CRIMISCOPE

IPSC - UNIL - 1015 Lausanne

Numéro 10 – septembre 2000

(II) RECONNAISSANCE DE LOCUTEUR: TRAVAIL POUR L'HOMME OU L'ORDINATEUR ?

En bref...

La reconnaissance de locuteurs en sciences forensiques est un problème non résolu. Plusieurs méthodes d'analyse sont utilisées : l'approche phonétique, spectrographique et automatique.

La première partie de cet article discute de leur validité et de leur fiabilité pour une utilisation dans le cadre.

La seconde partie de l'article décrira le fonctionnement d'un système automatique spécialement adapté aux sciences forensiques grâce à une interprétation de la preuve basée sur le théorème de Bayes. Il présentera quelques résultats provenant de la procédure d'évaluation de ce système. Finalement il mettra l'accent sur la nécessité de constituer des bases de données des populations pertinentes, et proposera l'utilisation de moyens techniques actuels pour la récolte des indices.

Présentation d'un système de reconnaissance automatique basé sur GMM

Avec une méthode de reconnaissance automatique de locuteurs, comparer un enregistrement provenant d'une écoute téléphonique à l'enregistrement de comparaison d'une personne suspectée revient premièrement à réaliser un modèle de la voix de la personne suspectée.

Enregistrement

La personne suspectée est enregistrée à partir du réseau téléphonique directement sur un ordinateur par l'intermédiaire d'une carte d'acquisition ISDN. Cette étape peut donc être réalisée depuis n'importe quel endroit muni d'un téléphone, comme un commissariat de police ou un lieu de détention préventive.

Deux types d'enregistrement sont réalisés. Une première série d'enregistrement est réalisée de manière à modéliser la voix de la personne suspectée alors que la seconde est réalisée dans le but d'évaluer l'intra-variabilité de la voix de la personne suspectée.

Modélisation de la voix de la personne suspectée

Pré-traitement

La série d'enregistrement réalisée pour la modélisation de la voix de la personne suspectée subi un pré-traitement automatique de suppression des silences, de manière à séparer les endroits de parole des endroits de silence.

Extraction des caractéristiques

Les caractéristiques dépendantes du locuteur sont ensuite extraites du signal prétraité par une méthode dérivée de la prédiction linéaire, Linear Prediction Coding (LPC), qui permet une caractérisation de l'enveloppe spectrale du signal dans le domaine temporel. Cette méthode est appelée prédiction linéaire perceptuelle, Perceptual Linear Prediction (PLP) du fait que la composante fréquentielle du signal de parole n'est plus décrite sur une échelle linéaire, mais sur une échelle non linéaire dont la résolution est proche de celle de l'oreille humaine [HERMANSKY, 1990]. Dans les faits l'extraction est réalisée sur des segments de signal de 20 ms. Douze paramètres PLP sont extraits de chaque segment, ce qui signifie que le signal entier est décrit par douze séries de paramètres PLP. Dans chaque série, le nombre de paramètres est directement dépendant de la longueur du signal analysé.

Modélisation par GMM

Cette étape consiste à modéliser la distribution de chacune des douze séries de paramètre PLP par un ensemble de 64 fonctions de densité gaussiennes. Chaque fonction de densité gaussienne étant définie par sa moyenne et son écart-type, chaque distribution de paramètres PLP est modélisée par 64 moyennes et 64 écarts-type. Le signal de parole entier, et donc la voix qui l'a produit, est modélisé par douze séries de 64 moyennes et 64 écarts-types. Finalement, cette information est organisée dans une matrice de covariance, qui constitue le modèle de la voix de la personne suspectée [REYNOLDS, 1995].

Comparaison modèle-indice : détermination de l'élément de preuve

Le modèle ainsi réalisé peut être comparé à l'indice, de manière à mesurer leur proximité. Cette mesure constitue l'élément de preuve. L'indice subit aussi un pré-traitement, de manière à supprimer les silences, puis une procédure d'extraction des paramètres PLP. L'adéquation entre les paramètres extraits de l'indice et la distribution de ces mêmes paramètres, décrite dans le modèle, est calculée et synthétisée sous forme d'un indice de proximité ; d'un point de vue mathématique, le système calcule la probabilité de l'indice, dans l'hypothèse ou il provient du modèle. De manière plus générale, tous les systèmes de perception artificielle, comme les systèmes automatiques d'identification de traces digitales *Automatic Fingerprint Identification System* (AFIS) ou les systèmes de reconnaissance de locuteurs, fournissent soit un indice de distance soit un indice de proximité entre deux échantillons comparés.

Interprétation de l'élément de preuve

Introduction

Lorsqu'une recherche fondamentale est développée par des ingénieurs dans un domaine comme l'intelligence artificielle perceptive, l'application de cette connaissance en sciences forensiques se décompose en deux étapes. La première est théorique et vise à réaliser une méthode de reconnaissance valide, du point de vue méthodologique.

A partir de la valeur de l'indice de proximité calculé par un système de reconnaissance automatique de locuteurs comme GMM, est-il possible d'identifier ou d'exclure formellement un locuteur comme étant la source de l'enregistrement de test?

La réponse est non car un système de reconnaissance automatique de locuteurs est un outil de mesure et non un outil de décision binaire d'identification ou de vérification. Il fournit un élément de preuve, sous forme d'une valeur numérique, mais il manque le canevas permettant d'interpréter cette mesure.

Une analogie permet de mieux comprendre ces deux niveaux d'analyse. Comme un système de perception artificielle, un ruban métrique est un outil de mesure. Il permet de calculer la distance entre deux objets avec une certaine précision, mais il ne permet pas de répondre dans l'absolu si ces deux objets sont proches ou non. Tenter de répondre à cette question nécessite aussi la définition d'un canevas permettant d'interpréter cette distance.

Lausanne est-elle proche de Lucerne ? Dans l'absolu il est impossible de répondre à cette question, mais si la distance de référence est Lausanne - Evian, alors la réponse est non ; par contre si la distance de référence est Lausanne - Berlin, alors la réponse est oui.

De la même manière, déterminer si la proximité entre le modèle de la voix d'une personne suspectée et d'un indice est grande ou non nécessite la définition d'un canevas permettant une interprétation cohérente de cet élément de preuve.

Définition d'un cadre de référence

Evaluation de l'intravariabilité

Cette première étape consiste à définir la distribution des indices de proximité fournis par le système de reconnaissance lorsque le modèle de la voix de la personne suspectée est comparé à des enregistrements de comparaison de cette même personne, de manière à évaluer l'intravariabilité de sa voix (*within-source variability*) (Fig. 1).

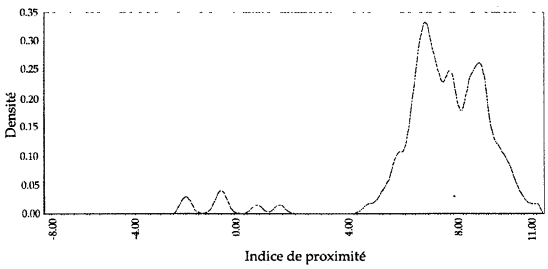

Fig. 1: Exemple de distribution des indices de proximité issu d'une procédure d'évaluation de l'intravariabilité

Pour atteindre ce but, le modèle de la voix de la personne suspectée est comparé au second type d'enregistrement réalisé avec cette personne, qui comprend des enregistrements de parole lue et de parole spontanée, l'usage de téléphones mobiles et fixes et éventuellement l'enregistrement de séquences distantes de quelques jours ou quelques

semaines, de manière à tenir compte de la modification de la voix dans le temps.

Evaluation de l'intervariabilité

Cette seconde étape débute par l'écoute de l'indice, afin de déterminer la langue parlée, l'accent et le genre du locuteur, lorsque cela est possible. Ces paramètres servent à définir la population potentielle des locuteurs qui ont pu produire l'indice, par exemple l'ensemble des locuteurs masculins s'exprimant en français avec un accent de Suisse Romande.

Comme il est impossible de recueillir un énoncé de parole de chacun des locuteurs suisse romands, la population potentielle est modélisée à l'aide d'une base de données, la plus représentative possible. En Suisse, l'entreprise Swisscom® possède p. ex. une base de données de modélisation des locuteurs suisses contenant 5000 francophones, 5000 italophones. Pour l'instant ce type de base de données n'existe que dans les pays économiquement les plus développés, alors que les langues présentes dans les écoutes téléphoniques sont généralement celles pratiquées dans des pays moins développés.

Une solution à ce problème consisterait à trier les enregistrements provenant des écoutes téléphoniques selon les langues et les accents, et de constituer ainsi des bases de données modélisant de manière très pertinentes les populations potentielles. Il est regrettable de constater que l'accès à la source des données les plus pertinentes pour la modélisation des différentes populations potentielles échappe aux auxiliaires de la justice dans le contexte légal actuel.

Lorsque la base de données existe ou a été constituée, un modèle de la voix de chaque locuteur de la base de donnée est constituée selon la méthode exposée supra. Cette base de données sert à définir la distribution des indices de proximité fournis par le système de reconnaissance de locuteurs, lorsque l'indice est comparé à un locuteur de même langue et de même accent, choisi au hasard dans la population potentielle (*between-source variability*) (Fig. 2).

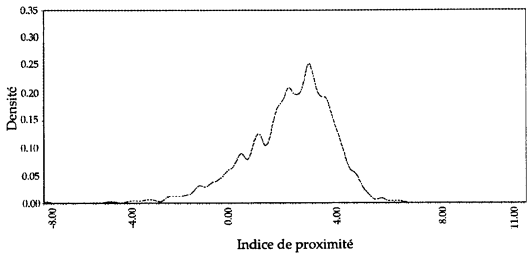

Fig. 2: Exemple de distribution des indices de proximité issu d'une procédure d'évaluation de l'intervariabilité

Interprétation de l'élément de preuve

L'évaluation de l'intravariabilité de la personne suspectée et de l'intervariabilité de l'indice dans la population potentielle rend possible l'interprétation de l'élément de preuve, dans un canevas cohérent. La vraisemblance de l'élément de preuve est calculée dans les hypothèses alternatives suivantes:

H1 : La personne suspectée est la source de l'indice

H2 : Une autre personne de la population potentielle est la source de l'indice

La vraisemblance de l'élément de preuve dans l'hypothèse H1 est calculée comme la probabilité de l'élément de preuve dans la distribution de l'intravariabilité. Si, par exemple l'élément de preuve vaut 8.00 alors sa vraisemblance est d'environ 0.17. Par contre si l'élément de preuve vaut 4.00, sa vraisemblance est proche de 0, environ 10^{-3} (Fig. 3).

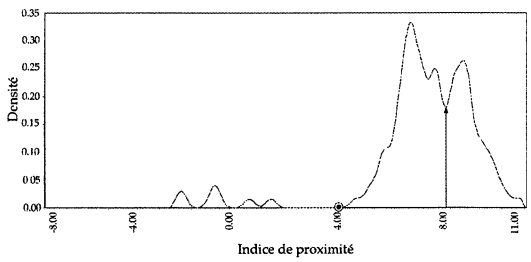

Fig. 3: Evaluation de la vraisemblance de deux éléments de preuve (4.00 et 8.00) dans la distribution de l'intravariabilité.

La vraisemblance de l'élément de preuve dans l'hypothèse H2 est calculée comme la probabilité de l'élément de preuve dans la distribution de l'intervariabilité. Si, par exemple l'élément de preuve vaut 8.00 alors sa vraisemblance est proche de 0, environ 10^{-3}. Par contre si l'élément de preuve vaut 4.00, sa vraisemblance vaut environ 0.15 (Fig. 4).

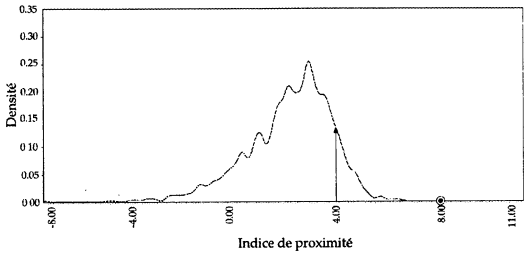

Fig. 4: Evaluation de la vraisemblance de deux éléments de preuve (4.00 et 8.00) dans la distribution de l'interva-riabilité.

Ces deux résultats permettent finalement d'évaluer un rapport de vraisemblance de l'élément de preuve dans le cas ou soit l'hypothèses H1, soit l'hypothèse H2, est vraie. Ce rapport de vraisemblance rend compte de la pertinence du choix de la personne suspectée comme source potentielle de l'indice.

Dans l'exemple proposé, lorsque l'élément de preuve vaut 8, sa vraisemblance est 170 fois ($0.17/10^{-3}$) plus grande dans l'hypothèse ou H1 est vraie que dans l'hypothèse ou H2 est vraie (Fig. 5). Le LR vaut 170. Par contre lorsque l'élément de preuve vaut 4, sa vraisemblance et est 150 fois ($0.15/10^{-3}$) plus petite dans l'hypothèse ou H1 est vraie que dans l'hypothèse ou H2 est vraie (Fig. 5). Le LR vaut 1/150.

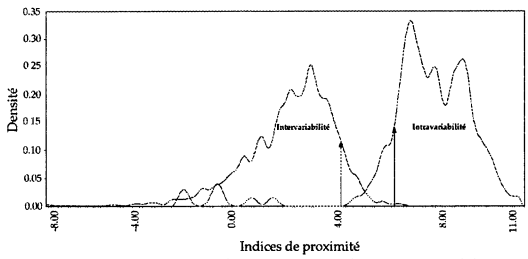

Fig. 5: Evaluation du rapport de vraisemblance de deux éléments de preuve (4.00 et 8.00) dans l'hypothèse ou soit H1 soit H2 est vraie.

Evaluation du système de reconnaissance

Introduction

Lorsque la validité de la méthode est acceptée, une seconde étape, plus pratique, consiste à évaluer sa fiabilité dans des conditions aussi proches que possibles de celles qui pourront être rencontrées dans une utilisation pratique.

Les variables susceptibles d'influencer la fiabilité d'une méthode de reconnaissance automatique appliquée en sciences forensique ont été largement développées supra. Une présentation des résultats pour quelques-unes de ces variables analysées permettent de mieux juger de la robustesse de la méthode GMM.

Principe

Le principe d'évaluation de la méthode consiste à estimer et à comparer la distribution des rapports de vraisemblance qui peuvent être obtenus à partir de l'élément de preuve, d'une part lorsque l'hypothèse H1 est vérifiée, c'est à dire lorsque la source du modèle et de l'indice est la même, et d'autre part lorsque l'hypothèse H2 est vérifiée, c'est à dire lorsque la source du modèle et de l'indice est différente.

Présentation des résultats

Pour la présentation des résultats, le choix s'est porté sur un mode de présentation proposé par EVETT ET BUCKLETON dans le domaine de l'interprétation de l'analyse génétique [EVETT ET BUCKLETON , 1996].

L'axe des abscisses est gradué en termes de valeurs croissantes de rapports de vraisemblance (LR). L'axe des ordonnées indique la probabilité que le résultat du test excède une valeur de LR donnée. Chaque représentation graphique comporte deux courbes ; la première rend compte de l'évolution des rapports de vraisemblance estimés lorsque l'hypothèse H1 est vérifiée et la seconde rend compte de la distribution des rapports de vraisemblance estimés lorsque l'hypothèse H2 est vérifiée. N représente le nombre d'expériences réalisées.

Résultats

Une première expérience montre l'influence du temps entre le moment de l'enregistrement de l'indice et du modèle de la voix de la personne suspectée. L'indice a été enregistré avant le modèle, et les deux enregistrements ont été réalisés à partir du même téléphone.

Logiquement les rapports de vraisemblance évalués par la méthode sont plus élevés lorsque l'hypothèse H1 est vérifiée que lorsque H2 est vérifiée, mais les performances décroissent faiblement lorsque l'intervalle entre l'enregistrement de l'indice et du modèle augmente (de 0 j. à 60 j.) (Fig. 6).

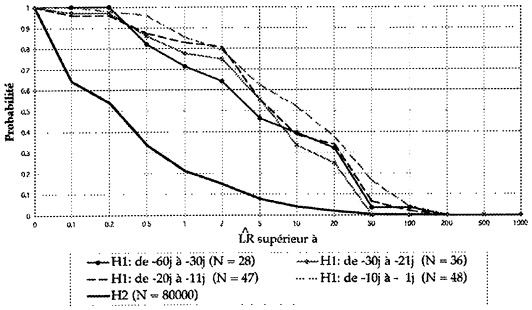

Fig. 6 : Évolution des rapports de vraisemblance en fonction de l'intervalle de temps séparant l'enregistrement du modèle de celui de l'indice

Une deuxième expérience montre l'influence de la présence d'un déguisement de la voix dans l'indice. Lorsque le même téléphone est utilisé pour l'enregistrement du modèle et de l'indice, les rapports de vraisemblance évalués par la méthode restent plus élevés lorsque l'hypothèse H1 est vérifiée que lorsque H2 est vérifiée, même si la présence d'un déguisement détériore les performances de manière importante (Fig. 7).

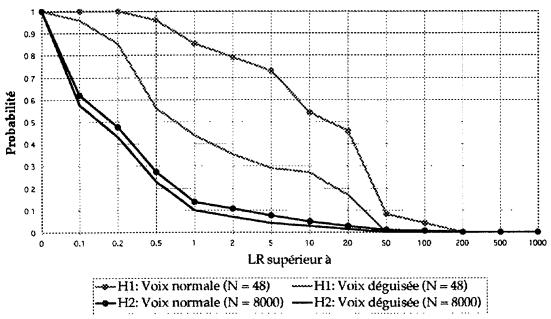

Fig. 7: Évolution des rapports de vraisemblance en fonction de l'absence ou de la présence de déguisement dans l'indice (même téléphone utilisé pour l'enregistrement du modèle et de l'indice)

Par contre, lorsqu'un téléphone différent est utilisé pour l'enregistrement du modèle et de l'indice et qu'il existe de surcroît un déguisement de la voix dans l'indice, les rapports de vraisemblance évalués par la méthode sont équivalents, que l'hypothèse H1 ou H2 soit vérifiée, ce qui signifie qu'aucun résultat probant ne peut être attendu dans de telles conditions (Fig. 8).

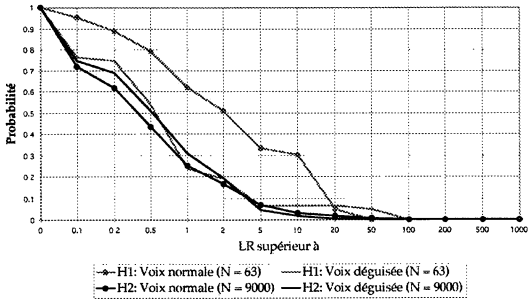

Fig. 8 : Évolution des rapports de vraisemblance en fonction de l'absence ou de la présence de déguisement dans l'indice (téléphone différent utilisé pour l'enregistrement du modèle et de l'indice)

Une troisième expérience montre l'influence de la présence d'un bruit de fond constitué de parole dans l'indice. Le rapport signal sur bruit (RSB) d'une ligne téléphonique normale est de l'ordre de 40 dB. Des conditions de fort bruit de fond (RSB = 12 dB ou 6 dB) détériorent les performances de la méthode de manière importante, mais les rapports de vraisemblance évalués par la méthode restent plus élevés lorsque l'hypothèse H1 est vérifiée que lorsque H2 est vérifiée. Par contre, lorsque le bruit de fond est aussi fort que le signal (rapport signal sur bruit de 0 dB), les rapports de vraisemblance évalués par la méthode sont équivalents, que l'hypothèse H1 ou H2 soit vérifiée, ce qui signifie qu'aucun résultat probant ne peut être attendu dans de telles conditions (Fig. 9).

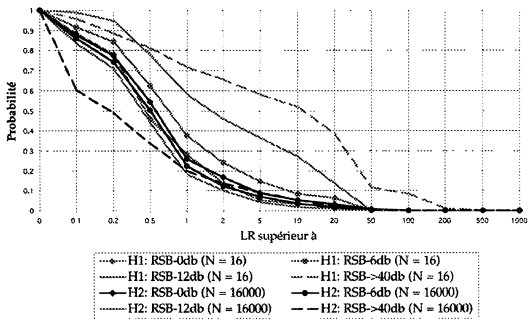

Fig. 9 : Évolution des rapports de vraisemblance en fonction de l'absence ou de la présence de bruit de fond lors de l'enregistrement de l'indice

Une dernière expérience montre l'influence du système utilisé pour l'enregistrement de l'indice. Dans les deux cas l'indice est enregistré par l'intermédiaire du réseau téléphonique ; dans le premier cas il est enregistré directement sur un ordinateur par l'intermédiaire d'une carte d'acquisition ISDN et dans le second il est enregistré sur l'enregistreur analogique d'une police cantonale.

Les résultats montrent que le système d'enregistrement analogique détériore la qualité de l'information à un point tel que les rapports de vraisemblance évalués par la méthode sont équivalents, que l'hypothèse H1 ou H2 soit vérifiée. Ceci signifie que l'information pertinente est perdue à cause du système d'enregistrement lors de la récolte de l'indice, alors que cette information est sauvegardée en cas d'utilisation d'enregistrement numérique (Fig. 10).

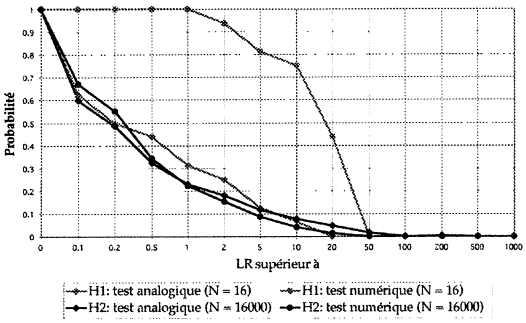

Fig. 10 : Évolution des rapports de vraisemblance en fonction du système utilisé pour l'enregistrement de l'indice

Conclusion

Perception naturelle ou artificielle ?

La voix est en premier lieu un comportement humain et ensuite seulement un signal, qui est la transposition imparfaite de ce comportement. Le système de perception naturel, le système auditif humain, est plus sophistiqué que n'importe quel système de perception artificielle. Grâce à une analyse parallèle et simultanée de plusieurs niveaux de communication il distingue à la fois la compréhension du contenu du message, l'état émotionnel du locuteur et son intention. De plus, le système auditif reste le seul système de perception efficace dans des conditions extrêmes de bruit de fond.

Analyse subjective ou objective ?

Le traitement de l'information auditive par le cerveau humain se révèle par contre imparfaite pour la reconnaissance de locuteurs dans une application forensique.

Lorsque cette tâche est réalisée par des non-experts, victime ou témoin, les résultats ont une valeur indicative dont l'incertitude est comparable à celle du témoignage.

Lorsqu'elle est réalisée par des experts, sur la base d'une analyse phonétique (auditive-perceptive et phonétique-acoustique), cette tâche nécessite une parfaite connaissance de la langue de l'indice par l'expert, sous peine de confondre des caractéristiques dépendantes de la langue ou de l'accent, et les caractéristiques réellement dépendantes du locuteur. Comme le mode de raisonnement humain par analogie s'accommode mal des probabilités statistiques, l'évaluation subjective de la rareté de caractéristiques dépendantes du locuteur peu fréquentes est variable selon les experts, d'où des évaluations potentiellement très différentes. Une interprétation des résultats dans un canevas cohérent reste difficile à envisager, à cause du coût très élevé de l'approche phonétique en temps/homme, si elle est réalisée dans règles de l'art. Finalement, un spécialiste ne peut développer des connaissances suffisantes que dans sa langue maternelle et dans un nombre limité d'autres langues.

L'utilisation de méthodes automatiques pour accomplir cette tâche de reconnaissance de locuteurs en sciences forensiques se heurte encore prioritairement à l'imperfection des systèmes de perception artificielle, qui restent sensibles aux divers types de variation du signal de parole rencontrés dans les cas pratiques. L'avantage de l'utilisation de ce type de méthodes réside principalement dans le traitement des données. Premièrement, n'importe quelle langue peut être abordée, puisque la connaissance de base dépend d'une base de données de locuteurs ; la méthode est donc susceptible de s'adapter très vite à de nouvelles conditions de criminalité. Deuxièmement, la rareté des caractéristiques extraites par une méthode automatique est évaluée selon des probabilités statistiques, plus proche de la réalité que les probabilités subjectives. Troisièmement une interprétation dans un canevas cohérent est seulement coûteuse en temps/machine en non en temps/homme, ce qui ne représente pas un obstacle au vu des performances actuelles et futures de l'outil informatique. Finalement, l'approche automatique permet d'effectuer rapidement et pour un coût très faible des simulations, de manière à pouvoir déterminer a priori, en analysant uniquement l'indice, la probabilité qu'une analyse amène un résultat probant (*preassessment*).

Récolte des indices et constitution de bases de données

D'un point de vue plus général, le besoin dans le domaine de la reconnaissance de locuteurs est croissant de la part du système pénal. Actuellement la demande avoisine une quarantaine par an, et il existe certainement un nombre important de demandes non explicitées, du fait que le manque de fiabilité des méthodes proposées est une information connue dans ce domaine.

Il est cependant nécessaire de mentionner ici que les résultats qui peuvent être attendus de n'importe quel type d'analyse dépend d'abord de la qualité de l'échantillon soumis pour analyse. La technologie utilisée par les polices suisses pour la récolte des indices ne répond plus du tout aux standards de qualité actuels et il est urgent de les équiper de matériel actuel si l'on veut pouvoir procéder à des analyses en reconnaissance de locuteurs.

Le système idéal devrait premièrement être capable d'enregistrer les deux interlocuteurs d'une conversation téléphonique sur deux canaux audio indépendants ; cette séparation des canaux permettrait d'éviter d'une part une corruption inutile du signal par du bruit de fond et d'autre part de lever tout doute quant à l'origine des énoncés. Deuxièmement le système devrait être capable d'enregistrer bit à bit de manière numérique le signal transmis sur le réseau téléphonique, en évitant toute compression des données avec perte, de façon à sauvegarder l'intégralité de l'information.

L'interprétation des résultats d'analyse nécessite l'utilisation de bases de données. L'idéal serait de pouvoir utiliser les enregistrements provenant des écoutes téléphoniques pour constituer des bases de données modélisant de manière très pertinente les populations potentielles de locuteurs, puisque la plupart des indices sont énoncés dans des langues pour lesquelles il n'existe aucune base de données.

Pour l'instant cette précieuse information transite par les autorités de police, chargées de la récolte des indices, mais elle n'est pas exploitée en temps que matière première pour la constitution de bases de données. Une utilisation de cette information dans ce but nécessite des aménagements juridiques et organisationnels, mais c'est à ce prix que des méthodes de reconnaissance de locuteurs réellement fiables pourront être proposées au système pénal.

Bibliographie

HERMANSKY H (1990) Perceptual linear predictive (PLP) analysis of speech. J. Acoust. Soc. Am. 1738 - 1752.

REYNOLDS DA (1995) Automatic Speaker Recognition Using Gaussian Mixture Speaker Model. The Lincoln Labora-tory Journal. 8 : 173 - 191.

EVETT IW and BUCKLETON JS (1990) Statistical analysis of STR data in: Advances in Hoemogenetics (eds: Carraredo A, Brickmann B, Bär W) Springer-Verlag, Heidelberg Germany.

Rédacteur de ce numéro:
Didier Meuwly

Rédaction: Prof. P. Margot et Prof. M. Killias, IPSC, UNIL, 1015 Lausanne

Veuillez adresser vos remarques et communications à:

Secrétariat de *Crimiscope*
UNIL - Institut de police scientifique et de criminologie
CH–1015 LAUSANNE

☎ (021) 692 46 42
Fax (021) 692 46 05
Int. (+ 41 21) 692 46 42