

## Dossier

# Ben Laden et le mythe de l'empreinte vocale



Contrairement aux idées reçues, il est impossible d'identifier avec certitude une voix enregistrée au téléphone.

**Notre société, dans sa quête de solutions à ses problèmes, suscite une multitude d'expertises. Or celles-ci, confrontées aux limites des connaissances scientifiques du moment, peuvent être dans l'impossibilité de donner des réponses claires. Encore faut-il bien vouloir le reconnaître... Ainsi, les spécialistes de la parole affirment de longue date que les expertises vocales sont infondées scientifiquement : il est impossible d'assimiler une personne à une voix enregistrée sans marge d'erreur. Nonobstant, la police et la justice persistent à recourir à des experts en acoustique pour confirmer l'identité de suspects et prendre des décisions parfois lourdes de conséquences.**

Régulièrement nous revenons sur les ondes des informations affirmant que le terroriste Oussama Ben Laden aurait livré, depuis sa cachette, un nouveau message audio menaçant l'Occident de ses foudres ; sa voix aurait été formellement reconnue ou serait en passe de l'être par les experts de la CIA. De même, l'analyse et la reconnaissance d'une voix enregistrée au téléphone détermine parfois l'issue d'affaires pénales. Un prévenu peut être ainsi placé en détention, voire condamné, sur la base de l'identification de sa voix par des experts (*voir l'encadré 1 ci-dessous*). Et des sites Internet (comme [crimelibrary.com](http://crimelibrary.com)) légitiment cette pratique en racontant à l'envi comment l'analyse par spectrogramme numérique de l'« empreinte vocale » de malfaiteurs a permis leur arrestation.

### Des expertises sans validité scientifique

Or, contrairement aux idées reçues, la voix n'offre pas des caractéristiques individuelles aussi stables et fiables que celles des empreintes digitales utilisées par la police depuis un siècle, et aussi difficilement réfutables que celles des empreintes génétiques, une technique découverte en 1985. La voix n'est pas une image du corps comme le sont les photos des crêtes papillaires, et encore moins une

#### Encadré 1. La voix comme élément de preuve



Pour établir l'identité des suspects, les enquêteurs peuvent utiliser comme éléments de preuve directs ou indirects des empreintes

représentation d'une partie du corps comme le sont les empreintes génétiques.

En clair, dans l'état actuel des connaissances, il n'existe pas de procédures permettant d'avancer avec certitude qu'une personne est – ou n'est pas – l'auteur d'un appel téléphonique ou d'un enregistrement audio. Les rapports d'expertise d'enregistrements vocaux n'ont donc aucune validité scientifique. C'est pourquoi les spécialistes de la parole réunis au sein du Groupe français de la Communication parlée (GFCP) de la Société française d'Acoustique (SFA), puis de l'Association francophone de la Communication parlée (AFCP), ont demandé à plusieurs reprises que les expertises vocales ne soient plus utilisées par la Justice tant qu'elles n'auront pas été validées scientifiquement.

Voyons plus en détails sur quels arguments se fonde cette position.

### Les aléas de la reconnaissance des voix

C'est une banalité de dire que l'homme est capable de reconnaître des voix familières, grâce à ses capacités cognitives. Ainsi, certaines lésions cérébrales attestées dans l'hémisphère droit pariétal entraînent la perte de cette compétence : c'est la phonagnosie. La tentation est donc grande de conclure que chaque voix possède des caractéristiques qui la rendent facilement reconnaissable et unique parmi toutes les voix possibles.

Ce serait une conclusion erronée ! En effet, la reconnaissance auditive des voix familières est loin d'atteindre des taux proches de 100 %. Expérimentalement, elle est de l'ordre de 60 % sur une phrase entière et de 80 % sur 30 secondes de parole. Dans des tâches auditives de discrimination consistant à indiquer si deux brefs échantillons de parole (5 secondes) proviennent ou non d'un même locuteur, les performances sont faibles, de 38 % à 76 % de bonnes réponses, soit 50 % en moyenne, pas mieux qu'un tirage au hasard.

Les machines doivent pouvoir résoudre ce problème, pensez-vous.

À tort, car les logiciels de reconnaissance de la voix n'atteignent pas non plus des scores très élevés. En 1998, dans le cadre d'évaluations de systèmes de reconnaissance du locuteur, conduites aux Etats-Unis par l'Institut national des normes et de la technologie, le NIST (National Institute of Standards and Technology), les évaluations des systèmes les plus performants ont été comparées à des tests perceptifs menés avec des auditeurs. Les systèmes techniques parvenaient à des taux d'erreur de l'ordre de 15 %, certes moins importants que ceux des humains (de l'ordre de 25 %), mais qui ne permettaient pas d'envisager une expertise juridique fiable (*voir l'encadré 2, deuxième partie*).

Ainsi, contrairement aux méthodes d'identification par « portrait robot », qui permettent de raviver le souvenir d'un témoin et d'en améliorer la précision, nous sommes encore loin de disposer d'une procédure de synthèse qui permettrait, à l'aide d'un jeu de caractéristiques sonores, de reconstituer la voix entendue et évoquée par un témoin. Des tests ont montré par ailleurs que la description de la voix n'est définitivement plus fiable si elle est effectuée par un témoin auditif plus de 24 heures après son écoute.

### Sens commun et empreinte vocale

Bien sûr, toutes ces conclusions vont à l'encontre du sens commun. Nous avons tous l'impression de reconnaître avec facilité les voix familières, y compris lorsqu'elles sont déformées par une liaison téléphonique. En fait, dans une conversation téléphonique, l'identification de l'interlocuteur fait appel à de nombreuses informations contextuelles et inconscientes : nombre limité des correspondants potentiels, heure et nature de l'appel, préoccupations du moment, relations circonstancielles, adaptation de la voix à l'interlocuteur, etc. Si bien que souvent, en décrochant le téléphone, on devine l'identité de la personne qui nous appelle.

digitales, des traces de sang, de sperme, des cheveux, des empreintes de pas, etc., mais aussi des enregistrements réalisés avec l'autorisation du procureur de la République par interception de communications téléphoniques.

Le juge d'instruction a le pouvoir d'ordonner une écoute téléphonique, à condition que la peine encourue soit supérieure ou égale à deux ans d'emprisonnement. La durée des écoutes ne peut excéder 4 mois (sauf renouvellement) et un procès-verbal doit être établi pour chaque enregistrement. Peuvent être surveillées, la personne qui a été mise en examen, mais aussi des tiers, la partie civile et même les avocats. De leur côté, la police et la gendarmerie peuvent procéder dans leurs laboratoires – Laboratoire d'Analyse et Traitement de Signal (LATS) de la police scientifique d'Écully, département Signal Image Parole de l'Institut de Recherche Criminelle de la Gendarmerie Nationale (IRCGN) de Rosny-sous-Bois – à une comparaison de la voix anonyme et de celle d'un suspect. Entre 1995 et 2000, le LATS a traité 154 affaires d'identification vocale, doublant presque chaque année le nombre de ces expertises très contestées.

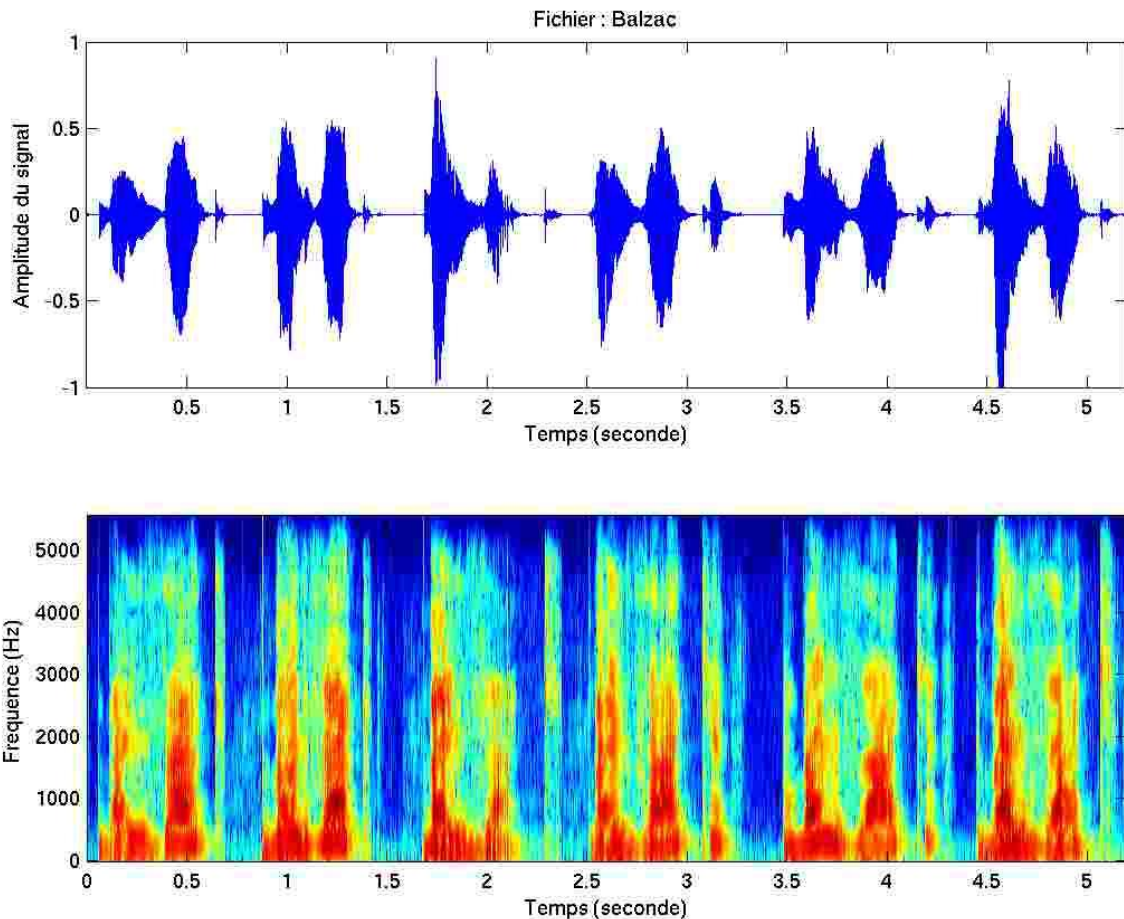
Le rapport d'analyse peut être versé au dossier de l'enquête préliminaire. De même, le juge d'instruction peut demander à être éclairé sur certains points par un spécialiste. Comme l'analyse de la parole n'existe pas en tant que spécialité auprès des tribunaux, les juges font appel à des experts qualifiés en « acoustique et vibrations » ou à des personnes censées posséder les compétences nécessaires.

Cette fausse impression s'appuie aussi sur un *a priori* intuitif qui ne date pas d'aujourd'hui : l'existence d'une « empreinte vocale » spécifique de chaque individu et facilement reconnaissable. On se souvient que Charles Lindbergh, en 1936, avait affirmé reconnaître la voix de l'assassin de son fils kidnappé, plus de deux ans après l'avoir entendue brièvement. L'idée qu'il pouvait se tromper n'effleura pas l'esprit des jurés, et ils condamnèrent à mort le prévenu. Avec les progrès techniques, ce préjugé gagna encore en évidence. Ainsi en 1962, Lawrence Kersta, ingénieur des laboratoires Bell, affirma, dans un article de la revue *Nature* intitulé *Voiceprint identification*, détenir les moyens d'identification de l'empreinte vocale (*Voiceprint*) de chaque individu. Il donnait à croire que l'analyse spectrale de la voix, sous forme de « sonagramme » (voir la figure ci-dessous), était tout aussi fiable que les empreintes digitales.

### Un taux élevé mais indéterminé d'erreur

Or en 1970, à la suite d'expertises judiciaires controversées qui avaient utilisé la méthode de Kersta, le comité technique sur la communication parlée de la Société américaine d'acoustique concluait que l'identification des locuteurs par la voix était sujette à un taux élevé mais indéterminé d'erreur. En effet, précisait le comité, les éléments spectrographiques présentent souvent des caractéristiques plus largement influencées par les mots qu sont prononcés que par l'identité du locuteur. En 1980, les spécialistes de phonétique du Royaume-Uni prenaient aussi position en ce sens en demandant aux candidats à l'expertise judiciaire de faire la preuve préalable de leurs compétences. À partir de 1990, les scientifiques français adoptaient à leur tour une telle démarche.

## Sonagrammes et identification des suspects



Le sonagramme est une représentation de la parole (en haut le signal de parole et, au dessous, le sonagramme proprement dit) : en abscisse le temps, et en ordonnée les fréquences composantes du signal, ici d'autant plus rouges que leur amplitude est importante. Au cours des années 1960, cette représentation – abusivement comparée à une empreinte digitale – a été considérée comme représentative des caractéristiques de la voix du locuteur. À la suite d'expertises judiciaires controversées, un rapport de l'Acoustical Society of America a conclu dès 1970 que l'identification fondée sur ce type de représentation entraînait un taux d'erreur important et

difficilement prévisible.

En fait, les ressemblances qui peuvent exister pour les sonagrammes correspondant à un mot donné, prononcé par deux locuteurs, peuvent être dues au fait qu'il s'agit justement du même mot. Et, inversement, un même mot prononcé dans des phrases différentes par un même locuteur peut présenter des tracés différents. La figure ci-dessus correspond à un même mot : « Balzac », prononcé six fois par une même locutrice et extrait de phrases différentes, enregistrées lors d'une conférence. Sur le sonagramme, les répétitions présentent de très nettes différences. Actuellement, en France, la police scientifique continue malgré tout à identifier les suspects en se basant, pour l'essentiel, sur ce type de comparaison, qui a pourtant fait la preuve de son inefficacité.

Cette position a été renforcée par l'évaluation scientifique de la fiabilité des empreintes digitales et génétiques, qui repose notamment sur l'existence de bases de données expérimentales très importantes : en France par exemple la police dispose de bases contenant des centaines de milliers d'empreintes digitales et, désormais, de milliers d'empreintes génétiques. Or, dans le domaine vocal, les bases de données disponibles actuellement ne comportent pas un nombre suffisant de locuteurs, de langues, de conditions d'enregistrement pour évaluer la fiabilité des méthodes existantes de reconnaissance.(...)

La raison de notre inaptitude à attribuer un profil vocal à une voix donnée est la suivante : un enregistrement de parole n'est que la capture indirecte de mouvements articulatoires complexes faisant intervenir les cordes vocales, la langue, le voile du palais, la mâchoire et les lèvres. Les mouvements des organes de la parole engendrent des variations de pression acoustique instantanée qui peuvent être captées par un transducteur et transformées en variations de tension électrique.

Or, comme tous les gestes de l'homme, les gestes de parole sont difficilement reproductibles à l'identique au cours du temps, sauf entraînement systématique. En effet, la vitesse d'articulation, l'intensité et la hauteur de votre voix varient beaucoup selon les conditions de communication (conversation familière, lecture, communication téléphonique, etc.), selon notre état psychologique et émotionnel, notre fatigue ou stress et, bien entendu, selon que nos cordes vocales et notre gorge se portent bien ou mal. La reconnaissance automatique de la parole, qui reste peu fiable encore aujourd'hui, est d'ailleurs directement confrontée à cette variabilité individuelle.

### Une gamme de techniques de déformation de la voix

À ces contraintes, liées au processus de production organique de la parole, s'ajoutent d'autres facteurs augmentant la difficulté de reconnaissance des voix : paramètres de transmission et d'enregistrement, qui dépendent eux-mêmes des appareils utilisés ; éventualité d'une superposition de plusieurs voix ou de bruits ; possibilité d'imitation et de déguisement de la voix, usage possible de toute une gamme de techniques de déformation allant du simple égaliseur de spectre (filtrage), au vocodeur et aux techniques de transformation de la voix par « morphing » (ou morphage), technique informatique qui permet de transformer progressivement un son (ou une image) en un(e) autre.

## Encadré 2. Des évaluations comme éléments de comparaison

Des protocoles d'évaluation de la reconnaissance du locuteur permettent de quantifier les performances des systèmes ayant pour but d'indiquer si une séquence de parole a bien été produite par un « locuteur cible » (*target speaker*) déjà enregistré et répertorié dans une phase d'apprentissage, ou s'il s'agit d'un imposteur. Ces évaluations sont menées chaque année par l'Institut national des normes et de la technologie des États-Unis (NIST), en collaboration avec d'autres laboratoires. Ainsi en 1998, elles ont impliqué 400 locuteurs cibles, 250 imposteurs, trois conditions d'apprentissage (1 à 2 minutes de signal) et de test (3, 10 et 30 secondes), avec ou non le même combiné téléphonique. Les enregistrements de parole sont recueillis de manière à représenter un échantillonnage statistique représentatif des conditions d'évaluation. Les principaux paramètres du protocole renvoient à deux types de conditions.

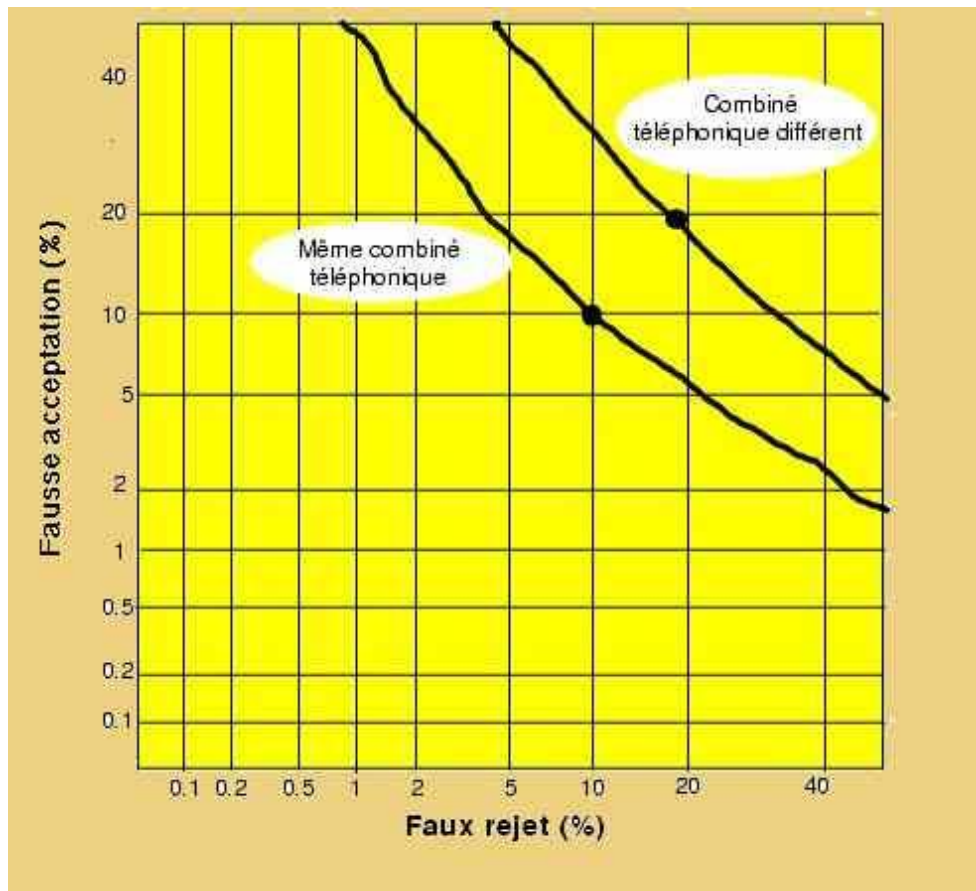
#### I Les paramètres modélisables :

- i le nombre de locuteurs cibles et le nombre d'imposteurs.
- i la dépendance au texte : soit le système sait ce que le locuteur est supposé prononcer, c'est-à-dire une séquence convenue (mot de passe) ou une séquence qu'il lui est demandé de prononcer, et il a été entraîné sur tous ces items possibles ; soit il n'a pas de connaissance a priori sur la séquence prononcée par le locuteur ; évidemment, dans ce dernier cas, les performances obtenues sont nettement moins élevées .
- i les conditions d'apprentissage : en particulier l'intervalle de temps qui sépare l'entraînement de la vérification, et le nombre de sessions d'enregistrement .
- i certaines conditions d'enregistrement et de transmission du signal : le rapport signal sur bruit, le type de bruit ambiant, les caractéristiques du microphone ; la nature de la transmission : analogique ou numérique.

#### I Les paramètres non maîtrisés durant l'expérience :

- i l'état émotionnel du locuteur au moment des sessions d'apprentissage et de test ; sa capacité à maintenir une certaine constance dans les caractéristiques de sa voix (variabilité intrinsèque) .

- i la coopération des locuteurs : les locuteurs sont toujours coopératifs ou tout au moins ne perturbent pas le système ; il n'est d'ailleurs pas possible de modéliser, hors contexte, des locuteurs non coopératifs .
- i les capacités d'imposture des locuteurs : dans la quasi-totalité des expériences, les locuteurs ne connaissent pas les autres locuteurs, et ne peuvent donc pas les imiter ; leur capacité d'imitation (instinctive ou apprise) ne peut être ni simulée, ni quantifiée.



Graphique DET (*Detection Error Trade-off*) montrant les performances d'un système utilisant des combinés téléphoniques identiques ou différents entre l'apprentissage et la reconnaissance, pour 30 secondes de signal test, en mode *text independent*.  
D'après M. Przybocki & A. Martin, NIST

Les systèmes sont évalués à partir de deux pourcentages d'erreur : le pourcentage de « fausses acceptations » (*false detection probability*) et le pourcentage de « faux rejets » (*miss probability*), calculés à partir du nombre de locuteurs cibles, du nombre de fois qu'ils n'ont pas été reconnus alors qu'ils auraient dû l'être, du nombre de tentatives d'« imposture » et du nombre de fois que des imposteurs ont été reconnus alors qu'ils n'auraient pas dû l'être.

Le taux de faux rejets correspond à la probabilité de ne pas détecter le locuteur cible : il représente le rapport entre le nombre de fois où un locuteur cible n'a pas été détecté et le nombre de tests l'impliquant. Le taux de fausses acceptations est la probabilité de détecter faussement un locuteur : soit le rapport entre le nombre de fois où le locuteur cible a été faussement détecté et le nombre de tests impliquant un « imposteur ».

Le graphique « DET (*Detection Error Trade-off*) » ci-dessus, tiré des évaluations internationales coordonnées en 1998 par le NIST, présente les résultats obtenus avec deux types de combinés téléphoniques différents. On constate que l'on peut diminuer le risque de fausse acceptation ; mais, dans ce cas, le nombre de faux rejets s'élève. Avec deux combinés téléphoniques différents, les erreurs sont de l'ordre de 20 %, une fiabilité loin d'être suffisante dans une affaire pénale.

Le message peut aussi avoir été enregistré par son auteur dans des conditions inaccessibles aux enquêteurs par exemple, pour détourner les soupçons, il peut avoir été fabriqué par le coupable à partir de la voix d'un autre locuteur. Certains téléphones digitaux permettent par exemple de modifier la hauteur et le timbre de la voix avec plusieurs dizaines de variantes. À moins de connaître les caractéristiques du traitement mis ainsi en œuvre, il est impossible d'identifier le locuteur, et même de déterminer son sexe !

Sur ce point, les machines ne font pas, actuellement, bien mieux que les humains. Encore faut-il préciser que cette conclusion a été établie, entre autres par le NIST aux États-Unis (*voir l'encadré ci-dessus*), avec des locuteurs qui ne cherchaient pas à déguiser leur voix et qu'aucun bruit n'était superposé aux enregistrements. Pour des conditions d'enregistrement très différentes tant techniquement (microphone, réseau téléphonique, caractéristiques des enregistreurs) que du point de vue de la situation de communication ou de la situation psychologique, avec des

voix déguisées et modifiées, avec du bruit superposé, et à des intervalles de temps de plusieurs mois, les performances seraient très nettement dégradées, au point d'être proches de celles d'un système opérant au hasard !

### Des probabilités trop floues pour conclure

Avec des enregistrements d'un même texte effectués dans les meilleures conditions techniques, est-il tout de même possible d'affirmer scientifiquement qu'ils proviennent ou non d'un même locuteur ? Même pas ! Bien sûr, il est possible de calculer des « distances » chiffrant quantitativement des ressemblances ou des différences pour un paramètre ou un jeu de paramètres donnés, d'évaluer les probabilités pour que deux voix proviennent d'un même locuteur dans un ensemble de personnes enregistrées au préalable. Mais il n'existe aucun consensus scientifique sur le choix des paramètres permettant de calculer ces distances, pas plus que sur les degrés de confiance des probabilités d'identification.

Les techniques de reconnaissance par la voix vont-elles progresser jusqu'à atteindre un haut niveau de fiabilité ? La voix est-elle trop facilement modifiable et trop variable pour que les analyses puissent atteindre ce niveau de performance ? Dans l'état actuel des connaissances et des recherches, il n'est pas possible de répondre à ces questions. C'est bien pourquoi la Justice devrait cesser de faire procéder à des soi-disant expertises vocales et à les utiliser comme preuve de culpabilité ou d'innocence.

**Louis-Jean Boë**

Institut de la Communication Parlée, INPG-Université Stendhal, CNRS, BP 25, 38040 Grenoble cedex 09

Pour contacter l'auteur  
boe@icp.inpg.fr

Pour aller plus loin

#### | Sites Internet

- | Association francophone de la Communication Parlée (Université d'Avignon),  
<http://www.afcp-parole.org>
- | National Institute of Standards and Technology (NIST),  
<http://www.nist.gov/speech/index.htm>
- | International Association of Forensic Linguists (IAFL),  
<http://www.iafl.org>
- | Biometric Consortium,  
<http://www.biometrics.org>
- | Liens sur la recherche en parole,  
<http://mambo.ucsc.edu/psl/speech.html>

#### | Ouvrages et articles

- | Boë L.J., Bimbot F., Bonastre J.F., Dupont P. (1999) « De l'évaluation des systèmes de vérification du locuteur à la mise en cause des expertises vocales en identification juridique », *Langues*, 2(4): 270-288.  
Version pdf, <http://www.afcp-parole.org/doc/Article-Langue.pdf>
- | Boë L.J. (2000) « Forensic voice identification in France », *Speech Communication*, 31 (2-3) : 205-224.  
Version pdf, <http://www.afcp-parole.org/doc/SpeComLJB.pdf>
- | Boë L.J., Bonastre J.F., Bimbot F. (2001) « Pourquoi la Justice doit arrêter les expertises vocales », *Justice* n°169, novembre 2001, Syndicat de la magistrature.  
Sommaire pdf, <http://www.syndicat-magistrature.org/Crew/Doc/74=somm169.pdf>
- | Bonastre J.F., Bimbot F., Boë L.J., Campbell J.P., Reynolds D.A., Magrin-Chagnolleau I. (2003) « Person Authentication by Voice: A Need for Caution ». 8th European Conference on Speech Communication and Technology, EUROSPEECH 1, 33-36, Genève, Suisse.
- | M. Epstein, A. Louvot, J.M. Pontaut, « Ben Laden Vivant ? »  
*L'Express* 21/11/2002,  
<http://www.lexpress.fr/express/info/monde/dossier/benladen/dossier.asp?ida=364516>
- | C. Cabal, *Les méthodes scientifiques d'identification des personnes à partir de données biométriques et les techniques de mise en œuvre*, Rapport de l'OPECST, juin 2003.  
En ligne sur le site de l'Assemblée nationale, <http://www.assemblee-nationale.fr/12/rap->

[oecst/i0938.asp](#)

| C. Champod, D. Meuwly (2000) « The Inference of identity in forensic speaker identification », *Speech Communication* 31:193-203.

| A.P.A. Broeders (2001) « Forensic Speech and Audio Analysis Forensic Linguistics. 1998 to 2001. A Review », 13e Symposium d'Interpol de science légale, Lyon, France, 16-19 octobre 2001.

Version

pdf,

<http://www.interpol.int/Public/Forensic/IFSS/meeting13/Reviews/ForensicLinguistics.pdf>

| L.M. Solan & P.M. Tiersma, « Falling on deaf ears », *Legal Affairs*, 2003.

A lire ici, [http://www.legalaffairs.org/issues/November-December-2003/story\\_solan\\_novdec03.html](http://www.legalaffairs.org/issues/November-December-2003/story_solan_novdec03.html)

#### | Revues

| *International Journal of Speech, Language and the Law*

<http://www.js-ijsll.bham.ac.uk/currentissue.asp>

| *Revue Language in the Judicial Process*

<http://www.outreach.utk.edu/ljp>

| *Speech Communication, Volume 31 (2-3) : Speaker Recognition and its Commercial and Forensic Applications*

<http://www.elsevier.nl/locate/specom>

© Vivant Editions – <http://www.vivantinfo.com>