La reconnaissance du locuteur : un problème résolu ?

Jean-François Bonastre, Driss Matrouf

Université d'Avignon, LIA
339 ch des Meinajariès, BP 1228, 84911 Avignon CEDEX 9, France
jean-francois.bonastre,driss.matrouf@univ-avignon.fr
http ://www.lia.univ-avignon.fr

ABSTRACT

This paper presents a brief summary of the progress made in the speaker recogniton field during the past decade. It tries to show that, even if these progress were drastic in terms of error rates, some questions are still open. It concludes with some proposal for future research works in the field of speaker characterization and recogniton.
Keywords: `Speaker recognition, GMM, SVM`

1. Introduction

La voix est une modalité biométrique compétitive pour plusieurs raisons bien connues. En particulier, cette modalité est souvent la seule disponible pour de nombreux types d'applications, dans le domaine commercial ou dans le champs de la criminalistique.
Durant ces dernières années, les progrès enregistrés dans le domaine de la reconnaissance du locuteur, notamment dans le cadre des campagnes d'évaluation internationales organisées par le NIST [10], ont été très impressionnants. En particulier, l'émergence de techniques capables de compenser les différences induites par l'usage de différents micros et de différentes liaisons téléphoniques - comme le Latent Factor Analysis (FA) ou Nuisance Attribute Projection (NAP) - autorise un niveau de performance très attrayant. Le niveau de performance atteint pousse à mettre en place des applications opérationnelles, ce qui est évidemment un souhait honorable.
Durant la même période, plusieurs scientifiques représentant des sociétés savantes ont attiré l'attention sur les dangers d'une transposition directe des résultats obtenus dans le cadre d'une application dans le milieu judiciaire [5], expliquant notamment qu'une évaluation scientifique - même menée de façon rigoureuse - reste appuyée sur une base de données représentant un nombre fini et limité de facteurs de variabilité et ne peut pas être généralisée à tout cadre applicatif, surtout dans le domaine criminalistique.
Cet article fait le point sur les progrès réalisés et interroge la communauté sur la pertinence des seuls taux d'erreur pour apprécier les progrès dans le domaine de la reconnaissance du locuteur. Le contexte expérimental constitué par le cadre des campagnes d'évaluation NIST est présenté dans la section 2. L'approche dominante en reconnaissance du locuteur, autour de la modélisation statistique GMM/UBM est décrite dans la section 3. La section 4 retrace les progrès réalisés ces dernières années. La section 5 pose le problème de l'évaluation des performances et de l'orientation générale des travaux de recherche, guidée par le seul objectif de réduire les taux d'erreur. La section 6 conclut et présente quelques pistes alternatives pour de futurs travaux.

2. Le cadre des évaluations NIST

L'histoire des campagnes d'évaluation NIST-SRE (Speaker Recognition Evaluation) a débuté en 1996 et s'est poursuivie par l'organisation de campagnes de façon quasi-annuelle depuis cette date. Ces campagnes ont pour objectif principal de proposer un cadre intégré d'évaluation de différents systèmes et approches : les participants utilisent les mêmes corpus et protocoles, les mêmes instruments de mesure des performances et sont synchronisés dans leur travaux par l'échéancier de la campagne. L'attrait majeur pour les participants est l'accès -gratuit- à des corpus oraux de grande taille, renouvelés ou étendus quasiment chaque année. Du coté du sponsor, le DoD, l'intérêt est double : d'une part il peut vérifier les progrès réalisés et les méthodes en devenir et, d'autre part, le sponsor influe sur l'orientation des recherches en faisant évoluer les protocoles. Le succès de ces campagnes se confirme d'année en année, avec plus de 30 participants en 2006 et une prépondérance des travaux reliés à ces campagnes dans les publications scientifiques.
Le contexte expérimental des évaluations NIST-SRE concerne une tâche de vérification du locuteur, basée sur de la parole téléphonique conversationnelle en mode "indépendant du texte". Il s'agit de répondre à la question "est-ce que le locuteur X a prononcé l'enregistrement y ?". Les enregistrements sont composés de conversations téléphoniques d'une durée moyenne de 5 minutes, soit 2m30 pour chaque locuteur, obtenus avec plusieurs téléphones pour chaque locuteur, majoritairement cellulaires. Nous nous concentrerons dans ce papier sur la tâche principale dite "1conv-1conv", dans laquelle les données d'entraînement et de test sont constituées par un seul enregistrement (soit en moyenne 2m30 de parole). L'ensemble des résultats présentés sur ce papier sont obtenus sur la base NIST 2005, réduite aux hommes.

3. L'approche GMM-UBM en reconnaissance du locuteur

L'approche GMM-UBM (Gaussian Mixture Model - Universal Background Model) est la technique prédominante en reconnaissance du locuteur, en mode indépendant du texte [3]. Lors d'un test, le système cherche à déterminer si l'enregistrement Y a été prononcé par un locuteur S. Cette prise de décision est modélisée par le rapport de vraisemblance :

$$\frac{p(Y|\lambda_{hyp})}{p(Y|\lambda_{\overline{hyp}})} \quad (1)$$

où Y est le segment de parole à tester, λ_{hyp} le modèle correspondant au cas où S a prononcé Y, $\lambda_{\overline{hyp}}$ corres-

pond au modèle de l'hypothèse inverse, $p(Y|M)$ est la fonction de vraisemblance et θ, le seuil de décision.
Les deux modèles sont des modèles à mélange de lois gaussiennes (GMM) :

$$p(x|\lambda) = \Sigma_{i=1}^{M} w_i N(x|\mu_i, \Sigma_i) \quad (2)$$

avec w_i, μ_i et Σ_i, les poids, les vecteurs moyennes et les matrices de covariance (en général diagonales) des différentes composantes du mélange. Le modèle $\lambda_{\overline{hyp}}$ est dénoté modèle du monde, ou Universal Background Model (UBM) quand il est indépendant de l'environnement. Il est estimé à l'aide d'une collection d'extraits vocaux provenant d'un grand ensemble de locuteurs, par l'algorithme EM, en maximisant la vraisemblance des données d'apprentissage. Le modèle λ_{hyp} est dérivé de l'UBM, par une technique de maximisation de la probabilité *aposteriori* (MAP), à partir d'un exemple de la voix du locuteur concerné. Seules les moyennes des composantes sont adaptées et les autres paramètres sont directement issus de l'UBM.

4. Evolution des performances

Malgré le facteur facilitant que représente l'organisation d'une série continue de campagnes d'évaluation par le NIST, l'évolution des tâches et protocoles au fil des années ne facilite pas le suivi des progrès sur une longue période. Nous nous limiterons ici à la période 2004-2008, pendant laquelle la difficulté de la tâche est restée quasiment constante (quelques résultats antérieurs seront tout de même proposés de manière simplifiée). Enfin, nous nous appuierons majoritairement dans cette partie sur les travaux réalisés au LIA[1], globalement représentatifs de l'évolution générale du domaine. Les différents systèmes présentés sont intégrés dans le système "open source" ALIZE/SpkDet [4].

Le système LIA-04 présenté lors de la campagne NIST 2004, a obtenu des résultats à l'état de l'art pour un système GMM-UBM. Après une optimisation de la paramétrisation acoustique (accroissement à 50 de la dimension des vecteurs acoustiques) qui a mené à une amélioration relative de 10% du taux d'égale erreur, ce système a servi de base aux travaux décrits dans ce papier, sous la dénomination "GMM-UBM". La majorité des résultats expérimentaux présentés dans ce papier sont issus de [8].

4.1. L'arrivée des systèmes mixtes GMM/SVM

A partir de l'intérêt soulevé par les classifieurs discriminants tels que les SVM, une évolution importante a été proposée, notamment par [7]. Elle met en oeuvre une approche mixte associant la robustesse de la modélisation statistique proposée par le GMM-UBM et le pouvoir discriminant des SVMs. Cette approche, dénotée GMM supervecteurs Svm à noyau Linéaire" (GSL en abrégé) utilise le GMM-UBM pour modéliser les données d'apprentissage ou de test et en extraire un condensé, sous la forme d'un supervecteur composé des coefficients des vecteurs *moyenne* de chacune des composantes du GMM. Ces vecteurs sont utilisés comme entrée d'un classifieur SVM (à noyau linéaire). Le tableau 1 montre les performances d'un tel système GSL, comparé au système de référence GMM-UBM. Le gain relatif entre le système GSL et le système UBM/GMM de référence est d'environ 8% en termes d'EER[2] et de 17% en termes de coûts de décision (minDCF).

4.2. La normalisation des effets dus à la session

Durant la dernière décennie, une très large majorité des efforts de recherche dans le domaine de la reconnaissance du locuteur ont été consacrés au problème de l'écart entre l'apprentissage et le test (*mismatch* en anglais). Cet écart est composé de tous les facteurs de variabilité autre que la variabilité inter-locuteur pouvant intervenir entre deux enregistrements : environnement, microphone et canal d'enregistrement et de transmission mais aussi état psycho-pathologique du locuteur, contenu linguistique, dérive temporelle de la voix, etc. Les travaux ont surtout porté sur les premiers facteurs de variabilités cités dans cette énumération, par l'orientation de NIST, guidée par les objectifs de leur sponsor. De nombreuses solutions ont été proposées pour normaliser les données acoustiques [14, 12] ou les scores [1].
Plus récemment, une nouvelle classe d'approches a été ouverte par les travaux sur le FA [9] au sein des modèles statistiques génératifs (GMM) et par Solomonoff et al [16] dans le cadre des classifieurs discriminants SVM, avec NAP. Le but commun de ces approches est de modéliser directement les différences inter-session plutôt que de compenser leurs effets. Cette modélisation est basée sur l'analyse de très grandes bases de données, dans lesquelles, par exemple, un locuteur donné prononce un grand nombre d'enregistrements au cours de nombreuses sessions et en utilisant des combinés téléphoniques différents. Dans l'exemple précédent, la variabilité due au type de combiné téléphonique (et de liaison) est directement visée. Il est à noter que, dans les deux approches, le problème sous-jacent, est situé dans l'espace des supervecteurs, soit dans un espace de grande dimension (plus de 25000 dans notre cas). Les deux approches ont été implémentées dans le toolkit ALIZE/SpkDet [8, 11].
Le tableau 1 présente les résultats pour FA et NAP. L'apport de ces approches est très clairement mis en évidence avec une réduction de moitié du minDCF et de l'EER, par comparaison avec le système UBM-GMM de référence.

Tab. 1: Performances pour différents systèmes avec et sans normalisation des écarts inter-session

Système	EER %	minDCF (*100)
LIA04	9.64	
GMM-UBM	8.67	3.37
GSL	8.02	2.79
GSL-NAP	5.28	1.69
GMM-FA	4.38	1.94

4.3. L'apprentissage non supervisé

Depuis la campagne 2004, une tâche de vérification en mode "apprentissage non supervisé" est proposée par le NIST. Le protocole correspondant traite les test dans l'ordre chronologique, locuteur cible par locuteur cible, et l'information provenant des

[1] Il est à noter que ces résultats sont souvent issus de collaborations avec d'autres partenaires, comme Swansea University.

[2] Par rapport au système LIA04, le gain relatif en termes d'EER est d'environ 17%.

tests précédents peut être utilisée pour répondre au test courant. L'idée sous-jascente est d'autoriser le système à améliorer les modèles des locuteurs cibles de façon non supervisée, i.e. nous ne savons pas si les tests précédents ont été prononcés ou non par le locuteur cible. Ce protocole est devenu le protocole principal de la campagne depuis 2006.
Différentes équipes ont proposé des solutions pour répondre à cette nouvelle piste [2, 17]. Le LIA a développé une approche dite "continue" [13] dans laquelle toutes les données observées sont intégrées à l'apprentissage du modèle du locuteur, au fur et à mesure de leur arrivée, associées à une mesure de confiance qui pondère leur influence. Cette approche a été appliquée au système GMM-UBM comme au système GSL. Le tableau 4.3 résume les résultats expérimentaux. Tous les systèmes présentés intègrent la normalisation de la session FA (les systèmes GSL associent ici un GMM avec FA et un SVM). Les

Tab. 2: Performance sans adaptation, avec adaptation non supervisée et avec adaptation en mode "oracle".

Système	EER %	minDCF(*100)
GMM (référence)	8.67	3.37
SFA	4.55	1.59
SFA-unsupervised	2.36	0.89
SFA-oracle	1.62	0.50
GSl-FA	4.48	1.62
GSL-FA-unsupervised	2.27	0.81
GSL-FA-oracle	1.71	0.56

résultats montrent que la quantité de donnée pour l'apprentissage des modèles est un facteur clé de performance. L'adaptation non supervisée proprement dite autorise une amélioration relative proche de 50%, pour l'EER comme pour le minDCF, par comparaison au système équivalent sans adaptation. Lorsqu'un oracle est utilisé (le système sait si un test provient ou non du locuteur ciblé avant de réaliser l'adaptation), l'EER est divisé par un facteur variant entre 2.8 et 2.6, le minDCF étant quand à lui divisé par un facteur de 3.2 pour le GMM à 2.9 pour le GSL. Comparé au système GMM-UBM de référence (sans FA), l'écart est encore plus fort, l'EER passe en effet de 8.67% pour le système de référence à 1.62% pour le SFA en mode "oracle".

5. Les taux d'erreur remis en question

Comme l'a montré ce papier, les systèmes récents de RAL se montrent capables d'exploiter des quantités grandissantes de données d'apprentissage, que ce soit pour modéliser et réduire les facteurs de variabilité entre entraînement et test ou pour caractériser un locuteur. Cette capacité permet - alors que la tâche visée est réputée difficile - d'atteindre des taux d'égale erreur très faibles, d'environ 2.3%, et cela avec une marge de progression encore importante (un EER de 1.62% est atteint en mode oracle et des progrès semblent aisément réalisables si la quantité de données d'apprentissage augmente, en terme de normalisation de la session comme de modélisation du locuteur).

5.1. La RAL, est-ce un problème résolu ?

Au regard de ces résultats, il est légitime de se demander si le problème de la reconnaissance du locuteur peut être considéré comme résolu. En effet, si il suffit dorénavant d'augmenter les quantités de données disponibles pour améliorer les performances, est-il utile de poursuivre les efforts sur le coeur des systèmes de reconnaissance du locuteur ?

5.2. Un facteur de doute

Dans [6], les auteurs cherchent à transformer artificiellement la voix d'un locuteur pour tromper un système de RAL (faire que le système reconnaisse à tord ce locuteur comme étant une personne donnée), sans que la modification ne soit détectable de manière auditive. La transformation est réalisée sur des fenêtres temporelles courtes et porte uniquement sur les paramètres du conduit, en s'appuyant sur le modèle "source-filtre". L'expérience, réalisée sur la base NIST, est résumée par le tableau 5.2. La technique de transformation non audible de la voix permet de tromper le système de façon convainquante, le taux de fausse acceptation (taux d'imposteurs acceptés par le système) passant de moins de 1% à près de 50% ! Ce résultat permet de voir les améliorations récentes

Tab. 3: Effet de la transformation artificielle, non audible, de la voix des imposteurs.

	Système de référence	Avec transformation
EER %	8.54	35.41
minDCF(*100)	3.58	9.41
Fausse % acceptation %	**0.88**	**49.72**
Faux rejet %	27.45	27.45

obtenues en RAL sous un autre jour : avons nous réellement progressé dans le champs de la reconnaissance du locuteur ? En effet, si il est possible de transformer une voix et, ainsi, de tromper un système de RAL sans que la transformation ne soit audible, cela peut indiquer que les informations utilisées par le dit système ne sont pas aussi liées aux spécificités des locuteurs qu'espéré...
Ce constat ne remet pas en cause l'intérêt des travaux réalisés en RAL, il est en effet clair que la communauté scientifique a beaucoup progressé dans la prise en compte des facteurs de variabilité, ce qui reste un point clé pour la RAL. Il impose cependant de ne pas voir les taux d'erreur comme critère nécessaire et suffisant pour évaluer la qualité d'un système et, par la même, les progrès réalisés en reconnaissance du locuteur.

6. Une autre voie de recherche pour la reconnaissance du locuteur ?

L'objectif majeur de ce papier est d'attirer l'attention de la communauté scientifique sur le danger que représente l'utilisation des taux d'erreur comme seul critère d'évaluation des performances mais aussi d'évaluation des progrès réalisés et, par là-même, d'orientation des recherches.
Le problème est plus critique dans le cadre criminalistique que dans le cadre commercial. En effet, les applications commerciales définissent généralement un environnement bien délimité, avec des facteurs

de variabilité connus, ce qui autorise une simulation expérimentale relativement fiable. De plus, comme toute technologie de ce type, l'erreur a un coût commercial qui peut être compensé par des gains.

L'évolution de la RAL, par la focalisation sur la réduction des taux d'erreur, a progressivement éloigné la communauté scientifique concernée de ses bases théoriques, phonétiques et analytiques. Il semble cependant possible d'analyser finement les systèmes actuels pour comprendre leurs réactions, voir les prédire. Plusieurs études permettraient de mieux comprendre les phénomènes sous-jacents :

1. Analyser les performances des systèmes en fonction des informations phonétiques présentes en entrée, pour mieux définir quelles informations utilisent les systèmes. Ce point n'est pas très novateur, il s'agit plutôt de reprendre des études datant d'une dizaine d'année, mais relativement délaissées, l'intérêt des chercheurs s'étant concentré sur les performances des systèmes dans le cadre des évaluations.
2. Travailler sur des données contrôlées, voire simulées. Dans la lignée de l'approche d'analyse par (re)synthèse, il semble intéressant de partir d'une voix synthétique ou naturelle et de synthétiser des stimulis en variant un par un les différents paramètres (hauteur de la voix, position du triangle vocalique, position et variations des formants, prosodie...). Une analyse perceptive serait également intéressante.
3. L'évaluation des performances doit intégrer plus d'hétérogénéité, de variabilité et d'inconnu, dans les protocoles de test, de manière à assurer une meilleure généralisation des résultats. L'étude des performances lorsque les systèmes sont confrontés à des données inconnues, d'un type non inclus dans les données de développement devrait être systématisée. Si l'ajout de données enregistrées reste coûteux, d'autres voies complémentaires sont envisageables, en utilisant les techniques de transformation de voix mais aussi de synthèse, tout en n'oubliant pas qu'il est - au vu des taux d'erreur annoncés - nécessaire de changer l'ordre de grandeur des bases de données, en travaillant sur des bases mettant en oeuvre des milliers de locuteurs. Enfin, une évaluation des possibilités "pratiques" des systèmes serait intéressante, en utilisant des données provenant d'une application réelle.
4. Si les systèmes actuels obtiennent en moyenne des résultats très encourageants, ils montrent aussi des réactions surprenantes, comme des tests "imposteur" - en très faible nombre - obtenant des scores supérieurs à la moyenne des scores des clients. Ces phénomènes sont souvent intrinsèques à l'approche statistique utilisée et difficiles à détecter et à réduire. Pouvoir définir à partir du signal si suffisamment d'information est présente pour apprendre un modèle de locuteur (ou l'adapter) ou évaluer un score de reconnaissance sont des points essentiels pour limiter les conséquences de tels phénomènes. Quelques travaux sur des mesures de confiance ont été récemment proposés, voir par exemple [15] mais ce thème est encore insuffisamment traité.

Références

[1] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas. Score normalization for text-independent speaker verification systems. *Digital Signal Processing*, 10 :42–54, 2000.

[2] C. Barras, S. Meignier, and J.L. Gauvain. Unsupervised online adaptation for speaker verification over the telephone. In *Speaker Odyssey*, 2004.

[3] F. Bimbot, J.-F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-Garcia, D. Petrovsk-Delacrtaz, and D. Reynolds. A tutorial on text-independent speaker verification. *EURASIP Journal on Applied Signal Processing*, 4 :430–451, 2004.

[4] J.-F. Bonastre and al. Alize/spkdet : a state-of-the-art open source software for speaker recognition. In *Speaker Odyssey*, 2008.

[5] J.-F. Bonastre, F. Bimbot, L.-J. Boe, J.P. Campbell, D.A. Reynolds, and I. Magrin-Chagnolleau. Person authentication by voice : A need for caution. In *Eurospeech*, 2003.

[6] J.-F. Bonastre, D. Matrouf, and C. Fredouille. Artificial impostor voice transformation effects on false acceptance rates. In *Interspeech*, 2007.

[7] W. M. Campbell, D.E. Sturim, and D. A. Reynolds. Support vector machines using GMM supervectors for speaker verification. *IEEE Signal Processing Letters*, 13, May 2006.

[8] B. G. B. Fauve, D. Matrouf, N. Scheffer, J.-F. Bonastre, and J. S. D. Mason. State-of-the-art performance in text-independent speaker verification through open-source software. *IEEE Transactions on Audio, Speech and Language Processing*, 15(7) :1960–1968, September 2007.

[9] P. Kenny and P. Demouchel. Eigenvoices modeling with sparse training data. *IEEE trans*, 13 :345–354, 2005.

[10] A. Martin and M. Przybocki. The NIST speaker recognition evaluation series, National Institute of Standards and Technology's website, http ://www.nist.gov/speech/tests/spk.

[11] D. Matrouf, N. Scheffer, B. Fauve, and J.F. Bonastre. A straightforward and efficient implementation of the factor analysis model for speaker verification. Interspeech 2007.

[12] J. Pelecanos and S. Sridharan. Feature warping for robust speaker verification. In *Odyssey*, 2001.

[13] A. Preti, J.-F. Bonastre, D. Matrouf, F. Capman, and B. Ravera. Confidence measure based unsupervised target model adaptation for speaker verification. In *Interspeech*, 2007.

[14] D. Reynolds. Channel robust speaker verification via feature mapping. In *ICASSP*, 2003.

[15] J. Richiardi, P. Prodanov, and A. Drygajlo. Speaker verification with confidence and reliability measures. In *ICASSP*, 2006.

[16] Alex Solomonoff, William M. Campbell, and Ian Boardman. Advances in channel compensation for SVM speaker recognition. In *ICASSP*, 2005.

[17] DA van Leeuwen. Speaker adaptation in the nist speaker recognition evaluation 2004. In *Interspeech*, 2005.