CRIMISCOPE

IPSC - UNIL - 1015 Lausanne Numéro 8 – avril 2000

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 criminalistique.

La seconde partie de l'article décrit 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 présente quelques résultats provenant de la procédure d'évaluation de ce système. Finalement il met l'accent sur la nécessité de constituer des bases de données des populations pertinentes, et propose l'utilisation de moyens techniques actuels pour la récolte des indices.

Introduction

Dans le film culte de Stanley Kubrick « 2001, l'Odyssée de l'espace » (1968), le héros franchit les frontières en présentant sa voix en guise de passeport. La science fiction et notre expérience journalière de la reconnaissance de personnes par leur voix laissent supposer que la reconnaissance de locuteurs est un problème résolu, tant pour l'homme que pour l'ordinateur.

Il est vrai que l'être humain a une bonne aptitude à identifier des voix de personnes familières et que les résultats de la reconnaissance automatique de locuteurs sont convaincants dans des conditions de laboratoire.

Cependant des circonstances aussi favorables sont rarement rencontrés dans des cas réels :

- Le locuteur, enregistré à travers la ligne téléphonique dans un cas de kidnapping ou de chantage, a souvent peur ou est stressé ; sa voix s'en trouve modifiée. De plus elle n'est pas familière de la police.
- Les trafiquants de drogue sont souvent l'objet d'écoutes téléphoniques, mais à cause de leur activité à la fois nomade et illégale, et parce qu'ils suspectent une surveillance policière, ils recourent à des cabines téléphoniques publiques ou à des téléphones cellulaires volés, utilisés dans des environnements bruyants, par exemple dans la rue, dans une voiture ou dans un bar.

La reconnaissance de locuteurs en sciences forensiques n'est donc pas un problème résolu dans des conditions aussi adverses. Dès lors, il est intéressant de connaître les procédures et le type de réponse qui peuvent être attendues dans ce domaine, de la part de l'homme, expert ou non, et de l'ordinateur. Parallèlement, il est nécessaire, voire urgent de s'interroger quant aux améliorations techniques à apporter dans la phase de récolte de l'indice que peut représenter l'enregistrement d'une voix.

La voix humaine

La production de la parole est composée de deux fonctions mécaniques de base : la phonation et l'articulation. La phonation est la production du signal acoustique par les cordes vocales. L'articulation inclut la modulation du signal acoustique par les articulateurs (principalement les lèvres, la langue et le palais mou) et la résonance de ce signal dans les cavités supra-glottiques (la bouche et le nez). Les phonèmes produits sont divisés en consonnes sourdes et voisées et en voyelles ; ils peuvent être caractérisés dans le domaine temporel, dans le domaine spectral et dans le domaine spectro-temporel. L'étendue spectrale du signal de la parole normale se situe entre 80 Hertz (Hz) et 8000 Hz, avec une étendue dynamique de 60 à 70 décibels (dB). La fréquence fondamentale moyenne de vibration des cordes vocales (F0), appelée « pitch », est située entre 180 et 300 Hz pour les femmes, entre 300 et 600 Hz pour les enfants et entre 90 et 140 Hz pour les hommes.

La perception de la parole est généralement décrite comme une transformation en cinq étapes du signal de parole en message : l'analyse auditive périphérique, l'analyse auditive centrale, l'analyse acoustique-phonétique, l'analyse phonologique et l'analyse d'ordre supérieure : lexicale, syntaxique et sémantique. L'oreille humaine est tout d'abord conçue pour la perception de la voix humaine. L'étendue de la percep-

tion spectrale humaine est située entre 16 et 20000 Hz, avec une excellente sensibilité entre 500 et 4000 Hz. La limite reconnue dans le domaine de l'intensité se situe entre 130 et 140 dB.

La voix comme indice

Récolte de l'indice

L'indice ne consiste pas en la voix elle-même, mais en une transposition obtenue par un transducteur, qui convertit l'énergie acoustique en une autre forme d'énergie : mécanique, électrique ou magnétique. Cette transposition est enregistrée sur un support de mémorisation, sur lequel elle est codée par une méthode de codage de l'information, analogique ou numérique. Dans un enregistrement analogique, l'intensité et la forme du signal sonore transcodé conserve une relation directe avec le son original. Dans un enregistrement numérique, par contre, le signal sonore transcodé est échantillonné, et chaque échantillon est traduit en un code binaire, comme ceux utilisés par les ordinateurs.

Type de l'indice

L'information enregistrée a généralement été acheminée par l'intermédiaire du réseau téléphonique, que ce soit dans le cadre d'une mesure officielle de surveillance [art. 179octies CP] ou d'un abus de téléphone [art. 179septies CP]. En Allemagne, 80% des demandes de reconnaissance de locuteurs concernent des cas impliquant une mesure de surveillance et les 20% restants concernent des abus de téléphone [KÜNZEL, 1994].

Dans le cas d'abus de téléphone, le message anonyme est généralement court, de quelques secondes à quelques minutes ; s'il s'agit d'un monologue, il peut avoir été préenregistré et volontairement modifié par un filtrage ou une procédure montage. D'un point de vue lexical, les thèmes sont ciblés, p. ex. extorsion, insultes, obscénités, menaces… Lors d'une mesure de surveillance, les enregistrements peuvent atteindre des centaines d'heures. Les modifications volontaires par filtrage ou procédure montage sont moins probables, à moins qu'elles ne soient l'œuvre de l'autorité qui procède à l'enregistrement. Leur contenu lexical est étendu et certaines expressions peuvent faire référence à des codes internes propres à un groupe ou à une organisation.

Variabilité du locuteur et déguisement

Indice

Lors d'un appel anonyme, la voix peut être d'une part altérée de manière involontaire par le locuteur pour des raisons particulières de stress et/ou de peur, liée à la commission d'une infraction. L'état de santé, la consommation de tabac et de substances psychotropes peuvent aussi affecter la voix. D'autre part, le locuteur peut aussi modifier délibérément sa voix, sa manière de parler et son langage.

Lors d'une mesure de surveillance, les locuteurs ignorent généralement qu'ils sont enregistrés. Dans de tels cas, il n'y a généralement aucune des altérations involontaires et délibérées décrites ci-dessus. Par contre, la spontanéité découlant de cette ignorance laisse au locuteur de substantielles possibilités d'adaptation de son discours en fonction du contexte, de son humeur et des relations qu'il entretient avec son interlocuteur.

Enregistrement de comparaison

Si une personne est suspectée et appréhendée sur la base d'un enregistrement téléphonique, il est possible de réaliser avec elle un enregistrement de comparaison de sa voix. Cet enregistrement devrait être réalisé dans des conditions similaires aux conditions d'enregistrement de l'indice, de manière à fournir une image représentative de la voix de cette personne.

Cette procédure est souvent réalisée plusieurs mois ou plusieurs années après l'enregistrement de l'indice et il est reconnu que la voix se modifie au cours du temps. La personne suspectée peut aussi se montrer non coopérative. Cet ensemble de raisons peut conduire à la limitation ou à l'impossibilité d'entreprendre une procédure de reconnaissance de locuteur efficace.

Variabilité du système d'enregistrement

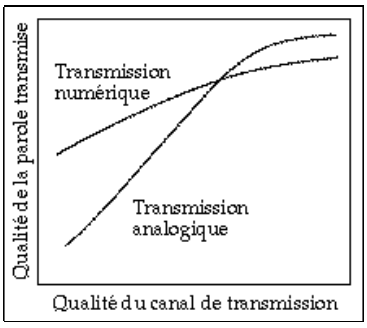
Distorsions du canal de transmission

La transmission téléphonique limite la bande passante de la parole entre 300 et 3400 Hz : plusieurs caractéristiques hautement dépendantes du locuteur de fréquence élevée, tout comme la fréquence fondamentale, ne sont pas transmises. L'auditeur perçoit tout de même la fréquence fondamentale par l'intermédiaire des multiples entiers de la fréquence fondamentale, appelées harmoniques.

L'étendue dynamique de la ligne téléphonique est limitée entre 30 et 40 dB et le signal souffre de distorsions non linéaires et d'addition de bruit de fond collecté par le microphone du combiné téléphonique.

La transmission téléphonique implique un nombre d'autres dégradations du signal de parole, qui dépendent à la fois des caractéristiques du microphone et du réseau téléphonique. Ce réseau peut être fixe ou mobile, et le système de codage et de transmission de l'information peut être analogique ou numérique. Dans les pays économiquement développés comme la Suisse, le codage et la transmission sont essentiellement numériques, alors qu'ils sont encore analogiques ou mixtes dans les pays moins développés (Fig. 1).

Fig. 1: Comparaison entre qualité de transmission téléphonique analogique et numérique



La qualité des réseaux numériques dépend particulièrement des algorithmes de codage et du débit de transmission de l'information.

Par exemple, la qualité du réseau fixe, dont le débit est de 64 kbit/s, est très supérieure à celle du réseau mobile, dont le débit se situe entre 8 et 16 kbit/s selon les systèmes.

Distorsions du système d'enregistrement

Le système d'enregistrement est le seul maillon de la chaîne qu'il est possible de contrôler lors de la récolte de l'indice, mais c'est malheureusement souvent le plus faible.

Même si le codage et la transmission sont numériques, les enregistrements réalisés dans le cadre de mesures de surveillance sont encore presque exclusivement réalisés sur des enregistreurs analogiques, avec une vitesse de défilement faible, ce qui a pour conséquence une qualité résiduelle faible. La plupart des cantons suisses utilisent encore des systèmes à cassette compacte analogique avec une vitesse d'enregistrement de 1,2 cm/s, alors que les appareils standard ont une vitesse de 4,75 cm/s ; de plus ces systèmes enregistrent simultanément un codage du temps sur la bande magnétique, sous forme d'impulsions sonores audibles et gênantes.

L'enregistrement de messages anonymes est réalisé sur le même type de systèmes d'enregistrements, lorsque le message vise un organisme officiel qui enregistre les conversations téléphoniques entrantes (police, service du feu, service d'urgence), ou sur un simple répondeur téléphonique, lorsque le message vise un particulier.

Reconnaissance de locuteurs

Types de reconnaissance de locuteurs

La reconnaissance de locuteurs renvoie à tout processus qui utilise des caractéristiques du signal de parole pour déterminer si une personne particulière est l'auteur d'un énoncé donné. Trois types d'approches peuvent être distinguées: la reconnaissance de locuteurs par audition, la reconnaissance de locuteurs par comparaison visuelle de spectrogrammes et la reconnaissance automatique de locuteurs.

- La reconnaissance de locuteurs par audition englobe l'étude de la manière dont les êtres humains associent une voix à un individu particulier ou à un

Bibliographie

AITKEN CGG (1995) Statistics and the evaluation of evidence for forensic scientists. John Wiley & Sons, Chichester.

BECKER RW, CLARKE FR, POZA FT AND YOUNG JR (1973) A semi-automatic speaker recognition system. U.S. Department of Justice, Law Enforcement Assistance Administration, National Institute of Law Enforcement and Criminal Justice, Washington.

groupe de locuteurs, et dans quelle mesure cette tâche de reconnaissance peut être remplie.

- La reconnaissance de locuteurs par comparaison visuelle de spectrogrammes s'intéresse à la qualité des décisions d'identité ou de non-identité rendues sur la base de comparaisons visuelles de spectrogrammes vocaux.
- La reconnaissance automatique de locuteurs repose sur des méthodes informatiques basées sur la théorie de l'information, la reconnaissance de forme et l'intelligence artificielle.

Les trois types d'approche sont actuellement utilisés en sciences forensiques. La reconnaissance de locuteurs est pratiquée soit par des experts, phonéticiens ou spécialistes en science de la parole, soit par des non-experts, principalement des victimes ou des témoins. La reconnaissance de locuteurs par comparaison visuelle de spectrogrammes est pratiqués principalement aux Etats-Unis par des examinateurs de spectrogrammes, alors que les méthodes de reconnaissance automatiques sont intégrées dans des systèmes semi-automatiques ou dans des systèmes assistés par ordinateur. La tendance actuelle est d'intégrer les résultats de la reconnaissance par l'audition et de la reconnaissance automatique et d'utiliser la spectrographie uniquement dans un but de visualisation du signal de parole ; toutefois rares sont les laboratoires qui maîtrisent et utilisent tous les types d'approche.

Procédure et méthodes

L'écoute est la première tâche. Ce peut être la seule si l'indice et l'enregistrement de comparaison diffèrent ostensiblement, par exemple si la fréquence fondamentale, le dialecte ou toute autre caractéristique facilement audible diffère. Par conséquent, une procédure de reconnaissance de locuteurs est plus facilement engagée lorsqu'une proximité auditive existe entre l'indice et l'enregistrement de comparaison, à moins qu'un déguisement soit suspecté. Lorsqu'il est engagé, le processus de reconnaissance est constitué de trois phases :

(1) l'extraction des caractéristiques dépendantes du locuteur,

(2) la comparaison de ces caractéristiques;

(3) le processus de décision.

Extraction des caractéristiques

Comme aucune caractéristique spécifique au locuteur n'est actuellement connue et que la répétition d'un même énoncé par le même locuteur varie d'un énoncé à l'autre, l'extraction de caractéristiques présuppose une connaissance des aspects du signal acoustique qui comportent les paramètres les plus dépendants de l'identité du locuteur [NOLAN, 1990]. L'existence de cette variabilité intralocuteur rend la reconnaissance de locuteurs analogue à l'identification d'écriture manuscrite. Les paramètres idéaux devraient : (1) montrer un grand degré de variation d'un locuteur à l'autre (grande variabilité intralocuteur), (2) être dotés de constance dans les énoncés de la même personne (sélectivité), (3) être de préférence insensibles à l'état émotionnel, de santé et au contexte de communication (faible variabilité interlocuteur), (4) résister aux tentatives de déguisement ou d'imitation (résistance), (5) être fréquents dans la parole spontanée (disponibilité), (6) ne pas être perdus ou affectés par la transmission téléphonique ou par le processus d'enregistrement (robustesse) et (7) pouvoir être extraits sans difficulté insurmontable (mesurabilité) [THEVENAZ, 1993]. Malheureusement aucune caractéristique dépendante du locuteur satisfaisant à tous ces critères n'a mise en évidence pour l'instant.

Comparaison des caractéristiques

A partir de l'enregistrement de comparaison, la voix du locuteur doit pouvoir être modélisée, c'est à dire représentée dans un espace à plusieurs dimensions, dont le nombre dépend du nombre de caractéristiques extraites. Un indice de proximité, approprié au type de caractéristiques, est mesuré pour estimer la proximité entre l'enregistrement de comparaison et l'indice. Cette estimation est réalisée implicitement dans une approche subjective et explicitement dans une approche objective.

Dans une approche subjective, comme l'analyse auditive ou spectrographique, l'élément de preuve est exprimé sous forme d'une opinion. Il représente une probabilité subjective qui tient compte des ressemblances et des différences entre l'enregistrement de comparaison et l'indice. Les résultats fournis par ce type d'approche sont contro-

BLACK B, AYALA FJ AND SAFFRAN-BRINKS C (1994) Science and the Law in the Wake of Daubert: A New Search for Scientific Knowledge. Texas Law Review 72 - 4 : 715 - 802.

BOLT RH, COOPER FS, DAVID EE, DENES PB, PICKETT JM AND STEVENS KN (1970) Speaker Identification by Speech Spectrograms: A Scientists' View of its Reliability for Legal Purposes. Journal of Acoustical Society of America. 47: 597-612.

BOLT RH, COOPER FS, DAVID EE, DENES PB, PICKETT JM & STEVENS KN (1973) Speaker Identification by Speech Spectrograms: some further Observations. Journal of Acoustical Society of America. 54 : 531 - 534.

versés, car une opinion subjective peut différer de manière significative des probabilités statistiques, aussi bien qu'entre différentes personnes.

L'élément de preuve fourni par une approche objective, comme la reconnaissance automatique de locuteurs, exprime la probabilité statistique d'une correspondance des caractéristiques dépendantes du locuteur extraites de l'indice et de l'enregistrement de comparaison. L'analyse phonétique-acoustique se trouve à mi-chemin entre une approche subjective et objective.

Processus de décision

La signification de l'élément de preuve, déterminé de manière objective ou subjective, doit être évaluée dans un canevas d'interprétation cohérent, du point de vue de l'inférence de l'identité de la source en sciences forensiques. L'approche par évaluation de rapports de vraisemblance, proposée comme canevas d'interprétation générique en sciences forensiques, notamment par KWAN et AITKEN, et plus spécifiquement proposé pour la reconnaissance de locuteurs en sciences forensiques par LEWIS, remplit ce critère de cohérence [KWAN, 1977 ; LEWIS, 1984 ; AITKEN, 1995].

La vraisemblance de l'élément de preuve est calculée d'une part dans l'hypothèse où la personne suspectée est effectivement la source de l'indice (H1) et d'autre part dans une hypothèse alternative, où la personne suspectée n'est pas la source de l'indice (H2). Le résultat de cette interprétation est exprimé sous la forme d'un rapport de vraisemblance, *likelihood ratio* (LR), des hypothèses H1 et H2.

La plupart du temps cependant, l'élément de preuve est interprété dans un canevas incohérent, du point de vue de l'inférence de l'identité de la source en sciences forensiques, notamment par l'intermédiaire de processus de décision binaires de classification ou de discrimination [CHAMPOD ET MEUWLY, 1998]. Dans ce cas, le résultat de l'interprétation est généralement exprimé sous la forme de décisions d'identification ou de vérification, assorties de taux d'erreurs de type I (faux négatif) et / ou de type II (faux positif) [KOVAL ET AL., 1998 / I].

Reconnaissance de locuteurs par audition

La reconnaissance de personnes familières et non familières sont probablement deux aptitudes humaines indépendantes et non associées. Identifier un locuteur familier consiste essentiellement en un processus de reconnaissance de forme dans lequel une voix unique est mise en correspondance avec un individu. Pour la discrimination entre locuteurs non familiers, le processus de reconnaissance s'appuie non seulement un processus de reconnaissance de forme, mais aussi une analyse des caractéristiques de la voix. Les performances de reconnaissance des êtres humains sont nettement meilleures lorsque les voix sont familières ; cependant, dans la plupart des cas réels, le locuteur inconnu n'est pas une personne familière de l'auditeur, qu'il soit expert ou non.

Reconnaissance par des non-experts

Les expériences conduites en présentant un groupe de locuteurs à des auditeurs non-experts montre que l'acuité de la discrimination pour les voix non familières dépend de plusieurs facteurs [CLIFFORD, 1980]:

- la taille et l'homogénéité auditive des voix présentes dans le groupe;
- l'âge et le sexe des locuteurs et des auditeurs;
- la quantité de parole initialement entendue;
- le délai entre l'écoute initiale et la procédure de reconnaissance;
- l'existence d'un déguisement de la voix;
- la qualité conjointe du canal de transmission et du système d'enregistrement;
- la présence ou l'absence d'un témoignage visuel concordant.

Même si tous ces facteurs sont favorables, la grande variabilité des performances individuelles restreignent l'usage des réponses de ce type d'expériences à une valeur indicative, dont l'incertitude est comparable à celle du témoignage.

BOLT RH, COOPER FS, GREEN DM, HAMLET SL, MCKNIGHT JG, PICKETT JM, TOSI OI & UNDERWOOD BD (1979) On The Theory And Practice of Voice Identification. National Academy of Sciences. Washington DC.

BUNGE E, (1977) Speaker recognition by computer. Philips Technical Review 37-8 : 207 - 219.

CHAMPOD C & MEUWLY D (1998) The Inference of Identity in Forensic Speaker Recognition. Proceedings of IEEE-ESCA Workshop on Speaker Recognition and its Commercial and Forensic Applications: Avignon. 125- 135.

CLIFFORD BR (1980) Voice Identification by Human Listeners: on Earwitness Reliability. Law and Human Behaviour. 4 : 373 - 394.

CUTLER PE, THIPGEN CR, YOUNG TR & MUELLER EB (1972) The Evidentiary Value of Spectrographic Voice Identification. The Journal of Criminal Law, Criminology and Police Science. 63: 343- 355.

DOHERTY ET (1976) An evaluation of selected acoustic parameters for use in speaker identification. J. phonetics 4 : 321 - 326.

EL MALIKI M. (2000) Speaker verification with missing features in noisy environements. Thèse de doctorat, Ecole Polytechnique Fédérale de Lausanne, Suisse.

FURUI S (1997) Recent Advances in Speaker Recognition. First International Conference on Audio- and Video- Based Person Authentification. Montana (Switzerland). 237 - 252.

KERSTA LG (1962) Voiceprint identification. Nature. 4861: 1253-1257.

KOVAL S, KAGANOV A & KITHROY M (1998/I) The Chart of the Standard Expert Actions and Decision Making Principles of Forensic Speaker Identification. Proceedings of the 8th COST 250 Workshop on Speaker Identification by Man and by Machine: Directions for Forensic Applications: Ankara. 62- 66.

Reconnaissance par des experts

Cette approche combine l'analyse auditive-perceptive et phonétique-acoustique. La plupart des caractéristiques dépendantes du locuteur considérées sont basées sur des mécanismes physiologiques de production de la parole ou sur des connaissances psycho-acoustiques [KOVAL ET AL., 1998 / II].

L'approche auditive-perceptive

L'approche auditive-perceptive consiste en une analyse auditive détaillée. Elle se focalise sur des paramètres (1) de la voix, comme la hauteur, le timbre, l'ampleur ; (2) de la parole comme l'articulation, la diction, la vitesse de parole, les pauses, l'intonation, les défauts de parole et (3) du langage comme la dynamique, la prosodie et le style.

Elle inclut aussi des observations linguistiques, notamment syntaxiques, idiomatiques et même des caractéristiques para-linguistiques comme le rythme respiratoire. Les principaux résultats de ce type d'analyse sont documentés dans une transcription effectuée à l'aide de l'alphabet phonétique international.

L'approche phonétique-acoustique

Grâce aux techniques de visualisation informatiques et aux algorithmes spécifiques au traitement numérique des signaux, l'analyse phonétique-acoustique permet une quantification ou une description plus précise de nombre des paramètres étudiés dans l'approche auditive perceptuelle. Dans le domaine fréquentiel, la trajectoire ainsi que l'amplitude et la largeur de bande des formants sont étudiées (les formants, ou fréquences formantiques sont, dans les voyelles, des plages de fréquence où les résonances du conduit vocal sont concentrées).

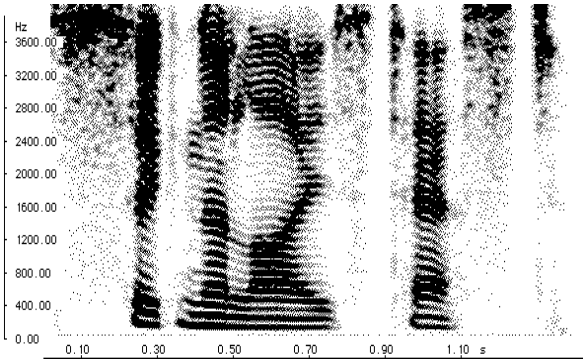
La distribution spectrale de l'énergie, la fréquence fondamentale moyenne et l'évolution temporelle de la fréquence fondamentale sont aussi étudiées. Dans le domaine temporel, l'analyse se focalise sur la durées des segments, le rythme et la variation cycle-à-cycle de la période de vibration des cordes vocales, appelé *jitter*.

Reconnaissance de locuteurs par comparaison visuelle de spectrogrammes

Technologie

Le spectrographe sonore est un instrument qui montre, sous forme graphique (Fig. 2), la variation du spectre à court terme de l'onde de parole. Dans chaque spectrogramme, le temps se déroule sur la dimension horizontale, les fréquences sur la dimension verticale, alors que la couleur ou l'intensité du trait représente l'intensité du signal sur une échelle compressée. Cette technique de reconnaissance de locuteurs, largement utilisée aux Etats-Unis est basée sur la supposition que la variabilité intralocuteur est moindre ou différente de la variabilité interlocuteur et que les examinateurs « d'empreintes vocales » peuvent détecter de manière fiable cette différence par comparaison visuelle de spectrogrammes.

Fig. 2: Spectrogramme à bande étroite (25ms - 40Hz) de l'énoncé «Signalyze test»



La méthode de KERSTA

En 1962, KERSTA propose en premier une méthode de reconnaissance par comparaison visuelle de spectrogrammes sous le nom d'identification par « empreintes vocales ». Il déclare que les spectrogrammes vocaux d'une personne sont aussi permanents et uniques que les empreintes digitales et permettent d'atteindre le même degré de certitude pour l'identification forensique [KERSTA, 1962].

En 1970, BOLT et ses collaborateurs réfutent ces assertions observant que, contrairement aux empreintes digitales, les ressemblances de forme des spectrogrammes vocaux dépendent premièrement de séries de mouvements acquis pour la production du langage, et seulement partiellement et indirectement de la structure anatomique du tractus vocal.

Les détails de forme des spectrogrammes sont tout aussi variables que leur forme générale, qui est affectée par la croissance, les habitudes et l'état de santé. Par conséquent, les différences avec les empreintes digitales sont plus significatives que leurs ressemblances, et la complexité inhérente au langage parlé rapproche plus la reconnaissance de locuteurs par comparaison visuelle de spectrogrammes de la discrimination auditive de voix non familières que de la discrimination dactyloscopique. BOLT et ses collaborateurs concluent que le terme « empreinte vocale » est une analogie fallacieuse aux empreintes digitales et que des recherches considérables demeurent nécessaires pour établir la validité et la fiabilité de cette méthode [BOLT ET AL., 1970].

L'étude de TOSI : une tentative de validation de la méthode de KERSTA

En 1972, TOSI et ses collaborateurs publient la seule étude à large échelle dans ce domaine, focalisée sur l'habilité des êtres humains à procéder à la reconnaissance de locuteurs sur la base de comparaison visuelle de spectrogrammes. Ils concluent que « si des examinateurs 'd'empreintes vocales' entraînés utilisent l'approche auditive et l'approche visuelle pour l'identification de locuteurs, leurs performances restent supérieures aux performances réalisées par des examinateurs non entraînés analysant des enregistrements réalisés dans des conditions de laboratoire, même lorsque les premiers analysent des enregistrements réalisés dans des conditions forensiques » [TOSI ET AL., 1972].

Cette conclusion, basée sur une extrapolation et non sur une démonstration, et l'assertion selon laquelle cette méthode est acceptée de la communauté scientifique, sont invalidées par BOLT et ses collaborateurs en 1973. Ces derniers déclarent que l'étude de TOSI a amélioré la compréhension de certains des problèmes liés à la reconnaissance de locuteurs par comparaison visuelle de spectrogrammes en indiquant l'influence de plusieurs variables importantes pour l'efficacité de l'identification. Pourtant, BOLT et ses collaborateurs concluent ensuite que dans nombre de situations pratiques la méthode manque d'une base scientifique adéquate pour aboutir à une estimation de sa validité et que les évaluations de laboratoire montrent une augmentation de l'erreur lorsque les conditions s'approchent des conditions forensiques réelles [CUTLER ET AL., 1972 ; BOLT ET AL., 1973].

Le rapport de l'Académie Nationale de la Science (*NAS*)

La vaste controverse suscitée autour de la validité de la reconnaissance de locuteurs par comparaison visuelle de spectrogrammes et de la recevabilité des expertises qui en résultent, incite l'*Académie Nationale de la Science des Etats-Unis* à conduire une étude en 1976, à la demande du *Federal Bureau of Investigations* (FBI). Le comité de l'Académie conclue en 1979 que les incertitudes techniques concernant cette méthode sont si importantes qu'une application forensique ne devrait être envisagée qu'avec la plus grande prudence. Il ne prend aucune position pour ou contre son utilisation forensique, mais recommande que, lorsque la méthode est présentée en tribunal, ses limitations soient clairement et explicitement mentionnées au juge ou au jury [BOLT ET AL., 1979].

Depuis la publication de ce rapport, une activité légale mixte existe aux Etats-Unis, certains cas favorisant la recevabilité de cette méthode, d'autres la rejetant.

Toutefois, en 1992 une série de règles concernant le témoignage scientifique est énoncée dans l'arrêt *Daubert v Merrel Dow Pharmaceuticals*[1] de la Cour suprême des Etats-Unis, qui précise notamment des exi-

[1] [Daubert v Merrel Dow Pharmaceuticals (1993, US) 125 L Ed 2d 469, 113 S Ct 2786]

KOVAL S, LLYINA O & KHITINA M. (1998/II) Practice of Usage of Auditive and Linguistic Features for Forensic Speaker Identification. Proceedings of the 8th COST 250 Workshop on Speaker Identification by Man and by Machine: Directions for Forensic Applications: Ankara. 23 – 29

KÜNZEL HJ (1994) Current Approaches to Forensic Speaker Recognition. Proceedings of ESCA Workshop on Automatic Speaker Recognition Identification and Verification: Martigny. 135-141

KWAN QY (1977) Inference of Identity of source. Ph. D. Thesis, University of California, Berkeley, CA, USA.

LEWIS SR (1984) Philosophy of Speaker Identification. Police Applications of Speech and Tape Recording Analysis: Proceedings of Acoustics. 6 :1 69 - 77.

gences concernant la méthodologie. Dans les faits, l'arrêt Daubert ne lie que les juridictions fédérales et son impact à long terme sur les régimes de recevabilité de la preuve scientifique aux Etats-Unis est inconnu [BLACK ET AL., 1994].

Reconnaissance automatique de locuteurs

Les méthodes de reconnaissance automatique de locuteurs peuvent être divisées en méthodes dépendantes et indépendantes du texte. Les méthodes indépendantes du texte sont prédominantes en sciences forensiques, ou des mots-clefs prédéterminés ne peuvent pas être utilisés.

Premières tentatives

Durant les années 1970, la recherche était orientée vers la découverte des facteurs déterminant la reconnaissance du locuteur dans le signal de parole et vers la sélection statistiques des paramètres les plus pertinents. Plusieurs systèmes semi-automatiques basés sur l'extraction d'événements acoustiques phonétiques explicites et sur des mesures statistiques simples furent développés pour un usage forensique. Les systèmes SASIS, *Semi-Automatic Speaker Identification System*, développé par *Rockwell International* aux Etats-Unis et AUROS, *Automatic Recognition Of Speaker by computer*, développé en collaboration par *Philips* et par le BKA, *BundesKriminalAmt*, en Allemagne ont été abandonnés à cause de leur manque de résultats dans le contexte forensique et de leur difficulté d'utilisation ; ils n'étaient utilisables que par des phonéticiens [BECKER ET AL., 1973 ; BUNGE, 1977]. Le système SAUSI, *Semi-AUtomatic Speaker Identification system*, est encore en développement à l'Université de Floride, mais son application dans des conditions forensiques réelles reste à établir [DOHERTY, 1976 ; HOLLIEN ET JIANG, 1998].

Dès le début des années 1980, les méthodes basées sur la localisation explicite d'événements acoustiques dans le signal de parole devinrent obsolètes et un consensus émergea, centré sur l'usage de paramètres dérivés du spectre à court terme du signal de parole. Le système CAVIS, *Computer Assisted Voice Identification System*, a été développé au *Los Angeles County Sheriff's Department* entre 1985 et 1989, avec la collaboration des Services Secrets des Etats-Unis. Il était basé sur l'extraction de caractéristiques temporelles et fréquentielles et sur une procédure de comparaison et pondération statistique. CAVIS a cependant été abandonné parce que les résultats de laboratoires encourageants, ne se répercutaient pas dans des conditions forensiques réelles [NAKASONE ET MELVIN, 1989].

L'état de l'art

A présent, la recherche se concentre sur des méthodes capables d'exploiter efficacement des groupes de paramètres implicites et abstraits extraits du signal de parole. Par conséquent, les progrès proviennent plutôt d'une amélioration dans les techniques de mesure et de modélisation de la distribution des paramètres que d'une amélioration des éléments dépendants du locuteur dans le signal de parole.

Plusieurs approches ont été utilisées pour modéliser le locuteur, principalement la quantification vectorielle (VQ), les modèles de Markov cachés ergodiques (HMM), les réseaux de neurones artificiels (ANN) et la modélisation par mélange de gaussiennes (GMM) [FURUI, 1997]. La plus récente évaluation des méthodes de reconnaissance de locuteurs indépendantes du texte effectuée par le *National Institute of Standards* (NIST) des Etats-Unis montre que les performances de la modélisation par mélange de gaussiennes (GMM) surpassent les résultats obtenus par les autres méthodes [PRZYBOCKI ET MARTIN, 1998].

Les limitations de la technologie

Les facteurs qui affectent le plus la reconnaissance automatique de locuteurs sont les variations d'un enregistrement à l'autre dues au locuteur lui-même, celles dues aux conditions de transmission et d'enregistrement et celles dues au bruit de fond. L'enregistrement de comparaison, la durée de l'indice, le sexe du locuteur, la variation entre l'enregistrement

NAKASONE H & MELVIN C (1989) C.A.V.I.S.: (Computer Assisted Voice Identification System) final report. National Institute of Justice, Grant no. 85-IJ-CX-0024.

NOLAN F (1990) The limitations of auditory-phonetic speaker identification, in: Texte zu Theorie und Praxis forensischer Linguistik (ed: Kniffka, H.) M. Niemeyer, Tübingen. 457 - 479.

PRZYBOCKI M & MARTIN AF (1998) NIST speaker recognition evaluation - 1997. Proceedings of RLA2C Workshop: Speaker Recognition and its Commercial and Forensic Applications. 120 - 124.

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

THEVENAZ P (1993) Résidu de prédiction linéaire et reconnaissance de locuteurs indépendante du texte. Thèse de doctorat, Université de Neuchâtel, Suisse.

de comparaison et l'indice affectent particuliè-rement les performances, tout comme les autres dégradations typiquement rencontrées en sciences forensiques. Il est reconnu que l'égalisation spectrale est efficace pour réduire les effets linéaires du canal de transmission et la variation spectrale à long terme, mais la normalisation peut être améliorée en utilisant des techniques plus récentes, comme la soustraction spectrale combinée à la compensation statistique des caractéristiques manquantes [EL MALIKI, 2000].

TOSI O, OYER H, LASHBROOK W, PEDREY C, NICHOL J & NASH EW (1972) Experiment on Voice Identification Journal of Acoustical Society of America. 51 : 2030 - 2043.

Rédacteur de ce numéro:
Didier Meuwly

Annonce

Le thème de ce *Crimiscope* a été étudié dans le cadre d'une thèse de doctorat qui sera présentée le lundi 22 mai 2000, à 17 heures, à l'auditoire C du Bâtiment de pharmacie de l'Université de Lausanne.

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