Analyse des scores d'un Système de VAL GMM-UBM

Salah-eddine Mezaache [1,2], *Driss Matrouf* [1] *and Jean-François Bonastre* [1]

[1] 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
[2]Centre Universitaire de Bordj-Bou Arréridj, Algérie
mez.salah@gmail.com

ABSTRACT

In speaker recognition, the performance of a system are usually computed globally on a large set of tests, even if it is well known that subgroups of test could show a very different behavior than the complete set. In fact, a small subset of tests could represent the main part of the reported errors. In this work, we highlight a such subset of tests, where the impostors obtain some very high recognition scores. We evaluate if the problem comes from the envolved speakers, from the voice extracts or from the client model estimation technique. We also propose a strategy in order to minimize the effects of the observed phenomena on the overall performance of the system.

Keywords: Speaker verification, GMM-UBM, Reverse scoring

1. Introduction

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 NIST[1], ont été très impressionnants. En particulier, l'émergence de techniques capables de compenser les différences induites par l'usage de différents microphones et de différentes liaisons téléphoniques - comme le Latent Factor Analysis (FA) ou Nuisance Attribute Projection (NAP) - autorise un niveau de performance très attrayant. Le LIA a développé, en collaboration avec différents partenaires, un système basé sur l'approche UBM-GMM et le FA, à l'état de l'art [7]. Ce système, appelé ALIZE/SpkDet [5], est distribué sous forme de logiciel libre.
Cet article se concentre sur l'analyse des scores issus de ce système. Il met en évidence le fait que, même si les performances sont très encourageantes, certains tests imposteurs obtiennent des scores anormalement élevés. Ces tests ne représentent que $\simeq 0.7\%$ du total des tests, mais provoquent 40% des erreurs. Plusieurs hypothèses expliquant ce problème sont proposées et discutées. Enfin, une solution permettant de diminuer environ de moitié l'influence du problème sur les performances globales du système est proposée.
La section 2 de cet article est consacrée au contexte expérimental du travail, issu de NIST 2006. Le système de référence du LIA est également présenté. La section 3 met en évidence le problème visé dans ce travail : la présence d'un faible nombre de tests amenant la majorité des erreurs commises par le système. Les causes potentielles du problème sont également présentées et discutées. La méthode proposée pour diminuer l'influence de ces tests sur les performances globales du système est présentée dans la section 4. Enfin, la dernière section présente les conclusions issues de ce travail.

2. Contexte expérimental

Le protocole expérimental employé dans cet article est basé sur le contexte des évaluations NIST-SRE 2006[1]. Le protocole correspond à la tâche 1conv-1conv restreinte aux locuteurs masculins. Il comporte 354 locuteurs et 22122 tests, dont 1594 tests clients. Les performances sont évaluées de manière traditionnelle, par le taux d'égale erreur (EER pour Equal Error Rate), le minimum de la fonction de coût de NIST (minDCF) et des courbes DET (taux d'erreur de type I en fonction du taux d'erreur de type II).
Toutes les expériences présentées dans cet article ont été effectuées en utilisant le système de reconnaissance du locuteur développé par le LIA, ALIZE/SpkDet[2], diffusé en "logiciel libre" [4, 5]. Les détails concernant les paramètres du système peuvent être trouvés dans [7] (configuration utilisant la normalisation des écarts inter-sessions par Symetrical Factor Analysis [9]-SFA-, basée sur les travaux de [8]).
Ce système met en oeuvre une approche statistique par mélange de gaussiennes (GMM : Gaussian Mixture Model) [10]. Elle est dénotée GMM-UBM car cette modélisation nécessite l'utilisation d'un modèle générique appelé modèle du monde, ou UBM (pour Universal Background Model) [3].

3. Analyse des résultats

Le tableau 1 résume les performances du système de base en fonction de la normalisation des scores employée. La figure 1 montre l'histogramme des distributions des scores client et imposteur pour le système de base après une normalisation de type TNorm [2]. Cette figure montre qu'une proportion non négligeable des scores imposteurs obtient des valeurs trop élevées, souvent supérieures à la moyenne des scores obtenus par les clients du système. Nous nous concentrons sur ces tests imposteurs à haut score. Un sous-ensemble des tests imposteurs, dénommé "problémati-

[1]http ://www.nist.gov/speech/tests/spk/2006/sre-06_evalplan-v9.pdf

[2]http ://mistral.univ-avignon.fr/

Tab. 1: Résultats (DCFmin,EER%) du sytème de base NIST06

système	DCFmin(x100)	EER(%)
nonorm	2.25	5.08
Tnorm	2.03	4.64
Znorm	2.16	4.32
ZTnorm	2.06	3.95

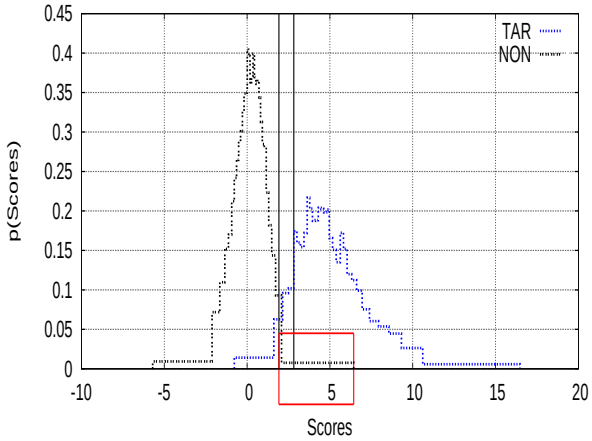

Fig. 1: Distribution des scores du système NIST06 en TNORM. *le rectangle en gras délimite la zone des tests problématiques*

que" est isolé, par sélection des tests imposteurs montrant des scores supérieurs à un seuil (θ). Dans notre analyse, (θ) est le seuil optimal qui minimise la fonction de coût DCFmin (le seuil est déterminé *a posteriori*).
Le tableau 2 décrit ces tests imposteur "problématiques" en fonction de la normalisation des scores employée. Sur les 22122 tests effectués, 20529 sont des tests imposteurs et entre 126 et 172 tests sont identifiés comme "problématiques", soit environ 0.7%. Ces tests sont constitués par un couple (segment de test, locuteur client). Ils impliquent dans la condition NONORM (sans normalisation des scores) seulement 118 segments de test différents et 72 locuteurs client différents. Il est intéressant de noter que, lors de nos expériences, le nombre de segments de test et de locuteurs client concernés restent quasiment constant quelque soit la méthode de normalisation des scores employée. Ce fait tend à démontrer que le problème est lié au moteur de reconnaissance du locuteur lui-même et non à un simple problème de normalisation des scores. Nous utiliserons la normalisation Tnorm dans le reste de cet article, celle-ci amenant la meilleur valeur en termes de minDCF. Pour tenter d'évaluer l'impact de ces scores problématiques en termes de performance, nous avons réalisé une expérience en mode "magicien d'oz", dans laquelle nous avons simplement retiré les tests problématiques (moins de 1% des tests imposteur, soit environ 0.7% du total des tests) avant de mesurer les performances du système. La figure 2 montre la courbe DET pour l'expérience complète et pour l'expérience dans laquelle ces tests sont supprimés. Les résultats montrent un gain relatif de 40% en

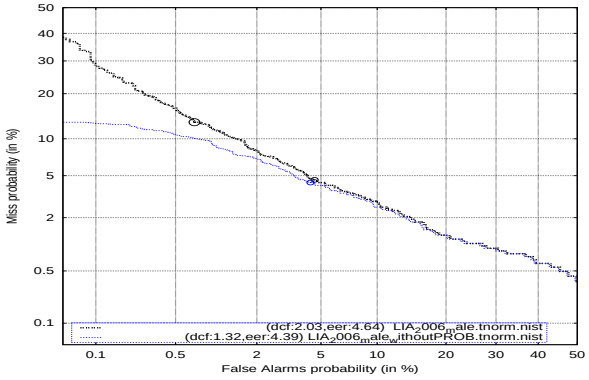

Fig. 2: Courbes DET. a : expérience de base ; b : expérience en supprimant les tests problématiques

DCFmin en éliminant seulement, environ, 0.7% des tests imposteur. Cet écart en terme de performance est significatif, d'autant qu'il intervient dans la zone de la courbe DET jugée primordiale dans le cadre des évaluations NIST. Différents facteurs peuvent expliquer la présence de ces tests "problématiques". Nous nous proposons d'analyser deux facteurs potentiels :

- Problème au niveau des modèles des locuteurs client. La quantité ou la qualité des données d'entraînement des modèles est insuffisante.
- Problème au niveau des segments de test. Le segment de test ne contient pas assez d'information (contenu phonétique pauvre ou niveau de bruit élevé).

Une première hypothèse consiste à estimer que la quantité d'information est insuffisante, soit au niveau de l'entraînement des modèles, soit au niveau des segments de test. Pour évaluer la pertinence de cette hypothèse, nous avons analysé la quantité de données (longueur des énoncés) des segments problématiques. Le tableau 3 présente la durée moyenne et l'écart type pour tous les énoncés d'entraînement et de test, de manière générale et de manière différenciée pour les segments appartenant aux tests problématiques. A part un petit nombre de fichiers d'entraînement, qui sont effectivement assez courts et qui appartiennent toujours au sous-ensemble "problématique", aucune différence significative n'est à signaler. Le facteur "durée" n'est visiblement pas une explication plausible du problème souligné dans ce travail.
Le deuxième point que nous abordons concerne la qualité des énoncés. La qualité d'un énoncé est difficile à quantifier, mais il est raisonnable d'avancer que le contenu linguistique (en terme de richesse phonétique) de l'énoncé et le rapport signal/bruit jouent un rôle important dans le contexte des GMM-UBM. Nous n'avons pas mené d'étude formelle sur ce fac-

Tab. 2: Nombre de tests problématiques en fonction de la normalisation des scores

	nonorm	Tnorm	Znorm	ZTnorm
Nbr de test	126	146	172	137
# modèles	69	72	93	74
# segments	105	130	127	109

Tab. 3: Moyenne et écart type des longueurs(nombre de trames) des segments de test et d'entraînement

	Moyenne	Écart type
total tests	7800	1900
tests problématiques	8100	900
total entraînement	7900	1200
entraînements problématiques	7900	1200

teur mais analysé auditivement les segments de test et d'entraînement concernés par le problème. Bien que quelques segments soient effectivement caractérisés par un contenu phonétiquement pauvre (par exemple, par des réponses de type enh, oueh, yah,...), ce facteur ne nous semble pas dominant.
En faisant référence à la ménagerie de Doddington [6], nous avons également cherché à déterminer si certains segments de test, i.e. certains locuteurs, présentaient des caractéristiques spécifiques, les rendant par exemple plus aptes à créer un faux positif. Pour cela, nous avons confronté le sous ensemble des segments de test "problématiques" à un large ensemble de modèles issus de la base de donnée NIST-2005 [3]. 1348 modèles de locuteurs masculins ont été utilisés lors de cette expérience, menant à 195460 tests imposteurs. L'analyse des résultats ne montre pas de signe caractéristique. En fait, la quasi-totalité des scores obtenus reste conforme avec la distribution classique des scores imposteurs. Il semblerait donc que les segments de test par eux-mêmes ne présentent pas des spécificités pouvant expliquer le phénomène étudié dans cet article. L'étude présentée reste cependant limitée : si nous avons utilisé un ensemble acceptable de modèles, ceux-ci restent issus d'un nombre limité de locuteurs, quelques centaines seulement.

4. Méthode REVERSE

Si l'analyse quantitative et qualitative des segments de parole impliqués dans le sous ensemble "problématique" ne permet pas de conclusions formelles, il reste néanmoins un facteur important, lié à ces caractéristiques, qui reste à explorer. Il est possible que nous nous trouvions en présence des limites intrinsèques de l'approche UBM-GMM : pour certains jeux de données, l'approche statistique n'est pas apte à modéliser les différences entre deux locuteurs. Comme il est connu dans la litérature que l'approche UBM-GMM est plus sensible au niveau de l'apprentissage des modèles qu'au niveau des tests, nous proposons d'inverser le processus normal de calcul des scores. Lors d'une procédure classique, les scores sont obtenus en calculant le rapport de vraisemblance suivant : $LLR(y, X) = LLK(y|X) - LLK(y|UBM)$. Où y représente les données de test, X le modèle du locuteur cible et $LLK()$ le logarithme de la vraisemblance. Un processus inverse, dénoté "REVERSE" dans cet article, consiste à apprendre un modèle Y à partir des données y (en utilisant MAP) et à calculer le LLR inversé : $LLR_{reverse}(y, X) = LLK(x|Y) - LLK(x|UBM)$, x étant les donnée d'apprentissage du locuteur considéré [1]. L'objectif visé par cette approche consiste à vérifier si le problème est lié aux données elles-mêmes ou à la façon dont l'UBM-GMM les prend en compte. La figure 3 présente la distribution des scores des tests problématiques avec la procédure de scoring classique et la scoring inversé Il est

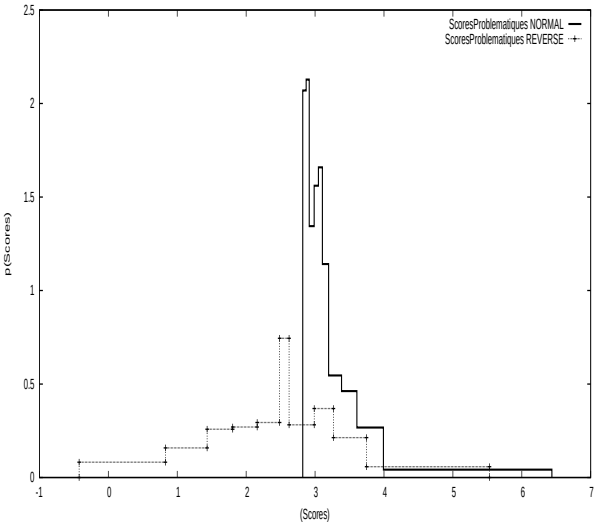

Fig. 3: Distribution des scores imposteurs problématiques en utilisant le scoring classique (NORMAL) et inversé (REVERSE)

aisé de remarquer qu'une large partie des scores situés au dessus du seuil de décision avec la procédure normale se trouvent au dessous avec la procédure inversée. Cette expérience tend à prouver que le problème se situe au niveau de la prise en compte des informations par l'UBM-GMM et non au niveau des données elles-mêmes.
Le phénomène remarqué dans le paragraphe précédent offre également une opportunité pour diminuer la perte de performance liée aux tests problématiques : appliquer un scoring inverse lorsque le scoring classique ne fonctionne pas. Pour valider cette solution, nous avons utilisé le scoring inverse seulement sur les tests problématiques, les autres tests étant scorés de façon classique. La figure 4 présente les résultats de cette expérience en mode "oracle". Un gain relatif de 25% apparait sur la valeur de DCFmin (rappelons que le gain maximum est estimé à environ 40%, cas dans lequel les tests sont simplement suprimés).

5. Fusion des scores

Dans la section précédente, nous avons montré que le scoring inversé appliqué aux tests problématiques amène un gain non négligeable en termes de DCFmin. Nous proposons deux solutions pour réaliser un système mettant en oeuvre ce principe : soit nous utilisons une simple fusion arithmétique (avec un poids égale à 0.5) des scores issus des deux procédés de scoring (normal et inversé), soit nous utilisons le score issu du scoring inversé dès lors que le scoring normal a proposé un score très élevé (le seuil de décision choisi

[3] http ://www.nist.gov/speech/tests/spk/2005/sre-05_evalplan-v5.pdf

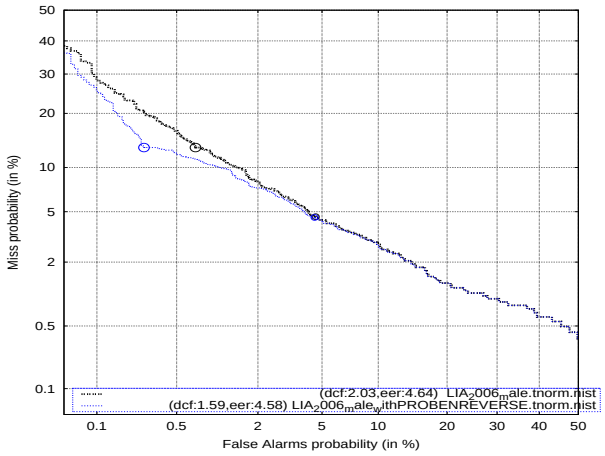

Fig. 4: Courbes DET du système NIST06 (TNORM), a : Système NORMAL ; b : système avec les tests problématiques en REVERSE

correspond au seuil menant au DCFmin).
Le tableau 4 présente les résultats obtenus par ces deux approches par fusion et pour les deux méthodes de scoring initiales. Le mode classique de scoring

Tab. 4: *Fusion des scorings NORMAL et REVERSE*

Système	DCFmin(x100)	EER(%)
NORMAL	2.03	4.64
REVERSE	2.19	5.26
FUSE	**1.81**	**4.20**
FUSE_THRESHOLD	2.13	5.07

montre des performances légèrement supérieures au mode inversé (4.64% d'EER pour le premier contre 5.26% pour le second). La fusion avec seuil n'apporte aucun gain. Le système de fusion arythmétique obtient un meilleur EER que l'oracle (4.2% à comparer à 4.64%). En DCFmin, cette fusion permet un gain relatif de 11% par rapport au système de base. Ce gain représente la moitié du gain potentiel mis en évidence par l'expérience en mode "oracle" (estimé à 22%).

6. Conclusion

Le travail présenté dans cet article s'appuie sur un système de reconnaissance du locuteur associant le paradigme GMM-UBM et une normalisation des écarts inter-session par le "factor analysis". Ce système a montré des performances à l'état de l'art. Néanmoins, certains résultats restent surprenants ; nous avons remarqué que 40% des erreurs commises par le système sont dues à un petit sous-ensemble des tests imposteurs, caractérisés par des scores très élevés. Nous avons proposé une analyse détaillée de ce phénomène, qui a permis de montrer que le problème provenait plus de la façon dont le système traitait les données que de caractéristiques liées aux données elles-mêmes. Nous avons exploité ce résultat pour améliorer les performances du système. En réalisant un scoring inversant l'usage des données d'entraînement et de test, fusionné avec la méthode classique, nous obtenons une réduction relative de la DCFmin de 11%, à rapprocher du gain potentiel optimal (amené par le scoring inversé), que nous avons évalué à 22% par une expérience en mode "oracle". Quelques gains complémentaires semblent facilement atteignables, avec un travail plus approfondi au niveau de la fusion des scores. Ce travail ouvre de nombreuses questions qui feront l'objet de futures recherches. Plus particulièrement, il nous semble intéressant d'étudier *pourquoi* le scoring inversé améliore les résultats, soit comprendre pourquoi l'apprentissage des modèles par MAP échoue dans certaines situations, alors que l'information contenue dans les données concernées semble suffisante.

Références

[1] M. Carrey and E. S. Parris. Cross validation in speaker recognition. In *Workshop on Speaker Recognition and its Commercial and Forensic Applications RLA2C*, pages 161–164, Avril 1998.

[2] Auckenthaler et al. Score normalization for text-independent speaker verification systems. *Digital Signal Processing*, 10 :42–54, 2000.

[3] Bimbot et al. A tutorial on text-independent speaker verification. *EURASIP Journal on Applied Signal Processing, Special issue on biometric signal processing*, 2004.

[4] Bonastre et al. Alize, a free toolkit for speaker recognition. In *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP 2005), Philadelphia, USA*, Philadelphia, USA, March 2005.

[5] Bonastre et al. Alize/spkdet : a state-of-the-art open source software for speaker recognition. In *Speaker Odyssey, South Africa, January 2008*, 2008.

[6] Doddington et al. Sheep, goats, lambs and wolves : a statistical analysis of speaker performance in the nist 1998 speaker recognition evaluation. In *Proceedings of International Conference on Spoken Language Processing (ICSLP 98)*, 1998.

[7] Fauve et al. State-of-the-Art Performance in Text-Independent Speaker Verification Through Open-Source Software. *IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING*, 5(7) :1960–1968, 2007.

[8] Kenny et al. Factor Analysis Simplified. In *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP 2005), Philadelphia, USA*, volume 1, 2005.

[9] Matrouf et al. A straightforward and efficient implementation of the factor analysis model for speaker verification. In *INTERSPEECH Conference, Antewerp, Belgium*, 2007.

[10] D. A. Reynolds. Speaker identification and verification using gaussian mixture speaker models. In *Speech Communication*, volume 171-2, pages 91–108, 1995.