La voix : un atout utile à l'identification d'individus ?

Patrick PERROT[12], Gérard CHOLLET[2]

[1]IRCGN, Institut de recherche criminelle de la gendarmerie nationale, 1 boulevard Théophile Sueur, 93110 Rosny sous Bois

[2]CNRS-LTCI-Institut Telecom -Telecom ParisTech 48, rue Barrault, Paris 75013
patrick.perrot@gendarmerie.defense.gouv.fr, chollet@tsi.enst.fr

Résumé

Messages de revendication, messages malveillants, les possibilités de transmettre des messages sous le couvert de l'anonymat sont aujourd'hui particulièrement fréquents. Les enregistrements de Ben Laden, les revendications terroristes de toutes origines, les appels à des fins de menaces ou de discrédit, voilà nombre de situations où la détermination de l'identité d'un individu à partir de sa voix peuvent s'avérer capital. Pourtant la voix est-elle unique, peut-on utiliser la voix comme une empreinte digitale ou génétique ? Les progrès en traitement du signal, la meilleure connaissance des mécanismes de la parole ont permis d'accroître significativement les performances des systèmes automatiques. Il existe encore des voix contraires à l'utilisation de la parole dans des applications autres que commerciales en raison d'un manque de fiabilité. L'objet de cet article est donc de présenter tout d'abord l'état de l'art en matière de reconnaissance automatique du locuteur, puis d'évaluer les performances de ces systèmes face aux techniques d'imposture possibles. La reconnaissance de locuteur constitue un terme générique qui comprend à la fois les notions d'identification et de vérification. L'identification consiste à vérifier l'identité d'un individu au sein d'un ensemble fermé en général, c'est-à-dire une comparaison de l'enregistrement de question avec les voix de N individus. La vérification consiste à reconnaître un individu au sein d'un ensemble ouvert, c'est-à-dire une comparaison de l'enregistrement de question à la voix d'un suspect uniquement. La reconnaissance automatique repose sur trois phases : apprentissage, test et décision. L'apprentissage propose la construction de modèles à partir de paramètres significatifs du locuteur afin de construire un modèle de la voix d'un individu ou un modèle de voix générique, représentatif de l'ensemble des voix. La phase de test a pour objet de calculer une distance de similarité entre les paramètres de la voix de question et les modèles préalablement établis. La décision repose sur la comparaison d'un rapport de vraisemblance par rapport un seuil établi précisément. Il existe aujourd'hui de nombreux systèmes automatiques dont les performances sont pourtant très hétérogènes. Nous présenterons donc les critères à prendre en compte dans le cadre d'une approche automatique de la reconnaissance de locuteur.

Face à cette problématique d'identifier ou de vérifier l'identité d'un individu, il existe des possibilités délibérées d'imposture. Celles-ci reposent sur des techniques de transformation ou de conversion de la voix. L'objectif de ces impostures est de ne pas être reconnu en modifiant les caractéristiques de sa voix par des techniques de déguisement, ou en essayant d'imiter une voix cible et ainsi se faire passer pour un autre individu. La détection des voix déguisées fondée sur la classification automatique, mais aussi le comportement des systèmes automatiques face à la problématique de l'imposture feront également l'objet d'un développement détaillé. Après une première partie qui définira la problématique de l'identification de la voix en matière de sécurité, nous dresserons un état de l'art du domaine reconnaissance de locuteur avant de présenter les travaux en matière de conversion de la voix et d'identification de déguisement.

Abstract

Messages claim, malicious messages, the potential to transmit messages under cover is today particularly common. Recordings of bin Laden, calls for threats, there are number of situations where the determination of an individual's identity from his voice can be crucial. Yet, can we consider voice as unique as fingerprints for instance? Advances in signal processing, and in the understanding of the speech mechanisms have significantly increased the performance of automated systems. There are still voices against the voice use in applications other than commercial due to a lack of reliability. The purpose of this article is to present first the state of the art in the field of automatic recognition of the speaker, and then to evaluate the performance of these systems respond to possible fraud techniques. Faced with this problem of identification or verification, there are many possibilities of deliberated fraud. These are based on techniques for processing or voice conversion. The aim of these forgeries is to false his own identity by modifying the characteristics of its voice through techniques of disguises, or trying to imitate a voice target and impersonate another individual voice. The detection of voice disguised based on the automatic classification, but also the behaviour of automated systems deal with the problem of fraud will also be detailed. After a first part, which sets a speaker recognition state of the art we will present the issue of voice security, before describing different works on the question of voice forgery.

1. Introduction

Nombre d'applications se développent aujourd'hui autour de la parole en reconnaissance automatique comme en synthèse. Les techniques de recherche dans l'identification ou la vérification d'un individu sont un sujet à l'ordre du jour. Les circonstances géopolitiques ont placé le terrorisme à l'ordre des préoccupations majeures. Or, la voix s'est avérée comme un outil, voire le seul, d'identification des locuteurs présents sur un enregistrement vidéo et bien sûr audio phonique. L'essor de la biométrie constitue elle aussi une des raisons de cet intérêt croissant pour la voix. En effet, parmi l'ensemble des modalités biométriques, la voix présente l'intérêt d'être l'une des plus faciles à exploiter, en terme d'acquisition, au sein d'une application biométrique, mais aussi l'une des plus faciles à transformer. Enfin, la forte demande d'expertise judiciaire en reconnaissance de locuteur exige un niveau de reconnaissance performant et si possible robuste à l'imposture. Ainsi, nous allons dans une première partie, nous intéresser aux enjeux de la reconnaissance du locuteur dans ses principes et mais aussi le niveau de performance des systèmes automatique. Dans une seconde partie nous nous intéresserons aux applications liées à la reconnaissance de locuteur dans le domaine judiciaire comme dans les technologies de sécurité avant d'en préciser les limites respectives. Enfin nous présenterons la problématique de l'imposture délibérée qui consiste à transformer sa voix ou à imiter la voix d'un individu cible.

2. Reconnaissance du locuteur

2.1 Principe de fonctionnement

En matière de reconnaissance automatique, nous pouvons distinguer deux grandes applications : l'identification et la vérification que nous regrouperons sous le vocable de reconnaissance. L'identification consiste à comparer un enregistrement anonyme (que nous qualifierons de question) à la voix de N individus. C'est une discrimination 1 contre N.

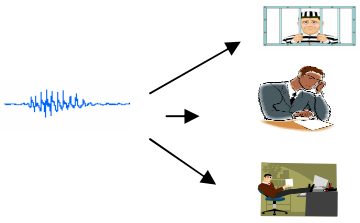

FIG n°1 : Principe de l'identification de locuteur (N=3)

Dans le cadre de la vérification nous comparons un enregistrement de question à la voix d'un individu, nous sommes donc dans une discrimination de type 1 contre 1.

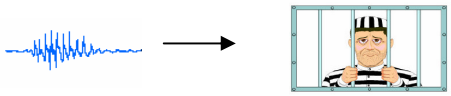

FIG n°2 : Principe de la vérification du locuteur

La reconnaissance automatique se classifie également à partir de la dépendance au texte. Le système de reconnaissance pourra être dépendant ou indépendant du texte prononcé. Le principe de la reconnaissance automatique de locuteur repose sur un test d'hypothèse statistique qui consiste à prendre une décision entre deux possibilités. A partir d'un enregistrement de question X, il convient de déterminer si :

H1 : l'enregistrement X a été prononcé par la personne I

H2 : l'enregistrement X a été prononcé par une autre personne.

Dès lors le test d'hypothèse statistique est le suivant :

$$p_0(X) = p(H_1/X) \text{ et } p_1(X) = p(H_2/X)$$

et dans le domaine logarithmique pour une facilité de calcul :

$$\log\ p_0(X) - \log\ p_1(X) \gtrless_{<} seuil$$

L'approche automatique va consister à modéliser ces tests d'hypothèse et l'évaluation du système permettra de déterminer le seuil au plus juste. La figure n° 3 représente un synoptique de système d'identification et la figure n°4 de vérification:

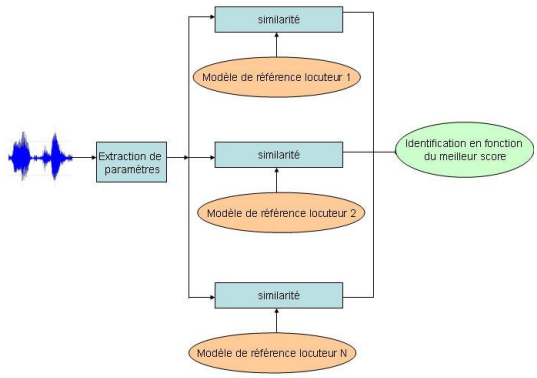

Figure n°3 : Identification de locuteur

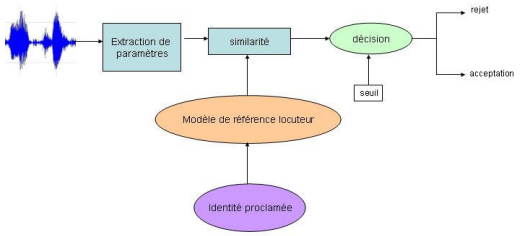

Figure n°4 : Vérification du locuteur

Il existe différents paramètres caractéristiques de la voix humaine. Nous pouvons penser aux paramètres les plus évidents comme la fréquence de vibration des cordes vocales ou encore la mélodie de la voix appelé prosodie et qui regroupent outre la fréquence fondamentale, l'énergie, et la durée syllabique ou phonémique. Ces paramètres sont appelés suprasegmentaux. Ils présentent néanmoins l'inconvénient majeur d'être peu robuste et peu discriminant du locuteur. D'autres paramètres sont donc utilisés : ceux caractéristiques du conduit vocal ou encore de l'enveloppe spectrale. Ce sont les cœfficients de prédiction linéaire LPCC (Linear Predictive Cepstral Coefficients) et leurs transformations ou encore les MFCC (Mel Frequency Cepstral Coefficients) et leurs transformations (dérivées et/ou accélération). Ces paramètres sont issus d'une analyse en bancs de filtre. En général 13 MFCC sont calculés auxquels s'ajoutent les 13 dérivées et les 13 dérivées des dérivées (accélération) pour parvenir à environ 39 coefficients.

A partir de cette phase d'extraction de paramètres nous distinguerons les deux grandes étapes des systèmes automatique : l'apprentissage et le test.

L'apprentissage consiste à modéliser les paramètres à la fois du (ou « des » en identification) locuteur proclamé, sensé modéliser l'hypothèse H1, et du modèle générique de l'ensemble des voix, sensé modéliser l'hypothèse H2. Nous appellerons ce modèle générique le modèle du monde.

Ces modèles sont sur les systèmes à l'état de l'art, constitués à partir de mélanges de gaussienne : les GMM (Gaussian Mixture Model) [9]. Le but est de modéliser la distribution des paramètres par une somme pondérée de composantes gaussiennes. Les paramètres extraits d'une base de données importante en nombre et diversifiée en voix d'individus permettront de modéliser le modèle du monde. Ce dernier sera adapté pour obtenir le modèle de voix d'un locuteur spécifique à partir des données d'apprentissage. Cette adaptation s'effectue en général au maximum de vraisemblance et par déclinaison en fonction du nombre de données disponibles par l'algorithme MAP (Maximum A Posteriori). La phase de test quant à elle consiste à calculer une distance de similarité entre les paramètres extraits de l'enregistrement anonyme et les modèles précédemment cités. Cette distance est une log vraisemblance d'appartenance d'un échantillon à un modèle.

2.1 Evaluation des performances

Le domaine de la reconnaissance automatique du locuteur fait l'objet d'évaluation annuelle depuis 1997 où nombre d'institutions (en général universitaires) présentent le résultat de leur système à partir d'un corpus de parole fournis par les organisateurs. La plus commune des évaluations est la campagne « Speaker Recognition Evaluation » organisée par le NIST (National Institute of Standards and Technologies)[7]. Initiées par 9 laboratoires en 1997, les campagnes NIST regroupent aujourd'hui plus d'une quarantaine de participants. Elles consistent à effectuer un classement des systèmes à partir d'une tâche spécifique comme par exemple :

- 2mn 30s de temps d'apprentissage
- 2mn30s de temps de test

Le détail des différentes tâches à exécuter est disponible sur le site des campagnes NIST SRE. L'idée est de transmettre à l'ensemble des participants une base de données commune de grande taille et de définir un calendrier précis de début d'évaluation et de restitution des résultats. Ces derniers sont présentés en général deux mois après les évaluations par chaque participant. La restitution des résultats est obligatoire dès lors que le laboratoire s'engage à participer. L'évaluation est d'autant plus significative que la base de données est importante mais aussi que le nombre de participants est conséquent.

Les résultats sont analysés à partir de courbe DET (Detection Error Tradeoff) telle que présentée sur la figure n°5.

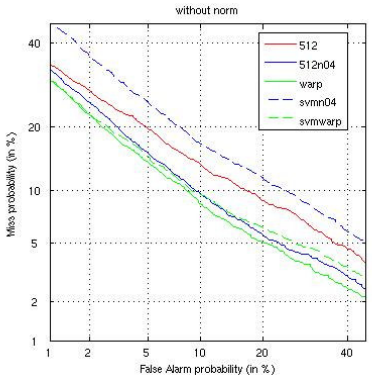

FIG n°5 : Exemple de courbe DET

Cette courbe est construite à partir de la distribution des score (figure n°6) client (individus devant être reconnus par le système) et imposteur (individus ne devant pas être reconnus par le système).

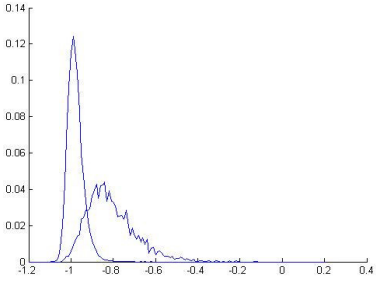

FIG n°6 : Distribution des scores

Deux types d'erreurs peuvent être commises par un système de reconnaissance automatique. Soit l'individu client n'est pas reconnu comme un client, nous sommes alors dans le cas d'un faux rejet. Soit l'individu imposteur est reconnu par le système comme un client, nous sommes dans le cas d'une fausse acceptation. Cette ambiguïté réside à l'intersection des gaussiennes client et imposteur. L'intersection entre la bissectrice et les courbes sur la figure n°5 représente l'EER (Equal Error Rate). En ce

point, nous avons une égalité entre les faux rejets et les fausses acceptations.
Aujourd'hui en regard des dernières évaluations NIST (2006) les meilleurs systèmes proposent la courbe ci-dessous :

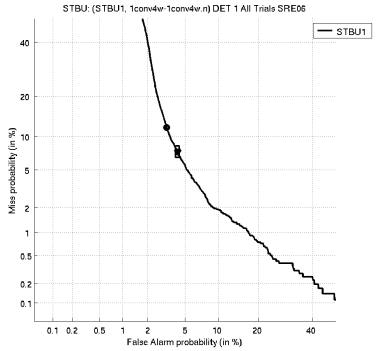

FIG n°7 : courbe DET NIST SRE 2006

Il existe d'autres évaluations, parfois plus adaptées à des conditions spécifiques comme celle organisées dans un cadre criminalistique : les évaluations organisées par NFI-TNO en 2003[7].

3. Reconnaissance du locuteur et sécurité

Il existe nombres d'applications de la reconnaissance de locuteur dont certaines sont directement liées aux technologies de sécurité. Nous nous attacherons donc à envisager la reconnaissance de locuteur dans le cadre des affaires judiciaires mais aussi comme un enjeu dans le domaine de la sécurité

3.1 La voix en matière d'expertise criminelle

En matière d'expertise judiciaire, la voix est un sujet qui fait débat. Pouvons-nous utiliser une telle modalité alors que le niveau de performance des systèmes n'est pas aussi fiable que pour les empreintes digitales ou l'empreinte génétique. Le mythe de l'empreinte vocale [5] a vécu et nous ne pouvons plus considérer la voix comme tel. Pour autant, nous sommes en général capables de reconnaître une voix familière au téléphone. Cela tend donc à prouver que la voix possède des caractéristiques propres au locuteur. En fait, la voix peut être considérée comme une modalité biométrique comportementale c'est-à-dire dépendante du comportement de l'individu, que ce comportement soit volontaire ou non. Ainsi considérée, la voix peut se rapprocher de la façon d'écrire, de la signature, ou du visage c'est-à-dire d'un ensemble de modalités biométriques lié au comportement. Dès lors, nous devons nous poser la question de savoir s'il est possible de prendre en compte la voix comme une caractéristique objective de reconnaissance d'un individu. Dans le cadre de l'expertise judiciaire deux approches prédominent : l'approche phonétique et l'approche automatique. Dans le cadre de cet article nous nous concentrerons sur l'aspect automatique de la reconnaissance. Pour autant, nous ne pouvons occulter l'approche phonétique qui apporte également des résultats pertinents. Elle présente des avantages certains en vue de reconnaître des particularités identifiantes de notre façon de parler. Ce sera des caractéristiques idiolectales, sociétales, régionales, liées aux défauts d'élocution, tics verbaux ou autres… Cette méthode présente cependant l'inconvénient d'être très liée à la compétence de l'expert et ne permet pas de définir un score de discrimination d'un individu par rapport à un autre. Il existe des sociétés savantes internationales comme l'IAFPA (International Association for Forensic Phonetics and Acoustics) qui permettent un échange entre experts du domaine. La reconnaissance automatique du locuteur quant à elle, présente ce double intérêt d'être à la fois reproductible et évaluable. Comme nous l'avons préalablement décrit, nous pouvons distinguer deux grandes applications en matière de reconnaissance automatique: l'identification et la vérification que nous regrouperons sous le vocable de reconnaissance. En général, la vérification est la tâche la plus commune en criminalistique. L'objectif est de comparer un enregistrement anonyme à la voix d'un suspect.
Le principe des systèmes automatiques a été décrit ci-dessus. Dans le cadre de l'approche criminalistique, ce qui nous intéresse c'est d'identifier les limites de la reconnaissance automatique, afin non pas de s'en affranchir car elle apporte indéniablement des éléments utiles à l'enquêteur, mais de maîtriser son usage dans les cas les plus adaptés.
Il existe malheureusement des limites objectives à l'emploi de cette modalité. Certaines de ces limites pourront être compensées, d'autres, en revanche, interdiront pour le moment toute analyse.

Ainsi, il apparaît nécessaire avant toute analyse forensique de prendre en compte différents critères que nous détaillons ci-après.

- qualité de l'enregistrement : la reconnaissance du locuteur s'effectue à partir d'enregistrements de bonne qualité de façon à ne pas avoir à compenser une dégradation du signal due à un bruit trop important. Cette qualité pourra être mesurée à partir du rapport signal sur bruit de l'enregistrement. En outre, il convient d'identifier dans la mesure du possible le canal de transmission de façon à disposer du même canal entre l'apprentissage et le test. Cela n'est malheureusement pas toujours possible.

- qualité de la parole : il convient également de prendre en compte les problématiques liées à l'intra variabilité du locuteur c'est-à-dire que la voix peut varier en fonction d'un état émotionnel ou de santé variable par exemple. Ensuite, cette parole doit être suffisamment riche en diversité et en quantité.

- La qualité de la base de données : la base utilisée doit être le plus proche possible des enregistrements des

individus suspectés. Cette base doit également être riche en diversité comme en quantité de voix pour être significative.

En matière de sécurité, les perspectives sont différentes car l'analyse consiste non pas à juger un individu à partir des éléments de sa voix mais à orienter les investigations, où à identifier un individu qui se prête généralement à l'identification.

3.2 La voix : un enjeu de sécurité

En matière de sécurité la reconnaissance de locuteur possède des applications potentielles très importantes comme :

- le contrôle d'accès à distance
- les services bancaires à distance
- l'accès à des informations classifiées
- biométrie
- incarcération à domicile

Contrairement aux applications forensiques, les applications mentionnées ci-dessus se distinguent par la volonté de l'individu d'être reconnu. En général, si un individu souhaite accéder à une pièce contrôlée, il collaborera volontairement à la phase de reconnaissance de même s'il souhaite accéder à son compte à distance. En outre, nous sommes en général dans une situation d'identification et non de vérification car le système possède un modèle de la voix de l'individu qui se présente. Les différents domaines d'application ci-dessus présente un intérêt évident pour l'utilisation des systèmes de reconnaissance automatique. La voix apparaît dans ces applications plus encore que dans le domaine forensique, comme une modalité biométrique utilisable par son niveau de performance qu'il faut toujours accroître, mais aussi sa facilité d'utilisation. En effet, dans ces différents cas le principal avantage réside dans la possibilité de contrôler l'environnement, le type et la qualité des enregistrements comme du protocole d'enregistrement. Comme nous l'avons vu en 3.1 cela permet de compenser un nombre significatif de limites. Pour autant, en raison de la sensibilité de ces applications, il est évident qu'il convient de s'intéresser dès aujourd'hui aux techniques d'imposture qui demeurent une menace très présente. Nous aborderons un certain nombre de possibilités d'imposture en 4.

Au-delà de l'expertise judiciaire la voix peut constituer un atout considérable pour authentifier un enregistrement, que celui-ci soit audio ou vidéo. L'enregistrement vidéo adressé par les FARC (Forces Armées Révolutionnaire de Colombie) et transmettant un message d'Ingrid Betancourt, mais aussi bien entendu les différents enregistrements de Ben Laden constituent des exemples très concrets. Ces deux cas particuliers font l'objet de couverture médiatique mais il existe nombre d'applications, où la voix peut significativement orienter les enquêteurs vers l'identification d'individu particulier.

4. Les techniques d'imposture

En matière de sécurité, il est indispensable de s'intéresser très précisément aux phénomènes d'imposture et d'usurpation d'identité. En effet en science criminelle où la collaboration de l'individu n'est pas évidente a priori comme dans le cadre d'applications de sécurité où cette collaboration est évidente, la problématique de l'imposture apparaît comme un sujet particulièrement sensible. Nous allons nous attacher dans cette partie à décrire les techniques d'imposture délibérée possibles mais aussi à en mesurer l'impact sur les systèmes de reconnaissance automatique.

En reconnaissance de locuteur, il est possible de modifier sa voix délibérément de différentes façons. Nous nous intéresserons au déguisement volontaire de la voix, c'est-à-dire à la façon de changer sa voix afin de la rendre méconnaissable, mais aussi aux techniques de conversion de la voix c'est-à-dire aux techniques d'imitation.

4.1 La question du déguisement

En matière de déguisement, les techniques ne manquent pas, de l'utilisation d'un mouchoir devant la bouche, à la modification de la fréquence de parole en passant par l'utilisation de moyens électronique et/ou informatique, le délinquant a à sa disposition nombre de possibilités. Il est intéressant de mesurer tout d'abord l'impact du déguisement sur la reconnaissance de locuteur (figure n°8). Pour cela nous avons développé un outil de reconnaissance basique, fondé sur une modélisation à mélange de gaussiennes pour créer des modèles de voix d'individus et nous avons mesuré l'évolution de la vraisemblance de la reconnaissance avant et après déguisement. Nous nous intéressons dans le cadre de cette étude à quatre déguisements parmi les plus employés : main devant la bouche, le nez pincé, la voix aigue et la voix grave. Nous pouvons en effet nous poser la question avant de travailler sur la reconnaissance d'un locuteur, de la possibilité du déguisement de sa voix.

L'évolution constatée avant et après déguisement est la suivante :

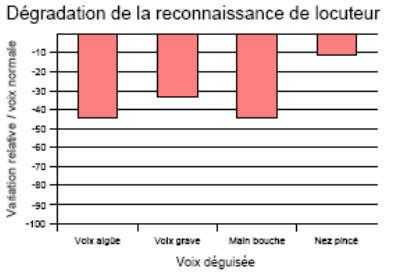

FIG n°8 : Impact des voix déguisées sur la reconnaissance

Nous constatons donc le type de déguisement qui semble altérer le plus significativement la reconnaissance de locuteur. En effet, en dehors du nez pincé nous constatons une dégradation du niveau de performance de reconnaissance de plus de 35%. Cette étude préliminaire a été effectuée sur une base de 15 locuteurs. Il convient donc de déterminer en préalable à la reconnaissance de locuteur la possibilité d'un déguisement. Cette étape est une étape de détection. Dès lors que celle-ci est réalisée, nous nous attacherons à identifier le type de déguisement effectué parmi les quatre étudiés.

La détection, comme l'identification, repose sur une phase d'extraction et une phase de classification des paramètres. Les paramètres choisis sont les MFCC (Mel Frequency Cesptral Coefficient).

4.1.1 Principe algorithmique de classification

Les techniques de classifications employées sont les suivantes :

- k-plus proches voisins
- GMM (Gaussian Mixture Models)
- SVM (Support Vector Machine – Séparateurs à vastes marges)

Nous avons étudié les performances de ces trois algorithmes en matière de détection, et nous sommes concentrés sur la classification GMM pour l'identification.

La technique des k-plus proches voisins consiste à affecter un élément à une classe en fonction de la classe d'appartenance de ses voisins. Dans le cas de la figure n°9 l'échantillon vert à classer appartiendra à la classe bleue si nous considérons la classe de ces deux plus proches voisins, alors qu'il appartiendra à la classe rouge si nous considérons la classe de ses cinq et plus « plus proches voisins ».

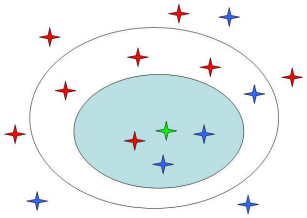

FIG n° 9 K plus proches voisins

La classification fondée sur les séparateurs à vastes marges ou Support Vector Machine (SVM) repose sur l'existence d'un classifieur linéaire binaire dans un espace adapté. C'est une méthode de classification par apprentissage supervisé introduite en 1995 [12]. La première étape consiste à apprendre les paramètres du modèle à partir du jeu de données de façon à définir la frontière ultérieure de décision. Cette étape utilise des fonctions appelées « noyaux » qui optimisent la séparation des données. Deux cas sont possibles : les données sont linéairement séparables et alors le classifieur linéaire est rapidement déterminé ou les données sont linéairement non séparables et alors, il convient d'accroître la dimension de l'espace des données. Nous sommes dans la plupart des cas dans cette situation. Le but est de déterminer non plus une droite (en dimension 2) capable de discriminer les classes, mais un hyperplan dans le nouvel espace. Nous comprenons que plus nous augmentons la dimension de l'espace, plus nous augmentons la probabilité de déterminer un hyperplan séparateur optimal (celui qui maximise la distance minimale aux exemples d'apprentissage). C'est cette transformation non linéaire qui est réalisée à partir d'une fonction noyau paramétrable. Différents types de noyau (polynomiale, gaussien, sigmoïde ou Laplacien) pourront être évalués afin de déterminer le plus adapté à l'application.

Après cette phase d'apprentissage il convient dans la phase de test de prédire la classe de nouveaux paramètres à partir de la classification effectuée et de l'hyperplan déterminé.

La classification à partir des GMM est également une classification fondée sur les méthodes d'apprentissage supervisé. La phase d'apprentissage consiste à modéliser une base de données de voix normales et une base de données de voix déguisées à partir d'un mélange pondérée de gaussiennes. L'estimation des paramètres propres au mélange de gaussienne (moyenne, matrice de covariance, coefficient de pondération) est obtenue à partir de l'algorithme itératif EM (Expectation – Maximisation). Suite à la phase d'apprentissage nous disposerons donc d'un modèle de voix normales et de voix déguisées pour la partie détection et d'un modèle propre à chaque déguisement pour la partie identification.

La phase de test consiste à mesurer une distance de similarité entre les nouveaux paramètres à classer et les modèles. Cette distance est le maximum de vraisemblance d'appartenance d'un échantillon à une classe. La décision d'appartenance sera prise à partir de la plus grande vraisemblance obtenue.

4.1.2 Expérimentation et résultats

Le corpus de parole utilisé dans la partie détection est le suivant :

Apprentissage : deux modèles de voix (normale et déguisée) sont créés à partir d'un échantillon de 5mn de parole lue. Ces 5 mn comprennent 16 locuteurs.

Test : 26 locuteurs test à partir de segment de parole de 20s : chaque locuteur prononce des phrases lues phonétiquement équilibrée dans chacun des déguisements et avec une voix normale. En conséquence les tests ont été effectué sur 26 voix normales et sur 102 voix déguisées.

Préalablement à l'extraction des paramètres MFCC, nous effectuons une étape d'extraction des portions de silence à partir d'une représentation bi gaussienne de l'énergie.

L'utilisation des « 20 plus proches voisins » apporte des résultats satisfaisants en matière de détection de voix déguisées mais encore un peu juste pour détecter les voix normales. Cependant, dans le cadre d'une application pré reconnaissance de locuteur il est préférable de détecter une voix déguisée même si certaines voix normales seront confondues avec une voix déguisée.

TAB. 1 : détection 20 plus proches voisins

Type de voix	Normale	Déguisée
Normale	**62%**	38%
Déguisée	22%	**78%**

Le véritable inconvénient de cette méthode est le temps de calcul qui altère considérablement l'utilisation des Kppv.

L'emploi des GMM s'est quant à lui révélé peu satisfaisant en matière de détection de voix normales. En effet, la plupart des voix ont été reconnu comme déguisées. Cette différence trop importante rend peu intéressante l'emploi des GMM pour cette partie détection.

TAB. 2 : détection GMM

Type de voix	Normale	Déguisée
Normale	**15%**	85%
Déguisée	6%	**94%**

Les SVM enfin apporte quant à eux des résultats très encourageants.

TAB. 3 : détection QV (2048)+SVM

Type de voix	Normale	Déguisée
Normale	**66%**	34%
Déguisée	4%	**96%**

Ces résultats sont, en effet, encourageants, car il ne résulte pas uniquement de l'emploi des SVM mais de l'utilisation d'une quantification vectorielle avant l'application des SVM. Le résultat affiché au tableau n°3 présente donc la classification après une quantification vectorielle faisant usage de 2048 centroïdes puis usage des SVM.

Or, même si la détection des voix déguisées est efficace, nous nous apercevons que le nombre de bonne détection de voix normale croît avec le nombre de centroïdes (voir tableau n°4,5,3). Cela est parfaitement cohérent, car l'usage des centroïdes déplace le choix des bons vecteurs supports. L'emploi de la quantification vectorielle était motivé par le gain en temps de calcul et une moindre complexité. Il convient donc de prolonger cette étude par l'emploi des SVM uniquement.

TAB. 4 : détection QV (512)+SVM

	Normale	Déguisée
Normale	**43**	67
Déguisée	9	**91**

TAB 5 : détection QV (128)+SVM

	Normale	Déguisée
Normale	24	76
Déguisée	6	**94**

En terme d'identification, seule une classification à base de GMM a été utilisée. Les résultats sont là aussi satisfaisants en particulier pour certains déguisements spécifiques.

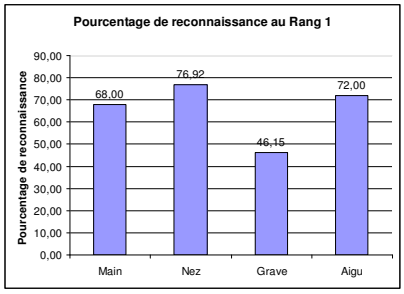

FIG n°10 : Identification par déguisement

La figure n°10 illustre l'identification de chaque type de déguisement en première reconnaissance.

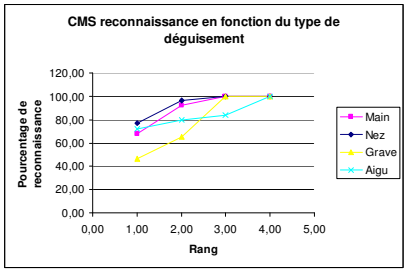

FIG n°11 : Identification cumulée

La figure n°11 illustre l'identification de chaque type de déguisement en reconnaissance cumulée.

4.2 La question de l'imitation

4.2.1 L'imitation par un professionnel

La première catégorie de conversion que nous présentons est celle qui apparaît comme la plus évidente, c'est-à-dire celle d'un imitateur professionnel. En effet, nous sommes tous habitués à ce type de prestation dans le cadre d'applications plutôt orientées vers le divertissement. Pourtant, l'utilisation de ce type de procédé vers des applications forensiques ou biométriques doit être envisagée. La question qui consiste à se demander quel serait le résultat d'une imposture réalisée par un imitateur professionnel n'est donc pas inintéressante. En outre, une telle étude permet de comprendre le mécanisme de l'imitation et de déterminer les paramètres auxquels nous sommes le plus sensible. En d'autres termes, nous devrions pouvoir extraire les paramètres les plus discriminants pour un locuteur donné. Cette question a été étudiée dans le détail par E. Zetterholm [2]. Elle s'est intéressée à sa langue d'origine : le suédois. Son étude s'appuie sur trois imitateurs professionnels différents. Nous savons qu'un imitateur imite non seulement la voix mais aussi la gestuelle et l'attitude de la personne imitée. Ici, ce ne sera pas le cas, seule la voix sera imitée et les résultats seront analysés via le canal téléphonique. Ils ne seront donc que perceptifs. Ce travail a montré que l'imitateur est en mesure de saisir différents aspects de la voix cible, parfois jusqu'à l'exagération. Ces diverses facettes de la voix sont le pitch, la qualité vocale, la prosodie et le style. L'impression globale pour celui qui écoute est largement satisfaisante même si l'ensemble des paramètres n'est pas reproduit, il en suffit de quelques-uns. A cela s'ajoute d'autres paramètres qui correspondent à des caractéristiques régionales, sociales, dialectales, ainsi qu'à des habitudes phonétiques. En effet, notre voix se construit à partir de l'évolution de caractéristiques physiques (conduit vocal…), mais aussi de notre environnement sociologique. Selon Anders Martensson, imitateur depuis dix ans, interrogé par E. Zetterholm, toutes les voix ne sont pas reproductibles. La plupart des voix de femmes et de jeunes enfants sont, pour un homme difficile à imiter. A cela s'ajoute la difficulté de parvenir à imiter certains dialectes et certains registres de pitch. Comme nous l'avons expliqué ci-dessus, l'imitateur s'attache à reproduire certains des paramètres caractéristiques de l'individu, cependant il n'a pas été possible d'établir un ordre de préférence dans le choix des paramètres à privilégier. Ils dépendent des aptitudes techniques et physiologiques de l'imitateur mais aussi des caractéristiques de la voix cible.

Cette étude révèle des résultats qui s'appuient sur des tests perceptifs. Capable de tromper une oreille humaine nous pouvons nous interroger sur l'efficacité de ces imitations face à un système automatique de vérification du locuteur. Une telle démarche a été envisagée par M. Blomberg (2004). Le système de vérification utilisé est dépendant du texte. Les mots prononcés sont des séquences de chiffres. Le calcul du score est basé sur le rapport des log vraisemblance. Il est utilisé comme résultat et la décision d'accepter ou de rejeter s'effectue à partir d'un seuil déterminé. Le système a été utilisé par Melin, Koolwaaij, Lindberg et Bimbot (1998). Ce travail s'est limité à un imitateur pour trois voix cibles dont les caractéristiques individuelles ont été choisies précisément. La première est perceptuellement qualifiée de très proche de la voix de l'imitateur, la seconde de moyennement proche et la troisième d'éloignée. Les résultats ont montré que les imitations étaient performantes pour perturber un système de reconnaissance automatique, particulièrement celle ayant la voix la plus proche de celle de l'imitateur.

Ainsi, cette première approche de conversion vocale montre des résultats capables de perturber à la fois l'oreille humaine et, sous certaines conditions, un système de reconnaissance automatique. L'avantage de ce type d'imposture est l'indépendance au texte et plus ou moins du locuteur. Cependant, les résultats obtenus par ce type d'expérience s'avèrent très liés à la qualité de l'imitateur et aux caractéristiques de la voix cible. De fait, c'est une méthode peu reproductible. En outre, il est à espérer que la finalité de l'imitation reste le divertissement et qu'un imitateur professionnel n'est pas a priori tenté par l'expérience de l'illégalité. C'est pourquoi, il est intéressant d'étudier d'autres possibilités de transformation, qui diffèrent totalement dans le principe de réalisation. Ce sont les conversions automatiques.

4.2.2 L'imitation par conversion automatique

La conversion automatique de la voix a fait l'objet, ces vingt dernières années, de nombreuses études. Les récentes campagnes de test des systèmes de reconnaissance automatique ont révélé la capacité de tels systèmes à rejeter des imposteurs, mais aussi malheureusement, la capacité à les accepter. Or, ces imposteurs que nous qualifierons de non clients ne cherchent pas à tromper le système. Il est donc légitime de s'interroger sur la robustesse d'une plateforme de reconnaissance automatique de locuteur lorsque l'imposture est délibérée.

La conversion automatique de voix peut se décomposer en deux étapes principales:

- apprentissage
- test

L'étape d'apprentissage consiste à déterminer la fonction de conversion entre un locuteur source et un locuteur cible. Pratiquement cette fonction est déterminée à partir de phrase identique prononcée par la source et la cible. Ces deux phrases doivent être préalablement aligné par DTW (Dynamic Time Warping). Dès lors, il convient de trouver la fonction qui permettra de transformer les paramètres d'une voix source vers ceux d'une voix cible.

L'étape de test, quant à elle, consiste à appliquer cette fonction de conversion à une phrase prononcée par la voix source. Il existe plusieurs possibilités de mesurer le niveau de performance de telles conversions. Des tests perceptifs tout d'abord permettent de se rendre compte à l'oreille du

résultat de la conversion. La difficulté est que ce type d'évaluation repose non seulement sur la qualité de la conversion mais aussi sur la qualité de la synthèse effectuée pour reconstruire le signal. Une autre possibilité est d'évaluer la conversion par le calcul d'une mesure de distance spectrale entre les paramètres source-cible, source transformée-cible, et source transformée-source. Enfin une dernière possibilité est de mesurer le déplacement de la courbe des scores vers les imposteurs et de constater le niveau de dégradation sur les courbes DET.

Une bonne transformation de la voix doit simuler les modifications des caractéristiques du conduit vocal, la prosodie, ainsi que l'excitation glottale. Nous ne sommes pas encore parvenus à satisfaire à l'ensemble de ces transformations. Les techniques développées se sont principalement intéressées à la transformation de l'espace acoustique d'un locuteur (source) vers un autre locuteur (cible). La performance d'un système automatique de conversion de la voix repose essentiellement sur les facteurs suivants:

- la quantité de données disponibles pour l'entraînement des modèles du locuteur source et du locuteur cible
- l'efficacité des modèles (quantification vectorielle, HMM (hidden markov models), GMM (Gaussian mixture model)) et des paramètres acoustiques (MFCC – LPCC)
- la technique d'élaboration de la fonction de conversion employée pour minimiser les différences acoustiques entre les deux locuteurs.
- La mise en œuvre de la synthèse (HNM, PSOLA…)

La combinaison de ces facteurs nous permet de dresser les différentes méthodes utilisées en conversion de voix. Plusieurs techniques apparaissent dans la littérature. Nous évoquerons les principales :

- conversion basée sur la quantification vectorielle [1]
- conversion basée sur une régression linéaire LMR (Linear Multivariate Regression). [11]
- conversion basée sur l'utilisation des GMM. [10]
- conversion basée sur l'indexation d'une mémoire cliente.[8][6]

Nous nous intéressons spécifiquement à deux techniques (celle fondée sur la LMR et celle sur l'indexation dans une mémoire cliente) et présentons leur niveau de dégradation. La LMR consiste à modéliser le mapping pour chaque classe après une transformation linéaire.

Soit $X_{s,q} = \{X_{s,q,j}\}$, avec j= 1 à M_q , l'ensemble des vecteurs de paramètres spectraux du locuteur source pour la q ième classe.

Soit $Y_{c,q} = \{ Y_{c,q,j}\}$, avec j = 1 à M_q, l'ensemble des vecteurs de paramètres spectraux du locuteur cible pour la q ième classe après alignement temporelle (DTW).

M_q est le nombre total de vecteur dans la q ième classe.

La LMR permet donc de trouver la transformation linéaire qui minimise l'erreur quadratique moyenne entre l'ensemble des vecteurs de la source et ceux de la cible. Si $m_{s,q,j}$ et $m_{c,q,j}$ sont les moyennes de la j ième composante des vecteurs spectraux respectivement de la source et de la cible, les vecteurs normalisés sont :

$$\tilde{X}\,s,q,k = Xs,q,k - ms,qk$$

et alors la transformation de régression linéaire Pq qui minimise l'erreur quadratique moyenne entre les deux vecteurs normalisés est celle qui minimise :

$$\sum_{k=1}^{Mq} \left\| \tilde{X}\,c,q,k - Pq\,\tilde{X}\,s,q,k \right\|^2$$

La solution à ce problème des moindres carrés est le produit de $X\,c,q,k$ par la pseudo-inverse de $X\,s,q,k$.

Pour appliquer cette technique à la conversion de la voix, on suit la procédure suivante pour chaque vecteur :

- détermination de la classe d'appartenance de chaque vecteur
- application de la transformation pour chaque vecteur
  - normalisation du vecteur : le vecteur transformé normalisé est obtenu par multiplication de la matrice Pq par le vecteur normalisé original.
  - « dénormalisation » du vecteur
  - calcul du spectre de puissance
  - reconstruction du signal

Les résultats obtenus sur 200 individus de la base de données BANCA sont présentés sur les figures n°12 et 13. Nous constatons un mouvement significatif de la courbe des imposteurs vers la courbe des clients.

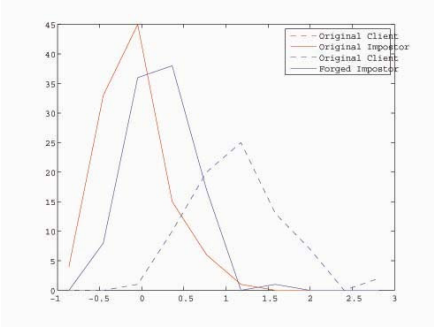

FIG n° 12 : Distribution des scores (LMR)

L'impact sur la courbe DET est également très significatif (figure n°13).

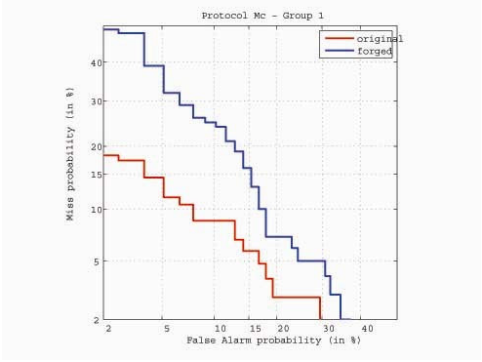

FIG n°13 : Courbe DET (LMR)

Outre cette technique de conversion, nous pouvons également présenter la conversion fondée sur l'indexation au sein d'une mémoire cliente. Cette technique est décrite en détail dans [8]. Elle s'appuie sur le codeur à très bas débit ALISP (Automatic Language Independent Speech Processing) [6]. Celui-ci nous permet de segmenter de façon automatique un signal de parole, mais aussi de construire une mémoire des segments acoustiques de la voix cible. Nous coderons ensuite la voix source afin d'en déterminer les caractéristiques acoustiques puis irons puiser dans la mémoire les segments de la voix cible pour remplacer les segments de parole de la voix source. Nous utiliserons la base de données BREF pour des essais préalables, puis la base NIST pour évaluer notre imposture. Cette évaluation sera effectuée à partir d'un système automatique de reconnaissance « Becars »[3]. L'utilisation d'un système automatique pour déceler l'imposture apparaît utile, non pour contester les performances d'un tel système mais plutôt pour en déterminer le comportement face à une imposture délibérée, pour à l'avenir, permettre d'effectuer des détections automatiques de ces conversions. Les résultats en terme de distribution des scores sont également très significatifs de même que l'impact sur la courbe DET.

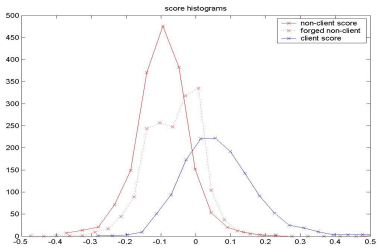

FIG n°14 : distribution des scores

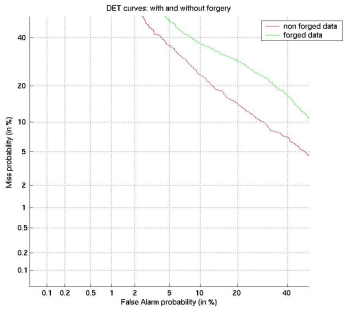

FIG n°15 : courbe DET (conversion par indexation)

Ainsi l'imposture qu'elle soit automatique ou manuelle (phonétique) constitue une véritable problématique pour les systèmes de reconnaissance de locuteur. Néanmoins étudier ces impostures, en mesurer l'impact sur les systèmes automatique mais aussi sur la perception, permettent de mieux comprendre les faiblesses de la reconnaissance et constitue donc une véritable aide à l'interprétation des résultats.

Conclusion

Comme nous pouvons le constater les systèmes automatiques ont largement progressé mais il convient de garder à l'esprit que l'application de la reconnaissance de locuteur dans le domaine opérationnel nécessite de bien maîtriser son sujet de façon à définir les limites de la reconnaissance et d'être en mesure d'interpréter le résultat le plus objectivement possible. Les possibilités d'imposture délibérées sont également très importante en ce qui concerne la modalité voix. En effet celle-ci qui présente un intérêt indéniable en terme de performance, de coût et de facilité d'emploi est malheureusement sujette à des possibilités d'imposture qui fragilisent significativement sa robustesse. L'étude de ces attaques est donc indispensable d'une part pour les détecter mais aussi mesurer leur influence sur les performances et atténuer leurs effets.

Références

[1] M. Abe, S. Nakamura, K. Shikano, H. Kuwabara – *Voice conversion through vector quantization* – ICASSP 1988

[2] Mats Blomberg, Daniel Elenius, Elisabeth Zetterholm Speaker verification scores and acoustic analysis of a professional impersonator Proccedings FONETIK 2004

[3] Raphaël Blouet and al. *Becars: a free software for speaker verification* Odyssey pages 145-148 – 2004

[4] R. Blouet – C. Mokbel – G . Chollet - Becars: un logiciel libre pour la verification du locuteur – Speaker Odyssey – 2004

[5] L.J. Boë - *Ben Laden et le mythe de l'empreinte vocale* – revue Vivant n°1 – 2003

[6] G. Chollet, J. Cernocky, A. Constantinescu, S. Deligne, F. Bimbot - *Toward ALISP : a proposal for automatic language independent speech processing* – Computational Models of Speech Processing – NATO ASI Series – 1997

[7] David A. van Leeuwen, Alvin Martin, Mark A. Prrzybocki, Jos. Bouten. *NIST and NFI-TNO evaluation of automatic speaker recognition* – Computer and Speech & Langage, vol 20, Issues 2-3, April-July 2006

[8] P. Perrot, G. Aversano, G. Chollet - *Voice forgery using ALISP*- Proceedings ICASSP 2005, Philadelphie

[9] D.A. Reynolds, R.C. Rose - *Robust text-independent speaker identification using Gaussian mixture speaker models* – IEEE – Trans. Speech and audio processing – 1995 –

[10] Yannis Stylianou, Olivier Cappé, Eric Moulines - *Continuous probabilistic transform for voice conversion* 1998

[11] H. Valbret, E. Moulines, J.P. Tubach – *Voice transformation using PSOLA technique* ICASSP 1992

[12] Vladimir Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, 1995.