'erreur (ER). Ce dernier paramètre offre des informations complémentaires au LR. Cette mesure prend en considération le risque relatif d'erreur pour un score obtenu en choisissant l'une ou l'autre des hypothèses. Le ER est le nombre de cas dans lesquels des enregistrements provenant de sources identiques sont faussement considérés comme provenant sources différentes sur le nombre de cas dans lesquels des enregistrement provenant de sources différentes sont faussement acceptés comme provenant de la même source si la valeur de E est employée comme un seuil dans une hypothèse test pour la concordance ou la discordance .
Si par exemple une valeur de ER de 10 est obtenue cela signifie qu'il est 10 fois plus probable de commettre une erreur en excluant le suspect comme étant à la source de l'enregistrement

indiciaire qu'en l'identifiant comme étant la source de la trace lorsque le score E est utilisé pour décider.
L'intérêt de la méthode repose sur la fait qu'il est possible de créer par avance les distribution H_0 et H_1 dans différentes conditions et d'évaluer un nouveau cas en regard de conditions similaires (Koolwaaji et Boves, 1999). Il s'agit simplement de déterminer si les conditions des bases de données sont similaires à celles du cas. Il est noter que la valeur du ER augmente en même temps que celle de E. L'ER donne alors des informations sur la qualité du match pour le score obtenu alors que le LR considère le nombre de fois ou la valeur E a été obtenue sous les deux hypothèses alternatives.

Une description complète de la base de données créée par le FBI est faite dans (Beck et al., 2004). Cette base de donnée contient 235 locuteurs dont 127 femmes. Elle comporte plusieurs modes d'élocution ainsi que plusieurs conditions d'enregistrements (variations des microphones), et chaque locuteur devait s'exprimer en anglais et dans sa langue natale. Deux sessions ont été réalisées avec chacun des locuteurs.

3.3 INTERPRÉTATION DES RÉSULTATS

L'interprétation de la valeur de l'indice a déjà été présentée au chapitre 2.3.3. Il convient ici de noter que l'approche par évaluation du LR peut se faire selon deux méthodes: la méthode des scores et la méthode directe.

La méthode des scores est celle présentée au chapitre 2.3.3

La méthode directe (Direct Method) (Alexander et Drygajlo, 2004) définit le LR comme la probabilité relative d'observer les caractéristiques de la trace dans la distribution des probabilités des caractéristiques du suspect et d'observer ces mêmes caractéristiques dans la distribution probabiliste du modèle de n'importe quel autre locuteur de la population potentielle.
Ici les exigences sont plus restreintes que pour la première méthode. En effet, il suffit de deux bases de données, en plus de la trace. La première qui constitue la base de donnée de référence **(R)** du suspect et la seconde qui constitue la base de donnée de la population d'intérêt **(P)**.

Si la première méthode détermine le LR en utilisant les distributions des vraisemblances obtenues en fonction des scores, la seconde méthode considère telles quelles les valeurs des vraisemblances qui sont générées par le système automatique et elle les utilise directement pour calculer le LR.
Il est à noter que la méthode des scores est utilisée dans bon nombre de domaines forensiques mais que la méthode directe ne peut être utilisée que dans les cas ou les résultats des analyses sont des vraisemblances.

Comme l'expliquent (Alexander et Drygajlo, 2004), chacune de ces deux méthodes est affectée par les incompatibilités qui peuvent survenir lorsque les conditions d'enregistrement des bases de données diffèrent.
Cependant, dans la méthode directe, la compensation de l'incompatibilité peut se faire soit dans l'espace des caractéristiques acoustiques, soit dans la modélisation statistique des caractéristiques. Dans la méthode des scores, une compensation statistique peut être entreprise en utilisant des bases de données différentes et incompatibles pour lesquelles la valeur de cette incompatibilité peut être estimée.

Dans les expériences réalisées par ces auteurs, il a été mis en évidence que les variations du LR sont plus faibles dans la méthode des scores mais par contre sa mise en œuvre est plus fastidieuse puisqu'elle nécessite la présence de trois bases de donnés ce qui va automatiquement allonger les temps de calculs.
La méthode directe fournit des valeurs de LR plus grandes mais elle ne permet pas d'estimer l'intravariabilité du suspect.

Les deux méthodes sont valides dans le calcul du rapport de vraisemblance. La première méthode paraît plus adéquate puisqu'elle permet de modéliser l'intravariabilité du suspect et l'intervariabilité de la trace. Cependant, cela nécessite de pouvoir réaliser non seulement des enregistrements de référence avec le suspect mais également des enregistrements de contrôle. Or, il n'est pas toujours possible pour l'expert d'avoir le suspect à sa disposition pour réaliser de tels enregistrements. Il se peut que le seul matériel qu'il possède soit un enregistrement incriminé et un enregistrement du suspect. Il ne lui est dès lors pas possible d'évaluer l'intravariabilité du suspect.

Dans ce dernier cas, en plus du rapport de vraisemblance, un autre rapport doit être utilisé qui donne des informations complémentaires. Il s'agit du "Rapport d'Erreur" ou Error Ratio comme présenté dans le paragraphe précédent.

4 BUTS DU TRAVAIL

Le but de ce travail de recherche est d'évaluer l'influence des conditions d'enregistrements dans la reconnaissance automatique des locuteurs en sciences forensiques. Cependant, afin de pouvoir atteindre ce but, la création d'une base de données permettant non seulement d'évaluer les performances du système automatique utilisé mais également de permettre une approche forensique de la reconnaissance des locuteurs a du être créée.
La création de cette base de donnée est au cœur de ce travail.

5 PARTIE PRATIQUE

5.1 CRÉATION DE LA BASE DE DONNÉE IPSC03

5.1.1 INFRASTRUCTURE, MATÉRIEL ET PROCÉDURE

L'enregistrement de la base de données s'est faite dans les locaux de l'Institut de Police Scientifique (IPS) de l'Ecole des Sciences Criminelle (ESC) de l'Université de Lausanne (UNIL).

Tous les enregistrements ont été réalisés en conditions contrôlées dans un local calme. Un téléphone fixe (raccordé au réseau RTCP) *Meridian, Nothern Telecom*® (France) et un téléphone cellulaire (raccordé au réseau GSM) *Nokia® 8310* ont été utilisés. Les deux téléphones employés utilisent le réseau de transmission de Swisscom®.
Les appels ont été transmis et enregistrés sur un serveur ISDN (standard européen DSS1) situé à l'Institut de Traitement des Signaux (ITS) de l'Ecole Polytechnique Fédérale de Lausanne (EPFL). Parallèlement à cela, un enregistreur numérique portable Sony® ICD-MS1 (échantillonné à 11025Hz), raccordé à un microphone à condenseur électret Sony® Cardio ECM-23, placé à environ 30 cm de la bouche du locuteur, a été utilisé.

Les locuteurs enregistrés, étudiants, assistants et Professeurs à l'IPS, sont tous des hommes âgées de 18 à 45 ans parlant le français. Deux enregistrements ont été réalisés avec chacun des locuteurs. Un premier avec le téléphone fixe et un second avec le téléphone cellulaire. En même tant que le locuteur parlait dans le téléphone, sa voix était captée par le microphone et enregistré sur le dictaphone numérique Sony®. Quatre fichiers audio de 10 à 15 minutes chacun ont ainsi été obtenus avec chacun des locuteurs, et cela sous trois conditions d'enregistrements différentes (Fixe, cellulaire et enregistrement direct sur support numérique (2x). Nous avons appelé ces conditions respectivement Fix, Cellular et Digital).

La personne enregistrée avait, devant elle, une présentation *PowerPoint*® sur support papier sur laquelle les instructions et les actions à réaliser étaient inscrites. Au niveau des actions, il s'agissait de prononcer des textes en employant différents modes d'élocution imposés.
Les modes d'élocution sont de trois type : le mode que nous avons appelé "normal" qui consiste simplement en de la lecture de texte. Le mode spontané, appelé "spontaneous" qui est obtenu par la description d'une image ainsi que par deux situations d'appels simulés. La première consiste en

l'appel à une personne de connaissance que le locuteur doit menacer de mort et le second, en un appel à la police pour signaler la présence d'une bombe dans des toilettes. Enfin, le troisième mode "dialogue" est obtenu par la lecture d'un dialogue que le locuteur enregistré devait lire en y mettant du ton.

Les fichiers audio enregistrés (format .MSV sur l'enregistreur numérique et .Wav sur le serveur) ont été "découpés" avec le logiciel CoolEdit pro 2® en différents "sous-fichiers" qui seront présentés dans le chapitre suivant.

5.1.2 CONTENU DE LA BASE DE DONNÉES

La structure de la base de donnée construite est la suivante :
Elle contient 11 enregistrements "Trace" **(T)**, 3 enregistrements "Reference" **(R)** et 3 enregistrements "Control" **(C)**.

Au niveau des enregistrements "Trace", 9 fichiers sont constitués de textes lus et 2 fichiers sont de la parole spontanée. Les 9 fichiers se découpent en trois groupes de trois fichiers, chaque groupe ayant un contenu linguistique similaire. En effet, pour chacun de ces groupes, un texte a été choisi dans la presse et les deux autres ont été construits sur ce modèle avec le souci de respecter le contenu linguistique. Les deux textes en parole spontanée sont des simulations d'appels telles que décrites dans le paragraphe 5.1.1.
Les enregistrements "Reference" contiennent uniquement du texte lu. Deux de ces enregistrements sont identiques l'un à l'autre. Ces différents textes ont été repris de IPSC02. Un lien avec cette base de données peut donc être envisagé.
Les enregistrements "Control" regroupent trois modes d'élocution différents. Le mode spontané, correspondant à la description d'une image, le mode normal, correspondant à la lecture d'un texte et le mode dialogue, pour lequel le locuteur devait lire un texte en y mettant le ton du dialogue.

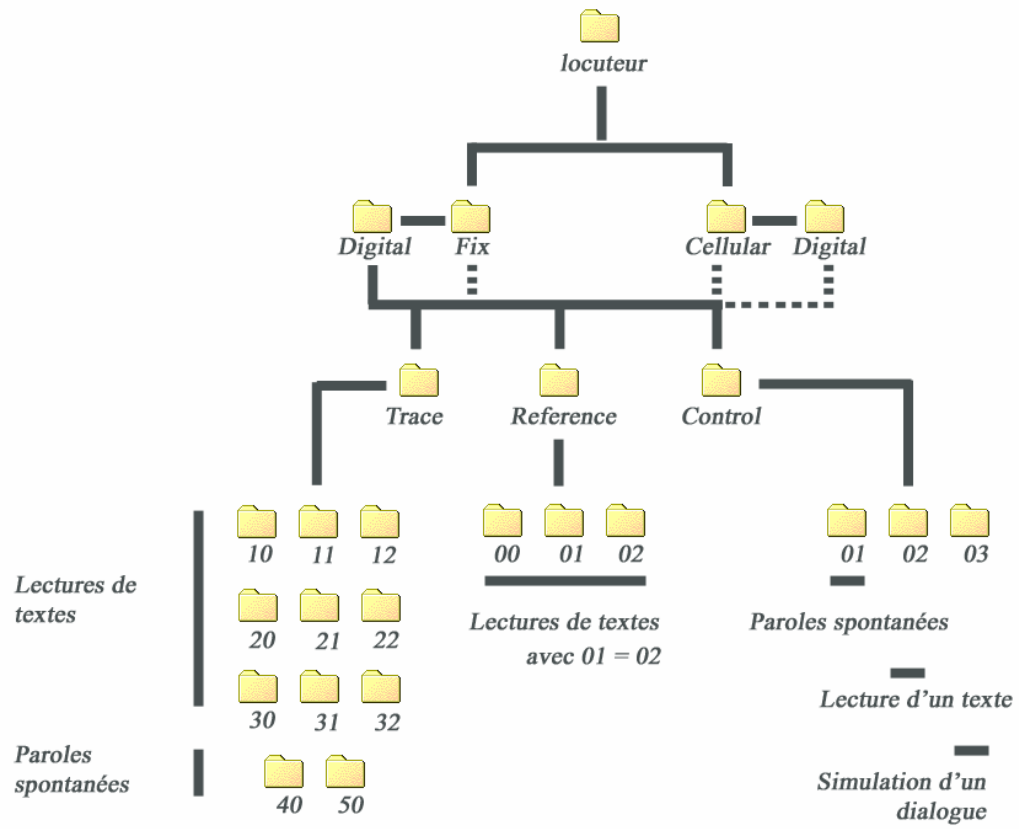


Figure 5.1.2.1 : Structure de la base de donnée

Remarque : La numérotation des fichiers est décrite ci-dessous dans : "Nomenclature des fichiers"

Ainsi, 17 fichiers ont été obtenus pour chacun des locuteurs dans chacune des conditions, soit un total de 68 fichiers/locuteur. Les temps d'enregistrements varient de quelques secondes pour les plus courts (T40 et T50) à environ deux minutes pour les plus longs (R01 et R02).

Le nombre de locuteurs enregistrés dans la base de données est de 73. Il est à noter que 63 fichiers sont complets et que 10 ne le sont que partiellement et cela pour diverses raisons techniques et pratiques que nous n'aborderons pas ici dans la mesure où ces problèmes peuvent tous être très facilement résolus par un nouvel enregistrement des sous-fichiers manquants avec les locuteurs concernés.

La somme de tous les fichiers audio contenus dans la base de donnée est de 4798. Cela représente un temps d'enregistrement d'environ 40 à 45 heures.

Nomenclature des fichiers :

- Un exemple avec le locuteur n°1, en condition "Fix", pour le fichier "Control 02" qui est en mode d'élocution normal:

M001FRFC02_NO

Signification des lettres utilisées :

- M = "Male".
- 001 = n° du locuteur. Ce numéro varie de 001 à 073.
- FR = car la langue dans laquelle les fichiers ont été enregistrés est le français.
- C = "Control". Cette lettre est remplacée par un T pour les fichiers "Trace" et par un R pour les fichiers "Reference".
- 02 = numéro du sous-fichier. Ce nombre peut prendre les valeurs de 10, 11, 12, 20, 21, 22, 30, 31, 32 pour les fichier T, 00, 01 et 02 pour les fichiers R et 01, 02 et 03 pour les fichier C.
- NO = normal. Ces lettres sont remplacées par SP pour le mode d'élocution "Spontanous" et par DL pour le mode d'élocution "Dialogue".

5.2 UTILISATION DE LA BASE DE DONNÉES

La base de données créée a été utilisée dans différentes applications. Les performances du système ASPIC ont été testées, des cas ont été simulés et l'approche bayesienne a été utilisée pour les évaluer. Enfin, cette base de données a été utilisée pour mener une recherche qui a abouti à la rédaction d'un article (Arcienega et al., 2005).

5.2.1 VÉRIFICATION DES PERFORMANCES

La vérification des performances du système est une étape essentielle et impérative. Elle permet de s'assurer que l'extraction des caractéristiques et les algorithmes de classification soient bien en mesure de discriminer deux locuteurs différents dans des conditions données (Alexander et al, 2004).

5.2.1.1. MÉTHODES UTILISÉES

Les performances du système peuvent être évaluées de deux manières. D'une part, à travers les courbes DET (Detection Error Trade-off) qui modélisent graphiquement le taux de fausse acceptation en fonction du taux de faux rejet pour chacune des vraisemblances considérées et d'autre part, à travers le calcul du coefficient de discrimination (DC).

Les courbent DET permettent non seulement de visualiser simultanément chacun des taux d'erreur pris en compte mais également de calculer le taux d'égale erreur (EER). Ce dernier représente le moment où le nombre de fausse acceptation et celui de faux rejet sont identiques. Graphiquement, il est l'intersection entre la courbe DET modélisée et la diagonale (positive) de la fenêtre graphique.

Le coefficient de discrimination exprime la capacité du système utilisé à différencier deux échantillons de voix qui proviennent de personnes différentes. Pour que la discrimination soit efficiente il faut, dans le cas où le suspect est l'auteur de l'énoncé incriminé – H_0 est vraie – que la distribution des scores de vraisemblance se situent dans une même gamme de valeurs. Cette gamme de valeurs doit se distinguer de celle obtenue lorsque le suspect n'est pas l'auteur de l'énoncé – H_1 est vraie – en question.
Plus la séparation entre les gammes de valeur est grande, meilleurs sont les performances du système et plus grand sera le coefficient de discrimination.

(Alexander et al, 2004) proposent de considérer le DC en faisant le rapport entre la différence des moyennes de H_0 et de H_1 et la somme des déviations standards selon :

$$DC = \frac{\mu_{H0} - \mu_{H1}}{\sigma_{H0} + \sigma_{H1}}$$

Comme l'explique (Alexander et al, 2004), une valeur de DC inférieur ou égale à 1 signifie que plusieurs valeurs appartenant à la distribution de H_0 appartiennent aussi à la distribution de H_1 est la séparation est mauvaise. Par contre, si la valeur du DC est comprise entre 1 et 2, la séparation est moyenne à bonne et si cette valeur dépasse 2 la séparation est très bonne.

5.2.1.2. APPLICATION DE LA MÉTHODE

Les 73 locuteurs ont été considérés. Les échantillons "Reference" ont été utilisés pour construire les modèles et les échantillons "Trace" ont été utilisés pour les comparaisons. Pour chacun des locuteurs, les 3 échantillons "Reference" et trois échantillons "Trace" (10, 20, 30) ont été employés. Les paramètres extraits de chacune des traces ont été comparés au modèle construit avec chacune des références. Au total, 9 scores sont obtenus pour chaque locuteur :

	TRACE 10	TRACE 20	TRACE 30
REFERENCE 00	1.49^{E+00}	1.56^{E00}	1.71^{E00}
REFERENCE 01	1.49^{E+00}	1.50^{E+00}	1.69^{E+00}
REFERENCE 02	1.28^{E+00}	1.37^{E+00}	1.52^{E+00}

Locuteur M001 ; CvsF ; H_0 vérifiée

Tableau 5.2.1.2.1 Représentation des scores sous H_0

L'ensemble des scores obtenus pour les 73 locuteurs peut être représenté sous la forme d'une matrice (219x219) dont les 657 valeurs constituant la diagonale (9 scores pour 73 locuteurs) représentent les cas où H_0 est vérifiée alors que les 47304 autres valeurs (219*219 – 657) représentent les cas où H_1 est vérifiée.

Remarque:

Nous avons observé que des valeurs de scores relativement faibles étaient obtenues, systématiquement, lorsque les traces étaient comparées avec les échantillons "Reference00". Cela

est probablement dû à la durée du signal sur cet échantillon. Si les échantillons "Reference01 et 02" possèdent environ deux minutes de signal chacun, la "Reference00" n'en comporte que quelques secondes.
Une normalisation des valeurs de la matrice a été entreprise afin de diminuer le poids de ces scores. La distribution de tous les scores représentatifs de la variabilité intralocuteur est ainsi plus homogène.

	M001			M001		
	Trace 10	Trace 20	Trace 30	Trace 10	Trace 20	Trace 30
Reference 00	5.509066	6.781853	5.355406	1.10E+00	1.29E+00	1.63E+00
Reference 01	12.145497	12.100731	11.617859	1.54E+00	1.28E+00	1.45E+00
Reference 02	13.121375	12.601819	11.643188	1.05E+00	1.10E+00	1.12E+00

Locuteur M001 ; CvsC ; H_0 vérifiée avant (droite) et après (gauche) normalisation

Tableau 5.2.1.2.2 : Représentation des scores avant et après normalisation

La procédure de vérification des performances a été réalisée pour les conditions suivantes avec les valeurs de scores normalisées:

- **FvsF** => **traces et références** obtenues en condition d'enregistrement **"Fix".**
- **CvsC** => **traces et références** obtenues en condition d'enregistrement **"Cellular".**
- **CvsF** => **traces** obtenues en condition d'enregistrement **"Cellular"** alors que les **références** ont été obtenues en conditions d'enregistrement **"Fix".**
- **FvsC** => **inverse** de la proposition précédente.

Les courbes DET, le EER et les DC ont été calculés avec le programme MatLab ®.

5.2.1.3. RÉSULTATS OBTENUS

Les résultats issus des comparaisons sont visibles dans le graphique suivant et les taux d'égale erreur ainsi que les coefficients de discrimination obtenus sont présentés dans le tableau ci-

	EER (%)	DC
FvsF	8.37	1.11
CvsC	10.35	0.93
FvsC	42.00	0.12
CvsF	42.77	0.12

Tableau 5.2.1.3.1 : EER et DC

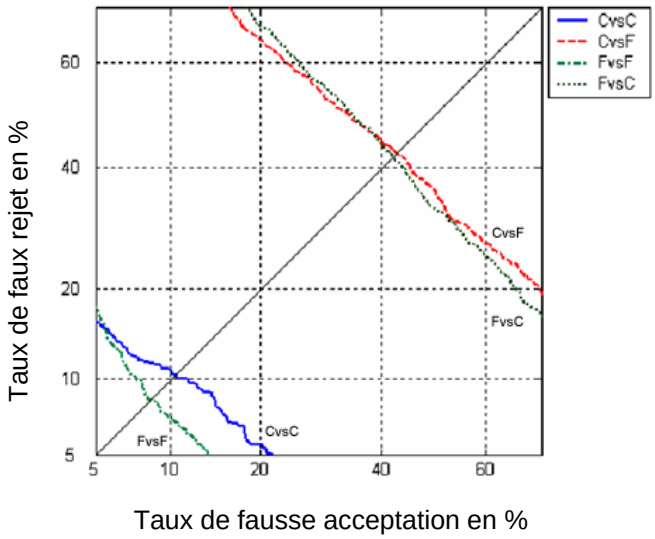


Figure 5.2.1.3.2 : courbes DET

5.2.1.4. DISCUSSION DES RÉSULTATS

La lecture des données représentées sur le graphique "Figure 5.2.1.3.2" doit se faire de la manière suivante : plus une courbe est proche du point origine (0,0), plus le taux de fausse acceptation et le taux de faux rejet sont bas et donc meilleure est la performance du système. Les points d'intersections entre la diagonale tracée sur le graphique et les courbes représentées donnent les taux d'égale erreur, relatifs à l'expérience que chaque courbe modélise.

Au niveau des résultats obtenus, nous pouvons immédiatement constater que les performances du système sont nettement meilleures lorsque les conditions d'enregistrement sont similaires. Cependant, même en conditions similaires, il apparaît que l'influence du réseau de transmission se fait sentir quant aux résultats fournis par le système automatique.
En effet, ceux-ci sont meilleurs lorsque le réseau de transmission utilisé est le réseau fixe (RPCT). Cela peut être constaté lorsque les taux d'égale erreur ainsi que le coefficient de discrimination sont pris en compte. Le EER pour le réseau fixe, se situe aux environs de 8% alors qu'il est d'environ 10% pour le réseau cellulaire (GSM).
Le DC, pour les enregistrements provenant du réseau fixe, est de 1.11. Cette valeur, interprétée selon les propositions de (Alexander et al, 2004), doit être comprise comme reflétant une performance moyenne du système de reconnaissance. Par contre, pour les enregistrements provenant du réseau cellulaire, le DC est de 0.93 et la performance du système est mauvaise. La distribution des valeurs de H_0 et de celles de H_1 sont mal séparées. Certaines valeurs appartenant à la distribution de l'une des deux hypothèses appartiennent également à l'autre.

Lorsque les conditions d'enregistrement sont différentes, la performance du système diminue et les résultats sont nettement moins bon. Que ce soit la trace qui provienne du réseau fixe et les comparaisons du réseau cellulaire ou que ce soit l'inverse ne semble pas avoir une grande influence sur les performances du systèmes. En effet, le taux d'égale erreur (42% pour FvsC et 42.77% pour CvsF) est quasiment le même pour les deux cas et la valeur du coefficient de discrimination est exactement la même (0.12 dans les deux cas). Les coefficients de discriminations obtenus pour les conditions d'enregistrements différentes montrent une mauvaise performance du système de reconnaissance automatique de locuteur.

5.2.1.5. CONCLUSION

Les résultats obtenus lors de l'évaluation des performances du système automatique montrent que le canal de transmission du signal de parole influence la performance du système. Ainsi, l'élément "conditions d'enregistrement" doit être pris en considération lorsque des comparaisons sont effectuées. Non seulement il s'agit de déterminer si les conditions d'enregistrement sont similaires ou différentes mais également à quel type de réseau on a affaire.

5.2.2 SIMULATION DE CAS; APPROCHE FORENSIQUE

La base de données créée a été utilisée pour simuler un certain nombre de cas permettant une approche forensique de l'interprétation de la valeur des scores rendus par le système de reconnaissance automatique.

5.2.2.1. MÉTHODE UTILISÉE

La méthode utilisée pour réaliser l'approche forensique est celle décrite dans le paragraphe 2.3.3. Il s'agit de la méthode des scores (Meuwly et Drygajlo, 2000).

Nous avons vu que cette méthode nécessitait, en plus de la trace, de disposer d'échantillons de référence et de contrôle, enregistrés avec le suspect, ainsi que d'une population potentielle.
L'emploi de ces bases de données est nécessaire pour permettre d'évaluer la force probante de l'indice au regard de deux hypothèses alternatives, par exemple:

H_0: Le suspect est à l'origine de la trace
H_1: Quelqu'un d'autre, appartenant à la population potentielle de référence, est à l'origine de la trace.

En effet, le calcule du rapport de vraisemblance (LR = $P(E/ H_0)/P(E/ H_1)$ nécessite de calculer alternativement la probabilité d'observer le score obtenu lors des comparaisons trace-modèle suspect, sachant que H_0 est vérifiée et la probabilité d'observer ce même score lorsque H_1 est vérifiée. Afin de mesurer la valeur de la première probabilité, l'intravariabilité au niveau de la voix du suspect doit être évaluée. Quant à la seconde probabilité, celle-ci ne peut être calculée qu'en connaissant les variations de la voix sur la trace par rapport aux voix de la population potentielle.

La représentation des résultats obtenus pour les rapports de vraisemblance calculés s'est faite dans ce travail sous la forme de Tippett plot (Tippett et al, 1968 et Evett et Buckleton, 1996 in Meuwly, 2001). Ceux-ci font figurer sur l'axe des abscisses, gradué de manière logarithmique croissante, les valeurs de rapport de vraisemblance. L'axe des ordonnées est gradué de 0 à 1 et représente la probabilité que la valeur du LR soit supérieure ou égale à la valeur indiquée en abscisse.
Sur le graphique, deux courbes sont représentées. La première modélise les cas ou H_0 est vérifiée et la seconde ceux ou H_1 l'est.

5.2.2.2. APPLICATION DE LA MÉTHODE

L'expérimentation consiste à prendre un locuteur et à le considérer comme étant le suspect alors que les 72 autres locuteurs sont considérés comme étant la population potentielle d'intérêt. L'expérimentation a été répétée avec 20 locuteurs.

Lors du processus d'évaluation du poids de l'indice par la méthode bayesienne, les données ont été choisies soit pour avoir H_0 vérifiée soit pour avoir H_1 vérifiée. À cette fin nous avons procédé en reprenant les matrices (219x219) comme décrites ci-dessus au paragraphe 5.2.1.2.
Pour chaque locuteur, 9 scores sont alors à disposition. Ils sont le reflet de l'intravariabilité de la voix de la personne enregistrée (3 "Trace" comparées avec 3 "Reference"). Nous avons considéré les échantillons "Trace" comme étant à la fois des contrôles et des traces.

Pour H_0 vérifiée:
De ces 9 scores, nous en avons pris 8 pour modéliser la variabilité intralocuteur (H_0) et 1 comme valeur de E. Les 216 valeurs (72 autres locuteurs) de la colonne à laquelle appartient E, ont été prises pour modéliser la variabilité interlocuteur (H_1).

		M001		
		TRACE 10	TRACE 20	TRACE 30
M001	REFERENCE 00	H_0	H_0	H_0
	REFERENCE 01	H_0	H_0	H_0
	REFERENCE 02	**E**	H_0	H_0
M002	REFERENCE 00	H_1		
	REFERENCE 01	H_1		
	REFERENCE 02	H_1		
M003	REFERENCE 00	H_1		
	REFERENCE 01	H_1		
	REFERENCE 02	H_1		
…	…	…		

Tableau 5.2.2.2.1 : Sélection des données pour H_0 vérifiée

Pour H_1 vérifiée:
Nous avons pris les 9 scores issus des comparaisons "Reference" - "Trace" pour modéliser la variabilité intralocuteur du suspect. La valeur E est le score obtenu lorsque la trace est comparée avec le modèle de référence du suspect. Les valeurs de H_1 sont toutes les autres valeurs qui se situent dans la colonne ou se trouve E (= 216 valeurs).

		M001			M021		
		Trace 10	Trace 20	Trace 30	Trace 10	Trace 20	Trace 30
M001	Reference 00	H_0	H_0	H_0			
	Reference 01	H_0	H_0	H_0			
	Reference 02	H_0	H_0	H_0	**E**		
M002	Reference 00				H_1		
	Reference 01				H_1		
	Reference 02				H_1		
…	…				…		

Tableau 5.2.2.2.2 : Sélection des données pour H_1 vérifiée

Exemple : Suspect = M001, Auteur = M021, Population potentielle = comparaison de la trace avec les modèles de tous les locuteurs à l'exception des 9 scores de la comparaison M001 – M021.

La même procédure a été appliquée avec les locuteurs M001 à M020. L'auteur étant toujours M021.

Remarque :

Le processus n'as été entrepris que pour les deux conditions d'enregistrement suivantes: CvsC pour les conditions similaires et CvsF pour les conditions différentes.
Le choix du téléphone cellulaire pour les conditions similaires repose sur les deux points suivants : d'une part les téléphones mobiles se sont largement répandu dans la population au cours des dernières années et leur utilisation dans le cadre d'activités criminelles est très courante et d'autre part, les performances du système de reconnaissance automatique de locuteur sont inférieures pour les téléphones cellulaires par rapport aux téléphones fixes.
En ce qui concerne les conditions différentes, nous avons vu dans le paragraphe 5.2.1.3 que les performances du système étaient similaires dans les cas FvsC et CvsF ainsi, la condition CvsF a été arbitrairement choisie.

5.2.2.3. RÉSULTATS OBTENUS

Les résultats, sous forme de Tippett plot, pour les comparaisons entre les conditions d'enregistrement similaires (CvsC) et différentes (CvsF) sont présentés dans les graphiques ci-dessous.

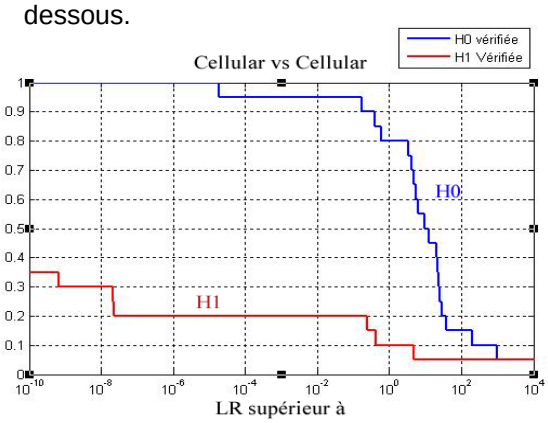


Figure 5.2.2.3.1 Tippet plots CvsC

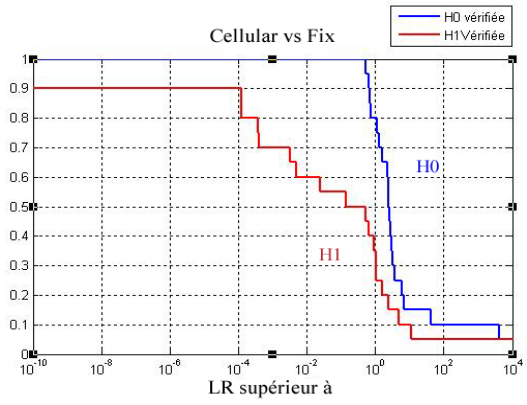


Figure 5.2.2.3.2 Tippet plots CvsF

5.2.2.4. DISCUSSION DES RÉSULTATS

Les clés pour lire le graphique sont les suivantes :

- Plus la séparation entre les courbes est grande, meilleures sont les performances du système. Idéalement, la courbe H_1 vérifiée devrait se trouver entièrement à gauche du point d'abscisse 10^0 (LR = 1) et la courbe H_0 vérifiée à droite de ce même point.
- La distance entre le point d'intersection de H_0 vérifiée et le sommet du graphique et le point d'intersection de H_1 vérifiée avec la base du graphique doit être la plus courte possible pour une bonne performance du système.

Les résultats montrent qu'en conditions d'enregistrements similaires CvsC, la séparation entre les courbes est bonne. Cependant, il apparaît également que la probabilité que la valeur du rapport de vraisemblance soit supérieue à 1, dans les cas où H_1 est vérifiée, n'est pas nulle. En effet, dans environ dix pour cent des cas cette valeur est supérieure à 1.
La probabilité d'obtenir une valeur de rapport de vraisemblance inférieure à 1, dans les cas où H_0 est vérifiée, n'est pas nulle non plus et survient dans environ 20% des cas.
Ainsi, la valeur du rapport de vraisemblance sera supérieure à 1 dans 10% des cas ou H_1 est vérifiée et dans 80% des cas où H_0 est vérifiée. Inversement, la valeur du rapport de vraisemblance sera plus petite que 1 dans 90% des cas où H_1 est vérifiée et dans 20% des cas où H_0 est vérifiée.
Il apparaît donc que le risque d'obtenir un faux positif (acceptation de la source commune de deux échantillons provenant de locuteurs différents) est plus faible que le risque d'obtenir un faux négatif (rejet de la source commune de deux échantillons alors qu'ils proviennent du même locuteur).

On constate encore que la distance qui sépare les intersections de la courbe H_0 vérifiée avec la partie supérieure du graphique et celle de H_1 vérifiée avec la partie inférieure du graphique est relativement importante. Les performances du système ne sont ainsi pas optimales.

Lorsque les conditions d'enregistrement sont différentes, la séparation entre les courbes est nettement moins bonnes que dans les conditions similaires.
Une valeur de LR plus grande que 1 est obtenue dans 80% des cas lorsque H_0 est vérifiée et dans 40% des cas où H_1 est vérifiée. A l'inverse, une valeur de LR inférieure à 1 est obtenue dans 40% des cas où H_0 est vérifiée et dans 60% des cas où H_1 est vérifiée.
Il est ainsi intéressant de constater que l'influence des conditions d'enregistrement différentes n'est pas vraiment importante dans les cas où H_0 est vérifiée. En effet le nombre de cas dans lesquels la valeur du LR est supérieure à 1 est identique à celui obtenu lors des comparaisons CvsC.
Par contre, dans les cas où H_1 est vérifiée, le nombre de cas pour lesquels la valeur du LR est inférieure à 1 n'est que de 60%.
Ainsi, la probabilité d'obtenir un faux positif est supérieur à celle d'obtenir un faux négatif. Cependant, cette probabilité reste tout de même élevée.

5.2.2.5. CONCLUSION

La recherche dans le domaine de la reconnaissance automatique de locuteur devrait donc viser le développement d'applications qui permettrait une meilleur extraction des paramètres caractéristiques de la voix et leur modélisation, tout en réduisant les influences des systèmes de codages et de transmission des réseaux téléphoniques. Ainsi, les taux de faux positifs et de faux négatifs pourraient être réduits et une meilleure séparation des valeurs des rapports de vraisemblance, dans les cas où chacune des hypothèses est vérifiée, serait obtenue.

5.2.3 AUTRE RECHERCHE MENÉE

(Arcienega et al., 2005) ont mené une recherche, en usant la base de données IPSC03, afin d'évaluer l'apport des réseaux bayesiens sur l'approche bayesienne "classique " telle que décrite au paragraphe 5.2.2.

5.2.3.1. MÉTHODE UTILISÉE

La méthode proposée par les auteurs combine la modélisation de l'enveloppe spectrale à celle des caractéristiques prosodiques du signal, afin de réduire l'influence des distorsions générées par les conditions d'enregistrements différentes du signal de parole.
Les performances de la modélisation statistique par GMM de l'enveloppe spectrale a été comparée avec une approche par réseaux bayesiens (Jensen, 2001), (Pearl, 1998). Celle-ci prend en compte, en plus de l'enveloppe spectrale, une autre caractéristique, dépendante de la première, la fréquence fondamentale.
Le réseaux bayesien utilisé est celui proposé par (Arcienaga et Drygajlo, 2003). La méthodologie repose sur la construction de models pour la fréquence fondamentale et l'enveloppe spectral qui sont conditionnels à une variable : le statu vocal. Cette dernière est introduite afin de mieux saisir les variations et de permettre une meilleur modélisation de la distribution des caractéristiques *vocales* et *aphones*. A un certain temps t, le statut vocale peut prendre soit la valeur 1 (*vocal*) soit la valeur 2 (*aphone*).
La modélisation de l'enveloppe spectrale se fait en excluant le plus possible les harmoniques liés à la fréquence fondamentale. Ainsi, cette dernière et l'enveloppe spectrale véhiculent des caractéristiques complémentaires et non corrélées.
Les caractéristiques de l'enveloppe spectrale sont modélisées par 2 GMM (*vocal* et *aphone*). La modélisation de la fréquence fondamentale est dépendante du statut vocal. Dans les zones *vocales* une modélisation des propriétés statistiques de la fréquence fondamentale est entreprise. Dans les zones *aphones*, il n'existe par réellement de valeur pour la fréquence fondamentale mais l'utilisation d'une table de probabilités discrètes en permet la modélisation.
Les probabilités du statut vocal sont définies par 2 *poids* qui expriment la probabilité de se trouver dans une zone *vocale* et la probabilité de se trouver dans une zone *aphone*.

5.2.3.2. APPLICATION DE LA MÉTHODE

Un certain nombre de cas ont été simulés afin de tester les performances de cette approche par réseaux bayesien. 10 locuteurs ont été sélectionnés dans la base de données et leurs modèles ont été construits pour chacune des conditions d'enregistrements ("Fix", "Cellular", "Digital"). 20 autres locuteurs ont été pris comme suspects afin de tester la performance de la reconnaissance vocale.
Environ deux minutes de signal de parole ont été utilisées pour entraîner les modèles de chacun des locuteurs. Par la suite, une minute de signal a été utilisé pour adapter les modèles des suspects. 6 échantillons d'environ 6 secondes chacun ont été utilisé pour les tests.
Les paramètres utilisés pour modéliser les caractéristiques de la voix sont les MFCCs (Mel-Frequency Cepstral Coefficients).

Des comparaisons du type FvsF, FvsC et FvsD ont été réalisées.

5.2.3.3. RÉSULTAS OBTENUS

Les résultats sont exprimés pour les deux approches, sous la forme du taux d'égale erreur (EER), dans le tableau ci-dessous :

	Approche classique EER(%)	Approche par réseaux bayesiens (EER%)
FvsF	4.8	3.3
FvsC	42.3	31.9
FvsD	37.5	22.5

Tableau 5.2.3.3.1 : EER et DC, comparaison des méthodes

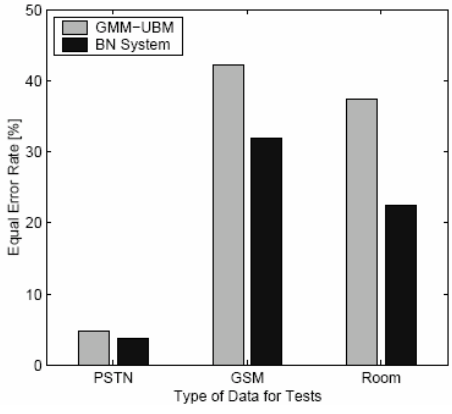


Figure 5.2.3.3.1 : FvsF, FvsC, FvsD, comparaisons des méthodes

	Approche classique EER(%)	Approche par réseaux bayesiens (EER%)
DvsD	4.8	3.3
DvsC	42.3	31.9
DvsF	37.5	22.5

Tableau 5.2.3.3.2 : : EER et DC, comparaison des méthodes

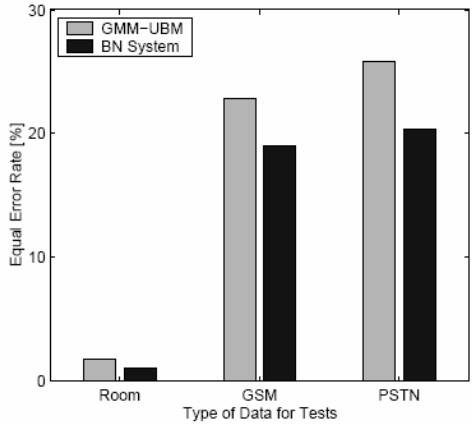


Figure 5.2.3.3.2 : FvsF, FvsC, FvsD, comparaisons des méthodes

Les performances du systèmes mesurées lors de l'approche forensique sont présentées dans les Tippett plot ci-dessous :

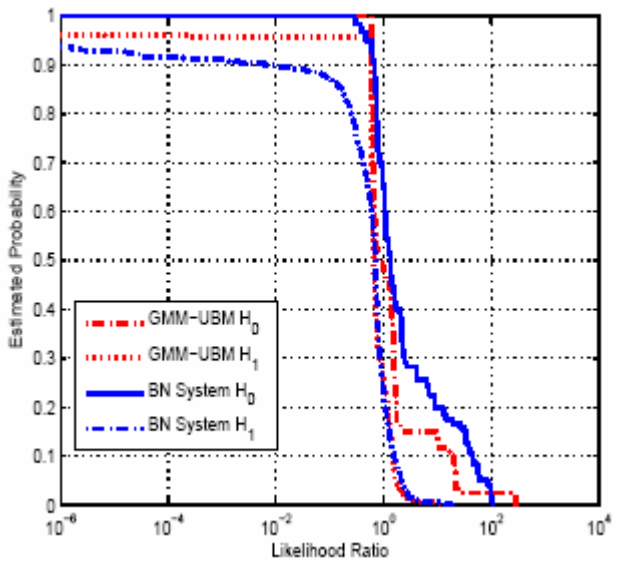


Figure 5.2.3.3.3 : Tippet plot CvsF

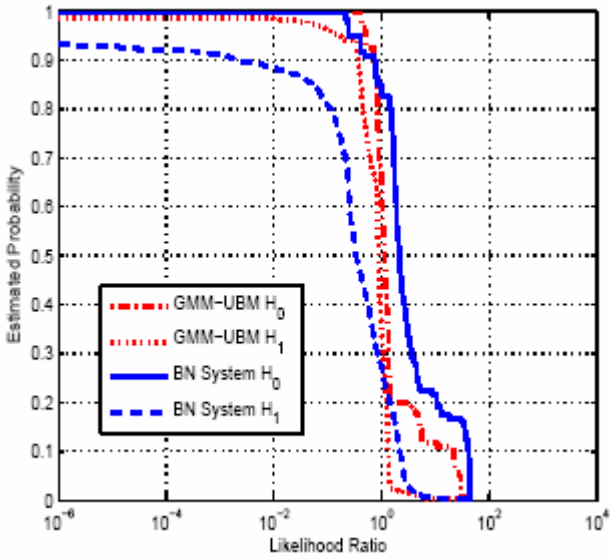


Figure 5.2.3.3.3 : Tippet plot FvsD

5.2.3.4. DISCUSSION DES RÉSULTATS

Lorsque l'on considère les EER obtenus lors des comparaisons présentées ci-dessus, il apparaît que tous les taux d'erreurs sont réduits lorsque la fréquence fondamentale est considérée. Bien que cette diminution soit plutôt faible dans les cas d'enregistrements en conditions similaires, elle devient plus importante dans les conditions d'enregistrements différents. Cela est principalement dû au fait que l'enveloppe spectrale de la voix subit plus de distorsions lorsque les conditions sont différentes.

Dans lesTippett plots présentés ci-dessus, on remarque que l'utilisation des réseaux bayesiens dans les cas où les conditions d'enregistrements sont différentes, une amélioration considérable dans la séparation des courbes intervient. Cela signifie donc que les performances du système sont meilleures. Cette différence intervient principalement lorsque les tests sont obtenus par le

réseau fixe et les modèles par enregistrement sur support numérique (FvsD). Lorsque les tests sont obtenus par le réseau fixe et les modèles par le réseau cellulaire, l'amélioration est présente mais moins significative que précédemment.
L'effet des conditions différentes d'enregistrement FvsC est ainsi plus important que l'effet FvsD.

5.2.3.5. CONCLUSION

L'approche par réseaux bayesien permet de prendre en considération des caractéristiques supplémentaires telles que la fréquence fondamentale ainsi que le statut vocal qui permettent une amélioration sensible des performances du système en conditions différentes d'enregistrement. Ainsi les informations véhiculées par la fréquence fondamentale sont non seulement solides au bruit (Arcienaga et Drygajlo, 2002) mais également aux distorsions générées par les lignes de transmissions du signal.

6 DISCUSSION GÉNÉRALE

6.1 INFLUENCE DES CONDITIONS D'ENREGISTREMENTS

Les divers expérimentations menées dans ce travail de recherche, avec la base de données IPSC03 créée à cet effet, ont montré que l'influence des conditions d'enregistrement dans la reconnaissance automatique de locuteur en sciences forensiques est un élément capital.
Cette influence se fait sentir aussi bien dans les cas où les conditions d'enregistrements sont similaires que dans les cas où elles sont différentes.

Lorsque les conditions sont similaires, une diminution de la performance du système est observée pour les enregistrements provenant de téléphones cellulaires par rapport aux enregistrements provenant de téléphones fixes. Cette diminution s'explique par la qualité moindre du signal transmis. En effet, le système de codage GSM réduit la quantité de données transmises. La seule contrainte à laquelle un opérateur est soumise est celle de l'intelligibilité du signal transmis. Tant que celle-ci demeure, l'opérateur remplit son contrat. Comme moins de données sont transmises par le réseaux cellulaire comparativement au réseau public commuté, moins de données caractéristiques de la voix du locuteur se retrouvent à la sortie du signal.
Lorsque les conditions sont différentes, les performances sont profondément altérées et cela que la trace soit transmise par le réseaux téléphonique publique commuté et le modèle du suspect par le réseau cellulaire ou que ce soit l'inverse qui survient.

L'effet de distorsion du signal par la ligne de transmission (en plus des autres effets dus à l'acquisition et à l'enregistrement) agit de manière déterminante sur la valeur probante de l'indice. C'est ainsi que la maîtrise de tous les paramètres d'influences apparaît comme un élément prépondérant dans l'interprétation de la valeur de l'indice.

6.2 BASE DE DONNÉES

L'interprétation de la valeur de l'indice ne peut se faire, pour la reconnaissance automatique de locuteurs, que si des bases de données sont disponibles. Certains paramètres essentiels doivent être pris en compte pour ces dernières. Il faut notamment que la taille soit suffisante pour assurer la représentativité statistique des données contenues. Il faut que la langue dans laquelle les échantillons de voix sont enregistrés correspondent à la trace de même que les conditions d'acquisition, de transmission et d'enregistrement du signal. Il faut encore que le contenu linguistique et les modes d'élocutions des éléments de la base de données correspondent aux éléments d'intérêts. Et finalement, il faut que ces bases de données possèdent plusieurs composantes qui permettent de choisir entre des échantillons de type : Traces, Référencs et

Contrôles. Ce n'est qu'alors qu'une approche bayesienne pourra être entreprise pour évaluer la valeur probante d'un indice.

La base de données développée dans ce travail de séminaire est de taille substantielle et sa structure est telle qu'une approche forensique de l'interprétation de la valeur de la preuve peut être entreprise. Cette base de données peut encore être développée en poursuivant les enregistrements afin d'augmenter le nombre de locuteurs qu'elle contient. Les performances de la base de données peuvent être évaluées en effectuant des tests avec le système automatique de reconnaissance. Les corrections nécessaire peuvent alors être apportées. Par exemple, lorsque nous avons vérifié les performances d'ASPIC (§ 5.2.1) nous avons dû normaliser les valeurs des scores obtenus en raison du temps plus court de l'un des enregistrements de référence par rapport aux deux autres. Il est ainsi possible que certains enregistrements ne soient pas idéaux, du point de vue de leur structure ou de leur contenu, et que des corrections doivent être apportées. Cela, soit en procédant à de nouveaux enregistrements avec les mêmes locuteurs, soit en construisant une nouvelle base de données qui tiendrait compte des éventuelles faiblesses de celle-ci.

La base de données IPSC03 a permis de mettre en évidence l'amélioration des résultats obtenus avec le système de reconnaissance automatique en conditions différentes, grâce à une approche par réseaux bayesiens et à la prise en compte de caractéristiques supplémentaires pour décrire l'enveloppe spectrale du locuteur. Ces caractéristiques sont la fréquence fondamentale et le statut vocal. L'évolution des systèmes de reconnaissance automatiques doit passer par la découverte de caractéristiques indépendantes de celles déjà utilisées et qui soient plus robustes aux distorsions afin de diminuer l'influence des conditions d'enregistrement et cela principalement lorsque celles-ci diffèrent.

7 CONCLUSION GÉNÉRALE

L'évaluation de l'influence des conditions d'enregistrement dans la reconnaissance automatique des locuteurs en sciences forensiques devait être abordée dans ce travail de recherche. Pour ce faire, il a été nécessaire de constituer une base de donnée répondant à certains critères. L'élaboration de cette base de donnée a constitué le cœur de ce travail.

Au niveau des conditions d'enregistrement, il est apparu que leur influence était prépondérante en matière de reconnaissance automatique et cela surtout lorsque elles étaient différentes.
Ainsi, lorsque une procédure de reconnaissance doit être abordée en sciences forensiques, il est capital de posséder les informations relatives aux conditions d'enregistrement de la trace, afin de pouvoir les reproduire pour créer des enregistrements de contrôle et de référence avec le suspect qui soient respectivement le plus proche possible de la trace et de la population potentielle.
Il est nécessaire de posséder une base de données de la population potentielle qui renferme les divers conditions d'enregistrements d'intérêt ainsi que d'autres paramètres, comme par exemple plusieurs modes d'élocution.
La base de données IPSC03 que nous avons créée permet de nombreuses applications. Elle peut notamment permettre de :

- Tester les performances d'un système de reconnaissance automatique.
- Evaluer la force de l'indice dans un cadre bayesien.
- Analyser l'influence de la quantité de signal à disposition (trace et suspect).
- Analyser l'influence du contenu linguistique, phonétique.
- Analyser l'influence du mode d'élocution.
- Analyser l'évolution des résultats en fonction de la taille de la base de données.
- Analyser la base de données afin d'en déterminer les forces et les faiblesses et d'évaluer les conséquences que cela peut avoir sur la construction d'une autre base de données ou sur le développement de celle-ci.

L'intérêt de cette base de données et donc de permettre, non seulement de servir à des applications multiples dans le domaine de la reconnaissance automatique de locuteurs mais également, de manière plus générale, à l'optimisation de bases de données.

Zimmermann Philipp

REMERCIEMENTS :

- Dr Andrzej Drygajlo
  - ITS-EPFL
- Alexander Anil, assistant doctorant
  - ITS - EPFL
- Botti Filippo, assistant doctorant
  - ESC – IPS
- Dessimoz Damien, assistant doctorant
  - ESC - IPS
- Les locuteurs enregistrés

8 BIBLIOGRAPHIE

Alexander A, Botti F, Drygajlo A. Handling Mismatch in Corpus-Based Forensic Speaker Recognition. The Speaker and Language Recognition Workshop, Toledo. *Speaker Odyssey* 2004: 69-74.

Alexander A, Botti F, Dessimoz D, Drygajlo A. The effect of mismatched recording conditions on human and automatic speaker recognition in forensic applications. *Forensic Science International* 2004; 146(1): 95-S99.

Alexander A, Drygajlo A. Scoring and Direct Methods for the Interpretation of Evidence in Forensic Speaker Recognition. Proceedings of 8th International Conference on Spoken Language Processing, Jeju, Korea, 2004.

Arcienaga M, Drygajlo A. Robust voiced(unvoiced Decision Associated to Continuous Speech Tracking in Noisy Telephone Speech for Robust Speaker Verification. *Lecture Notes in Computer Science* 2003; 2688: 78-85.

Arcienaga M, Drygajlo A. Scoring and Direct Methods for the Interpretation of Evidence in Forensic Speaker Recognition. Proceedings of the (th International Conference on Spoken Language Processing, Jeju, Korea, 2004.

Arcienaga M, Alexander A, Zimmermann P, Drygajlo A. A Bayesien Network Approach Combining Pitch and Spectral Envelope Features to Reduce Channel Mismatch in Speaker Verification and Forensic Speaker Recognition. Interspeech'2005 - Eurospeech - 9th European Conference on Speech Communication and Technology, September 2005, Lisbon, Portugal [to appear]

Atal BS. Automatic recognition of speaker from their voices. Proc. IEEE, 1976; 64 (4): 460.

Beck SD, Schwartz R, Nakasone H. A Bilingual Corpus for Language and Speaker Recognition (LASR) Services. *ODYSEE 2004*: 265-270.

Bertillon A. Une application de l'anthropométrie sur un procédé d'identification. Annales de Démographie Internationales. Paris : G. Masson, 1881.

Botti F, Alexander A, Drygajlo A. Evaluation of evidence in Forensic speaker recognition in Forensic Speaker Recognition with a Questioned recording and a Single suspect's Recording. Proceedings of the 3rd European Academy of Forensic Science Meeting, Instanbul, Turkey. *Forensic Science International* 2003; 136: 365-366.

Botti F, Alexander A, Drygajlo A. An Interpretation Framework for the Evaluation of Evidence in Forensic Automatic Speaker Recognition with Limited Suspect Data. The Speaker and Language Recognition Workshop, Toledo. *Speaker Odyssey* 2004: 63-68.

Braun A. Procedures and perspectives in forensic phonetics; Proceeding of the XII International Congress of Phonetic Sciences, Stockholm, 1995; 3: 146-153.

Bunge E. The role of pattern recognition in forensic science : an introduction to methods. In: Kube E, Störzer HU, Clarke RV, ed. Police Research in the Federal Republic of Germany. 15 Years Research with the Bundeskriminalamt. Berlin: Springer-Verlag, 254-265.

Champod C, Meuwly D. The Inference of Identity in Forensic Speaker Recognition. *Speech Communication* 2000; 31: 193 – 203.

Dessimoz D. Reconnaissance de locuteurs: Comparaison de performances entre la reconnaissance auditive par des profanes et des systèmes automatiques. Séminaire de 4ème année. Institut de Police Scientifique, Université de Lausanne, 2004.

Evett IW, Buckleton JS. Statistical Analysis of STR data. In: ed Carraredo A, Brinkmann B, Bär W. Advances in Forensic Haemogenetics. Heidelberg: Springer-Verlag 1996; 6: 79-86.

Hermansky H. Perceptual linear predictive (PLP) analysis of speech. *Journal of Acoustic Society American* 1990; 87 : 1738-1752.

Hermansky H, Morgan N. RASTA processing of speech; IEEE Trans. ASSP 1994; 2: 578, 589.

Jayant N. High-quality coding of telephone speech. In: Furui, Sondhi, ed. Advances in speech signal processing. New York: Dekker, 1992: 85-108.

Jensen FV. Bayesien Network and Decision Graphs. Springer-Verlag New, 2001.

Kersta LG. Voiceprint identification (A). *Nature* 1962; 4861: 1253-1257.

Kersta LG. Voiceprint identification (B). *Journal of Acoustic Society of America* 1962; 34:725.

Koolwaaji J, Boves L. On Decision Making in Forensic Case Work. *Forensic Linguistics, the International Journal Of Speech Language and the Law* 1999; 2:242-264.

Künzel HJ. Schprechererkennung : Grundzüge forensicher Sprachverarbeitungen, Heidelberg: Kriminalistic Verlag, 1987.

Künzel HJ. Currant approaches to forensic speaker recognition. Proceeding of ESCA Workshop on automatic speaker recognition, identification and verification 1994: 135-141.

Locard E. Les preuves de l'identité - le signalement. Lyon : Joannès Desvignes et ses fils, 1932

Meuwly D. Reconnaissance de locuteurs en Sciences Forensiques : L'apport d'une approche automatique. Institut de Police Scientifique, Université de Lausanne, 2001.

Meuwly D, Drygajlo A. Reconnaissance Automatique de Locuteurs en Sciences Forensiques: Modélisation de la Variabilité Intralocuteur et Interlocuteur. *5ème Congrès français d'Acoustique,* Lausanne, 2000: 522-525.

Meuwly D, Drygajlo A. Forensic Speaker Recognition Based on a Bayesien Framework and Gaussian Mixture Modelling (GMM) Proceedings of the « 2001: A Speaker Odyssey, The Speaker Recognition Workshop », Crete, Greece, June 18-22, 2001.

Ottolenghi S. Trattato di polizia scientifica. Milano : Società Editrice Libreria, 1910 : 272-276.

Pearl J. Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Ville: San Fransisco Morgan Kaufmann Publishers, 1998.

Pruzanski S. Pattern-matching procedure for automatic talker recognition. *Journal of Acoustic Society of America* 1963; 35: 2041-2047.

Reynolds DA. A Gaussian Mixture Modeling approach to Text-Independant speaker identification. Georgia Institut of Technology, Atlanta, USA 1992.

Reynolds DA, Rose RC. Robust text-independent speaker identification using Gaussian mixture speaker models. IEEE Trans ASSP; 3 (1): 72-83.

Ribary G. Etude des performances du système de reconnaissance du locuteur ASPIC sur la base de donnée de l'IASCP. Séminaire de 4ème année. Institut de Police Scientifique, Université de Lausanne, 2002.

Rossy Q. Simulation de cas réels de reconnaissance de locuteurs au moyen du logiciel ASPIC. Séminaire de 4ème année. Institut de Police Scientifique, Université de Lausanne, 2003.

Steinberg JC. Application of sound measuring instrument to the study of phonetic problems. *Journal of Acoustic Society of America* 1934; IV: 16-24.

Tippet CF, Emerson VJ, Fereday MJ, Lawton F, Lampert SM. The evidential value of the comparison of point flakes from sources other than vehicles. *Journal of Forensic Sciences* 1968*;* 8: 61-65.