AUGMENTATION DU TAUX DE FAUSSE ACCEPTATION PAR TRANSFORMATION INAUDIBLE DE LA VOIX DES IMPOSTEURS

Jean-François Bonastre, Driss Matrouf, Corinne Fredouille

LIA, Université d'Avignon
Agroparc, BP 1228
84911 Avignon CEDEX 9, France
{jean-francois.bonastre,driss.matrouf,corinne.fredouille}@univ-avignon.fr

ABSTRACT

This paper investigates the effect of a transfer function-based voice transformation on automatic speaker recognition system performance. We focus on increasing the impostor acceptance rate, by modifying the voice of an impostor in order to target a specific speaker. This paper is based on the following idea : in several applications and particularly in forensic situations, it is reasonable to think that some organizations have a knowledge on the speaker recognition method used and could impersonate a given, well known speaker. We also evaluate the effect of the voice transformation when the transformation is applied both on client and impostor trials.

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. Malgré une fiabilité largement inférieure comparée - par exemple - à l'iris, les progrès enregistrés dans les dernières décennies ont mené à l'émergence de systèmes automatiques utilisables dans des applications commerciales. Cependant, les nombreux facteurs de variabilité intervenant dans cette modalité, difficiles à contrôler et à prévoir, forment une limite aux performances des systèmes. Les différences dues au microphone, à l'environnement et aux différences inter-sessions sont les facteurs les plus présents dans la littérature, pour leur importance d'une part, mais également pour leur plus grande fréquence.
Durant la même période, dans le champ de la criminalistique, les juges, les avocats, les enquêteurs et les agences de sécurité nationale se sont montrés très demandeurs de techniques permettant d'identifier un individu par sa voix, pour confondre un suspect ou pour servir d'élément de preuve dans un tribunal [1][2]. En dépit du fait que la communauté scientifique, dans une large majorité, ait remis en cause les bases scientifiques d'une telle identification vocale [3][4][5] et du message fort de "besoin de précaution" émis par [6], les techniques d'identification vocale sont couramment utilisées dans la pratique criminalistique, particulièrement dans le contexte d'événements terroristes à l'échelle mondiale, avec l'utilisation de plus en plus fréquente de systèmes automatiques.
Ce papier vise à identifier une des limites des systèmes automatiques, ou plutôt de leur utilisation dans ce contexte spécifique : si vous connaissez la technique de reconnaissance utilisée pour identifier une voix, si vous avez également un exemple de la voix d'une personne X, est-il possible de transformer la voix d'une personne Y de telle façon que le système conclut à une identité entre X et Y, sans que cette transformation ne soit audible ?
Les objectifs visés dans ce papier sont proches de l'approche "voice-forgery" proposée dans [17], la différence principale réside dans l'interprétation de la transformation : dans notre cas, le système de reconnaissance doit être trompé quand le but est de synthétiser une voix proche de X au sens perceptuel pour la "voice-forgery". Ce papier étend les travaux préliminaires présentés dans [11][12].

La section 2 présente l'approche statistique couramment utilisée en reconnaissance du locuteur et dans laquelle s'inscrit ce travail. La section 3 décrit la technique de transformation de voix que nous utilisons. La section 4 décrit le protocole expérimental utilisé et les résultats sont présentés dans la section 5. Enfin, la section 6 conclut ce travail et ouvre quelques perspectives.

2. 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 [10]. Etant donné un segment de parole Y et un locuteur S, la tâche de vérification du locuteur consiste à déterminer si Y a été prononcé par S. Cette prise de décision est modélisée par l'estimation d'un rapport de deux probabilités : Y provient de S ($H0$) et Y a été prononcé par un inconnu ($H1$). Ce rapport (LR, pour Likelihood Ratio) est comparé à un seuil de décision θ ; $H0$ est désignée si le ratio est supérieur au seuil, $H1$ dans le cas contraire. De manière plus formelle, le ratio LR est donné par :

$$LR(Y, H0, H1) = \frac{p(Y|H0)}{p(Y|H1)} \quad (1)$$

avec Y, le segment de parole à tester, $p(Y|H0)$, la vraisemblance de $H0$, $p(Y|H1)$ la vraisemblance de $H1$ et θ, le seuil de décision.
Un modèle, λ_{hyp} représente l'hypothèse $H0$; ce modèle est appris à partir d'un exemple de la voix du locuteur concerné. Le modèle $\lambda_{\overline{hyp}}$ représente la seconde hypothèse, $H1$, et est généralement appris à partir d'une collection d'extraits vocaux provenant d'un grand ensemble de locuteurs.
Le ratio LR devient $\frac{p(Y|\lambda_{hyp})}{p(Y|\lambda_{\overline{hyp}})}$. 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

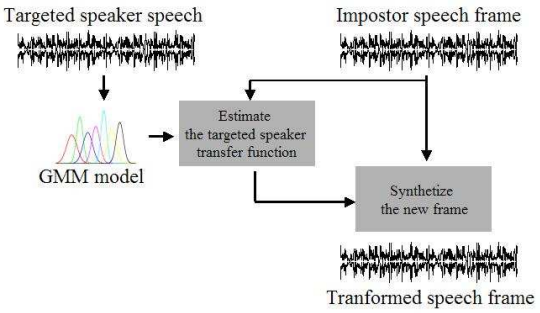

Fig. 1. Principe de la transformation, pour une trame de signal.

du mélange. Le modèle $\lambda_{\overline{hyp}}$ est dénoté modèle du monde, ou UBM quand il est indépendant de l'environnement et est estimé en maximisant la vraisemblance des données d'apprentissage correspondante. Le modèle λ_{hyp} est dérivé de l'UBM, par une technique de maximisation de la probabilité $aposteriori$ (MAP). De manière courante, seules les moyennes des composantes sont adaptées et les autres paramètres sont directement issus de l'UBM [13].

3. TRANSFORMATION DES VOIX

L'objectif de la transformation de voix présentée dans cette section est d'augmenter la vraisemblance d'un signal Y étant donné un locuteur X, ou plus exactement étant donné le GMM représentant ce locuteur. La voix transformée doit conserver un aspect naturel pour un auditeur.

Le principe retenu pour cette transformation consiste à analyser le signal Y trame à trame en utilisant un modèle source-filtre et à modifier la fonction de transfert pour la rapprocher du locuteur cible. Celui-ci est modélisé par un modèle GMM, appris sur un extrait de parole lui appartenant ; ce modèle servant à la transformation du modèle du locuteur cible. En fait, la fonction de transfert initiale est remplacée par une fonction de transfert estimée par une moyenne arithmétique des moyennes des différentes composantes du modèle cible, pondérées par leur probabilité $aposteriori$ pour la trame considérée. Le signal est enfin reconstitué par une simple approche "overlap-add". La figure 1 présente une vue générale du procédé de transformation pour une trame du signal Y.

En réalité, le procédé est un peu plus complexe, par le fait que chaque système de reconnaissance du locuteur utilise une représentation spécifique des données acoustiques, issue d'une étape de calcul de coefficients acoustiques suivie de différents étages de normalisation (de manière classique : sélection des trames de parole puis centrage et réduction des données sélectionnées) rendant parfois impossible l'estimation de la fonction de transfert cible. Pour contrer ce problème, nous avons appris deux modèles pour le locuteur ciblé : un modèle en respectant la représentation des données du système de reconnaissance du locuteur et un modèle utilisant une représentation des données plus simple, propre au système de transformation de voix. Ces deux modèles sont appris en parallèle, le premier guidant le second (i.e. les probabilités $aposteriori$ pour qu'une donnée soit liée à une composante du modèle - utilisées au sein de l'algorithme d'estimation EM/MAP - sont issues du premier modèle).

Soit y, une trame du signal à transformer, provenant d'un imposteur S' et x, la trame correspondante, appartenant au locuteur cible, S . Le modèle source-filtre amène :

$$Y(f) = H_y(f)S_y(f) \tag{3}$$
$$X(f) = H_x(f)S_x(f) \tag{4}$$

avec Y et X, les représentations spectrales de y and x, H_y et H_x, les fonctions de transfert correspondant respectivement à y et x, S_x et S_y, les transformées de Fourier du signal de la source pour x et y (il faut noter que le spectre n'est rien d'autre qu'une représentation compacte de la fonction de transfert). Pour rapprocher y de x - en termes de forme spectrale - il est suffisant de remplacer H_y par H_x dans l'équation 3 :

$$Y'(f) = H_x(f)S_y(f) = \frac{H_x(f)}{H_y(f)}Y(f) \tag{5}$$

Pour réaliser cela, si nous décidons de ne pas modifier la phase du signal original, nous appliquons le filtre suivant au signal y :

$$H_{yx}(f) = \frac{|H_x(f)|}{|H_y(f)|} \tag{6}$$

Dans ce processus de filtrage, pour chaque trame y (imposteur) nous avons besoin de connaitre H_x (et non x voir equation 6). Pour des commodités statistiques, nous préférons estimer une version LPCC de H_x que nous dénommons $Hlpcc_x$ (le passage de l'une à l'autre est trivial [12]) : $Hlpcc_x = Sigma_g[P(g|y) * m_{g,lpcc}]$. Les probabilités $P(g|y)$ sont calculées en utilisant le gmm cepstral (dans le domaine de représentation des données du système de verification du locuteur) du client. Les $m_{g,lpcc}$ sont les moyennes des composantes du modèle de ce client dans le domaine LPCC (la correspondance entre les composantes des deux modèles est maintenue lors de l'apprentissage de ceux-ci).

4. PROTOCOLE EXPÉRIMENTAL

Les expériences permettant de valider l'approche choisie dans ce papier ont été réalisées sur la base du protocole de la campagne d'évaluation NIST-SRE 2005 [14].

Le corpus utilisé correspond au corpus de NIST-SRE 2005, réduit à la partie "homme" de la tâche "one-conv/one-conv". Les messages vocaux utilisés pour l'apprentissage des modèles des clients et pour les tests sont d'une durée moyenne de 2mn30 de parole téléphonique et conversationnelle. Le protocole comporte 1231 tests "clients" et 12317 tests "imposteurs". Le modèle UBM ainsi que les données nécessaires à la normalisation des scores sont issus des corpus des campagnes SRE des années 2002 à 2004.

Trois expériences ont été réalisées :

- Une expérience de calibration, réalisée sans utiliser la transformation de voix (baseline).
- Une expérience où la transformation de voix est appliquée pour chaque test "imposteur", avec la connaissance du message vocal utilisé pour entraîner le modèle du locuteur ciblé (expérience 1).
- Une expérience où la transformation de voix est appliquée pour chaque test "imposteur", en utilisant un message vocal venant du locuteur ciblé, mais différent du message utilisé

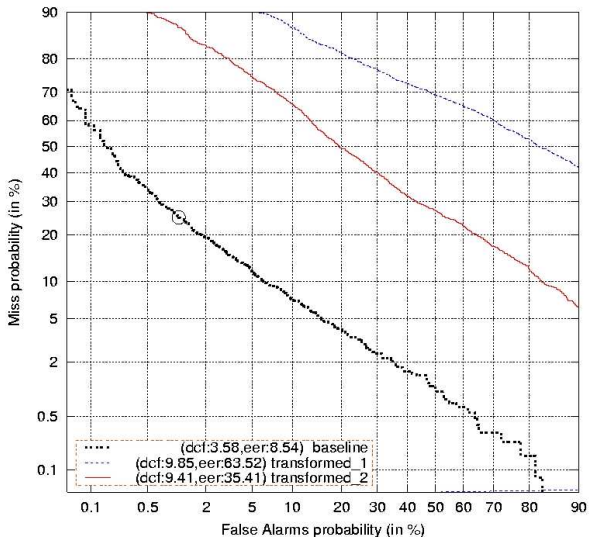

Fig. 2. Courbes DET pour le baseline, l'expérience 1 (utilisant les données d'entraînement pour la transformation des voix) et l'expérience 2 (utilisant un enregistrement différent pour la transformation de voix).

pour l'apprentissage, excepté pour un faible nombre de locuteurs, pour lesquels un seul enregistrement est disponible (expérience 2).

Pour toutes les expériences, le même modèle du monde a été utilisé, pour le procédé de transformation de voix comme pour les tests de reconnaissance.

Nous avons utilisé le système LIA_SpkDet [15], développé au LIA sur la base du toolkit ALIZE [18][16]. L'ensemble des logiciels utilisés est disponible sous forme de logiciels libres.
Le système LIA_SpkDet est basé sur l'approche UBM-GMM et implémente une normalisation des scores de type TNORM. La paramétrisation acoustique correspond à 16 coefficients cepstraux LFCC augmentés des 16 dérivées premières. Une sélection des vecteurs acoustiques, basée sur une modélisation multi-gaussienne de l'énergie est réalisée, avant de centrer et de réduire les coefficients. Le modèle UBM ainsi que les modèles de locuteurs comportent 2048 composantes. Durant les tests, les 10 meilleures composantes sont sélectionnées et utilisées.

5. RÉSULTATS

L'influence du procédé de transformation de voix est mesurée en termes de courbes DET, présentant les taux d'erreur de type I en fonction du taux d'erreur de type II, et à travers les distributions des scores imposteurs. La figure 2 présente les courbes DET pour le baseline, l'expérience 1, où les voix des imposteurs ont été transformées en utilisant les données d'apprentissage des locuteurs ciblés et l'expérience 2, pour laquelle des données différentes ont été employées pendant la transformation. La normalisation TNORM a été employée dans tous les cas. La figure 3 montre les distributions des scores imposteurs pour les trois expériences. Les figures 4 et 5 montrent les distributions des scores clients et imposteurs du

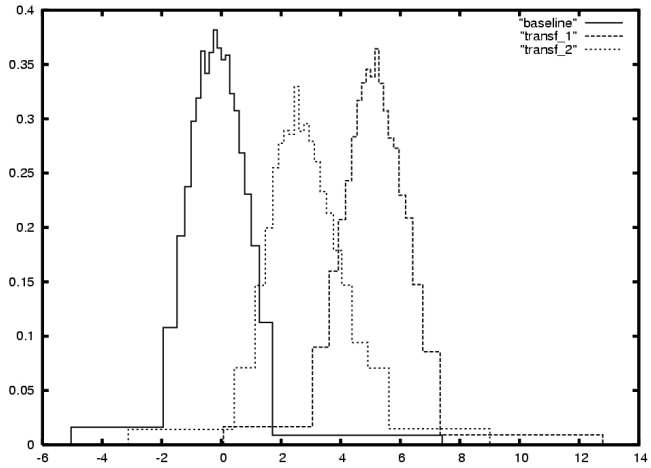

Fig. 3. Distribution des scores imposteurs pour le baseline (baseline), l'expérience 1 (transf_1) et l'expérience 2 (transf_2)

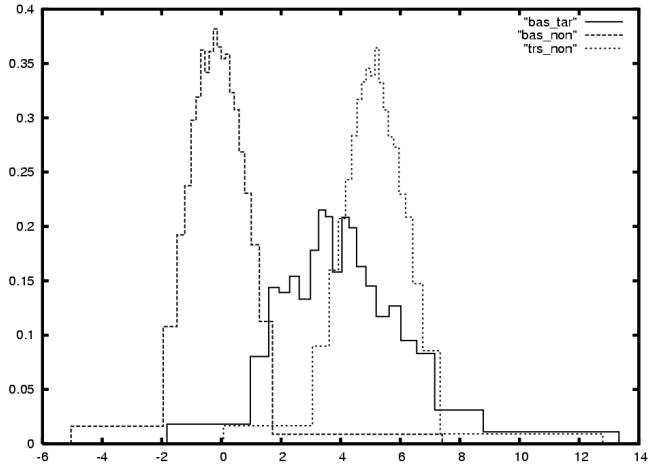

Fig. 4. Distributions des scores clients pour le baseline (bas_tar) et des scores imposteurs pour le baseline (bas_non) et pour l'expérience 1 (trs_non).

baseline, comparées respectivement à la distribution imposteur de l'expérience 1 et de l'expérience 2.
Un important déplacement de la distribution des scores imposteurs est clairement mis en évidence pour les deux expériences où la transformation de voix est appliquée, comparé au baseline. Le même effet est observé sur les courbes DET, qui montrent une très forte dégradation des performances. Bien entendu, l'expérience 1, qui utilise directement les données d'apprentissage pour la transformation, présente une dégradation plus importante que l'expérience 2, plus réaliste, dans laquelle un enregistrement différent de celui de l'apprentissage est utilisé.

Pour mettre en évidence l'évolution du taux de fausses alarmes, nous proposons dans la table 1 les taux de faux rejet et de fausse alarme pour un seuil $apriori$, fixé empiriquement sur un ensemble de tests différent. L'influence de la transformation de voix sur le taux de fausse alarme est clairement mise en évidence.

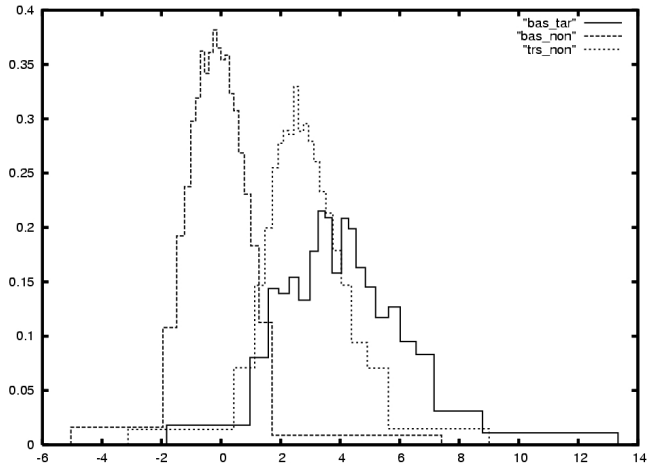

Fig. 5. Distributions des scores clients pour le baseline (bas_tar) et des scores imposteurs pour le baseline (bas_non) et pour l'expérience 2 (trs_non).

	False Alarm (%)	Miss Probability (%)
Baseline	0.88	27.45
Exp 1	96.55	27.45
Exp 2	49.72	27.45

Table 1. Taux de "False Alarm" et de "Miss Probability" en utilisant un seuil défini $apriori$, pour les 3 expériences.

6. CONCLUSION ET PERSPECTIVES

Dans ce papier, nous avons évalué les effets d'une transformation artificielle de la voix sur les taux de fausse alarme d'un système de reconnaissance du locuteur. L'objectif principal était de vérifier si il était possible de tromper un tel système lorsqu'on possède un exemple de la voix d'un locuteur cible et des connaissances sur le système utilisé. Sous ces hypothèses, qui semblent tout à fait réalistes par exemple dans le contexte d'événements terroristes de grande ampleur, un procédé très simple de transformation de la voix a permis, sans dénaturer la voix, de perturber très sensiblement le système de reconnaissance utilisé. L'approche et les résultats proposés mettent en lumière une des limites des approches utilisées en reconnaissance automatique du locuteur, dans le cadre criminalistique.
Dans cette étude, une large connaissance du système de reconnaissance du locuteur employé était utilisée pour la transformation de voix : technique de base, paramétrisation et modèle du monde notamment. Une première continuation de ce travail consistera donc à mesurer l'influence des différents éléments cités précédemment. Des améliorations du procédé de transformation sont également envisagées, en lissant les transformations par exemple.

7. REFERENCES

[1] R.H. Bolt, F.S. Cooper, D.M. Green, S.L. Hamlet, J.G. McKnight, J.M. Pickett, O. Tosi, B.D. Underwood, D.L. Hogan, "On the Theory and Practice of Voice Identification", *National Research Council, National Academy of Sciences, Washington, D.C.*, 1979.

[2] O. Tosi, "Voice Identification : Theory and Legal Applications",*University Park Press : Baltimore, Maryland*, 1979.

[3] R.H. Bolt, F.S. Cooper, E.E.Jr. David, P.B. Denes, J.M. Pickett, K.N. Stevens, "Speaker Identification by Speech Spectrograms : A Scientists' View of its Reliability for Legal Purposes",*Journal of the Acoustical Society of America*, 47, 2 (2), 597-612, 1970.

[4] J.F. Nolan, "The Phonetic Bases of Speaker Recognition",*Cambridge University Press : Cambridge*, 1983.

[5] L.J. Boë, "Forensic voice identification in France",*Speech Communication, Elsevier*, Volume 31, Issues 2-3, June 2000, pp. 205-224 (http ://dx.doi.org/10.1016/S0167-6393(99)00079-5).

[6] J.-F. Bonastre, F. Bimbot, L.-J. Boe, J.P. Campbell, D.A. Reynolds, I. Magrin-Chagnolleau, "Person Authentication by Voice : A Need for Caution", *Proceeding of Eurospeech 2003*, 2003

[7] C. Champod, D. Meuwly, "The inference of identity in forensic speaker recognition",*Speech Communication*, Vol. 31, 2-3, pp 193-203, 2000

[8] J. González-Rodríguez, J. Ortega, and J.J. Lucena, "On the Application of the Bayesian Framework to Real Forensic Conditions with GMM-based Systems", *Proc. Odyssey 2001 Speaker Recognition Workshop*, pp. 135-138, Crete (Greece), 2001

[9] P. Rose, T. Osanai, Y. Kinoshita, "Strength of Forensic Speaker Identification Evidence - Multispeaker formant and cepstrum based segmental discrimination with a Bayesian Likelihood ratio as threshold", *Speech Language and the Law*, 2003 ; 10/2 : 179-202.

[10] F. Bimbot, J.-F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-Garcia, D. Petrovska, D. A. Reynolds, "A tutorial on text-independent speaker verification", *EURASIP Journal on Applied Signal Processing*, 2004, Vol.4, pp.430-451

[11] D. Matrouf, J.-F. Bonastre and J.P. Costa, "Effect of impostor speech transformation on automatic speaker recognition", *proc. of COST 275 Workshop "Biometric on the internet"*, Hatfied, UK, 2005

[12] D. Matrouf, J.-F. Bonastre and C. Fredouille, "Effect of voice transformation on impostor acceptance", *Proc. of ICASSP 2006*, Toulouse, France, 2006

[13] D. A. Reynolds, T. F. Quatieri, R. B. Dunn, "Speaker verification using adapted Gaussian mixture models", *Digital Signal Processing (DSP), a review journal Special issue on NIST 1999 speaker recognition workshop*, vol. 10(1-3), pp 19-41, 2000.

[14] NIST Speaker Recognition Evaluation campaigns web site, http ://www.nist.gov/speech/tests/spk/index.htm

[15] LIA_SpkDet system web site, http ://www.lia.univ-avignon.fr/heberges/ALIZE/LIA_RAL

[16] J.-F. Bonastre, F. Wils, S. Meignier, "ALIZE, a free toolkit for speaker recognition", *Proceedings of ICASSP05*, Philadelphia (USA), 2005

[17] P. Perrot, G. Aversano, R. Blouet, M. Charbit, G. Chollet, "Voice Forgery Using ALISP : Indexation in a Client Memory", *Proceedings of ICASSP05*, Philadelphia (USA), 2005

[18] ALIZE project web site, http ://www.lia.univ-avignon.fr/heberges/ALIZE/