Analyse en Composante Principale pour l'extraction des i-vecteurs en vérification du locuteur

Anthony Larcher[1] Pierre-Michel Bousquet[2] Driss matrouf[2] Jean-Francois Bonastre[2]
(1) Institute for Infocomm Research, A*Star, Singapour
(2) Université d'Avignon - -CERI - LIA
alarcher@i2r.a-star.edu.sg

RÉSUMÉ

Nous proposons une alternative aux méthodes état-de-l'art développées récemment pour la vérification du locuteur dans le cadre du paradigme de Variabilité Totale. Les expériences présentées montrent que l'utilisation de l'Analyse en Composante Principale (ACP) en remplacement du Factor Analysis (FA), pour la réduction de dimension des super-vecteurs, peut amener a des performances équivalentes. Ainsi l'extraction des *i*-vecteurs selon le critère du Maximum de Vraisemblance en utilisant la matrice des vecteurs propres obtenus par une ACP permet de surpasser un système état-de-l'art combinant Factor Analysis et Analyse Discriminante Linéaire Probabiliste dans 3 des 8 conditions de NIST-SRE08. Nous montrons également que des *i*-vecteurs obtenus par simple projection orthogonale sur la matrice produite par ACP peuvent surpasser l'approche état-de-l'art dans deux des conditions évaluées. Les résultats présentés dans cet article illustrent le potentiel d'une approche déterministe pour la vérification du locuteur.

ABSTRACT

Principal Component Analysis for i-vector extraction in speaker verification.

In this work, we propose alternative algorithmic combinations for speaker verification based on the Total Variability paradigm. Experiments presented in this paper show that replacing Factor Analysis (FA) by a Principal Component Analysis (PCA) for super-vector dimensionality reduction can lead to state-of-the-art performance. Extracting the *i*-vectors according to the Maximum Likelihood criteria when using an Eigen Vector matrix resulting from a PCA outperforms a state-of-the-art system based on Factor Analysis and Probabilistic Linear Discriminant Analysis in 3 conditions of the NIST-SRE08 evaluation over 8. Computation of *i*-vectors by an orthognonal projection on the PCA matrix is also shown to outperform the state-of-the-art configuration in 2 of the 8 conditions. The results presented in this paper illustrate the potential of a Deterministic approach for speaker verification.

MOTS-CLÉS : Vérification du locuteur, I-vecteurs, Réduction de dimension.

KEYWORDS: Speaker verification, I-vectors, Dimension reduction.

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 1: JEP, pages 297–304,
Grenoble, 4 au 8 juin 2012.

1 Introduction

La vérification automatique du locuteur est le procédé biométrique qui consiste à authentifier une personne en utilisant l'information portée par sa voix. Durant une phase appelée enrôlement, un utilisateur fournit au système un échantillon de sa voix qui sera utilisé pour des comparaisons ultérieures. Durant une seconde phase, appelée test, les systèmes état-de-l'art, qui reposent sur

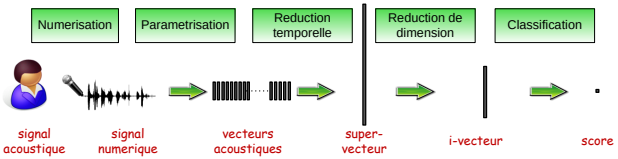

FIGURE 1 – Structure d'un système de vérification du locuteur suivant le paradigme de l'espace de Variabilité Totale.

le paradigme de l'espace de Totale Variabilité introduit par (Dehak *et al.*, 2011), suivent une structure en cinq étapes décrite par la Figure 1.

1. Numérisation du signal acoustique ;
2. Paramétrisation : les informations utiles sont extraites du flux de données sous la forme d'une série de vecteurs de durée variable appelés paramètres acoustiques ;
3. Réduction temporelle : la série de paramètres acoustiques est transformée en un super-vecteur dont la dimension élevée est indépendante de la longueur de la série de paramètres acoustiques ;
4. Réduction de dimension : la dimension du super-vecteur est réduite, le vecteur résultant de cette étape est appelé *i*-vecteur.
5. Classification : l'*i*-vecteur correspondant au segment de parole à évaluer est comparé à l'*i*-vecteur obtenu selon le même procédé à partir des données d'enrôlement de l'utilisateur. La comparaison de ces vecteurs produit un score qui permet de décider si oui ou non l'utilisateur du système correspond a l'identité qu'il clame.

Dans cet article, nous traiterons principalement des étapes de réduction de dimension et de classification.

Une grande majorité des systèmes proposés récemment dans la littérature réduisent la dimension des super-vecteurs grâce au Factor Analysis (FA) dont une description est donnée dans (Dehak *et al.*, 2011). Des travaux récents (Garcia-Romero et Espy-Wilson, 2011) ont montré que l'Analyse Discriminante Linéaire Probabiliste (ADLP) (Prince et Elder, 2007) obtenait d'excellentes performance lorsqu'elle était appliquée aux *i*-vecteurs obtenus par FA. Le but de nos travaux est de montrer qu'il est possible d'obtenir des performances équivalentes aux systèmes état-de-l'art en remplaçant le FA et l'ADLP par d'autres méthodes, notamment déterministes.

Dans la suite de cet article nous proposons d'utiliser l'Analyse en Composantes Principales (ACP) comme alternative au Factor Analysis pour l'extraction des *i*-vecteurs. L'Analyse Discriminante Linéaire Probabiliste est elle remplacée par un Radial-NAP suivi d'une distance de Mahalanobis proposées par (Bousquet *et al.*, 2011). Dans une quatrième partie, nous proposons quatre systèmes reposant sur l'ACP et la distance de Mahalanobis. Les performances de ces systèmes sont discutées dans une cinquième partie. Enfin nous présentons les conclusions et les perspectives issues de ce travail.

2 Réduction de dimension

La vérification du locuteur peut être réalisée directement dans l'espace des super-vecteurs (Wan et Campbell, 2000). Cependant, de meilleurs performances peuvent être obtenues en effectuant la classification dans un espace de dimension réduite (Kenny *et al.*, 2005; Dehak *et al.*, 2011). Parmi les méthodes usuelles de réduction de dimension, certaines comme le Factor Analysis sont non-discriminantes : l'Analyse en Composantes Principales (ACP) (Yaman *et al.*, 2011), l'Analyse en Composantes Indépendantes (ACI) (Hyvarinen et Oja, 2000), l'Analyse en Composantes Principales Probabiliste (ACPP) (Scheffer *et al.*, 2011) et d'autres au contraire cherchent des sous-espaces adaptés a la tâche considérée comme l'Analyse Discriminante Linéaire (ADL)(Dehak *et al.*, 2011), l'Analyse Discriminante Linéaire Probabiliste (ADLP) (Prince et Elder, 2007) ou la Régression PLS (Srinivasan *et al.*, 2011). Certaines de ces méthodes ne peuvent être utilisées en grande dimension pour des raisons calculatoires ou par manque de données comme dans le cas des méthodes discriminantes qui nécessitent pour la plupart un nombre de locuteurs trop important. Mais ces méthodes peuvent également être combinées afin d'optimiser la réduction de dimension comme dans (Dehak *et al.*, 2011; Senoussaoui *et al.*, 2011).
La comparaison de toutes ces méthodes dépasse le cadre de cet article et nous choisissons ici de remplacer le FA par une ACP. Dans (Tipping et Bishop, 1999), les auteurs montrent qu'il existe un lien entre Factor Analysis et Analyse en Composante Principale (Tipping et Bishop, 1999) puisque le maximum de vraisemblance du FA est obtenu lorsque les vecteurs propres de la matrice T sont les axes principaux obtenus par l'ACP.

2.1 Factor Analysis

Le développement des voix propres (Kuhn *et al.*, 1998) puis du Joint Factor Analysis (Kenny *et al.*, 2005) ont fait du Factor Analysis la méthode de réduction de dimension de référence en vérification du locuteur(Dehak *et al.*, 2011). Une représentation compacte $\mathbf{w}$, appelée *i*-vecteur, du segment de parole est obtenue en suivant un modèle linéaire Gaussien d'après lequel le super-vecteur $\mathscr{S}$ peut être décomposé selon l'équation 1

$$\mathscr{S} = \mathbf{m} + \mathbf{T} \cdot \mathbf{w} \tag{1}$$

où $\mathbf{m}$ est le super-vecteur du modèle UBM, $\mathbf{T}$ est une matrice rectangulaire et $\mathbf{w}$ est définit comme suivant une loi normale standard. La projection $\mathbf{w}$ d'un super-vecteur $\mathscr{S}$ est obtenue par :

$$\mathbf{w} = (\mathbf{I} + \mathbf{T}^t \mathbf{\Sigma}^{-1} \mathbf{N} \mathbf{T})^{-1} \mathbf{T}^t \mathbf{\Sigma}^{-1} \mathbf{F} \tag{2}$$

où N et F sont respectivement les statistiques d'ordre 0 et 1 du segment de parole calculés sur le modèle du monde, $\mathbf{\Sigma}$ est la matrice de covariance du modèle du monde et $\mathbf{I}$ est la matrice identitée.

2.2 Analyse en Composante Principale

L'ACP est un procédé déterministe largement utilisé pour la représentation de données et la réduction de dimension (Jolliffe, 2002). Pour la réduction de dimension, l'ACP consiste a trouver une base orthonormale d'un sous-espace, appelé espace propre, qui maximise la variance des

données après projection. Cette projection minimise également l'erreur quadratique moyenne. Étant donne un ensemble de super-vecteurs de dimension M, et de matrice de covariance C, la matrice C peut s'écrire :

$$C = \mathbf{QDQ}^t \tag{3}$$

où **D** est une matrice diagonale dont les termes, appelés valeurs propres, sont rangés par ordre décroissant et **Q** est la matrice des vecteurs propres de C rangés en colonnes.

La matrice de projection dans l'espace propre de rang k est la matrice rectangulaire **P** de dimension $k \times M$ dont les lignes sont les k vecteurs propres de C correspondant aux k plus grandes valeurs propres de **Q**. La projection **w** d'un super-vecteur $\mathscr{S}$ est obtenue par :

$$\mathbf{w} = \mathbf{P} \cdot \mathscr{S} \tag{4}$$

3 Classification

Nous décrivons dans cet article deux méthodes de classification ayant montré de bonnes performances dans la littérature (Larcher *et al.*, 2012; Kenny, 2010). Ce choix ne présage cependant pas des possibilités d'utiliser d'autres approches (Dehak *et al.*, 2011).

3.1 Analyse Linéaire Discriminante Probabiliste

L'ALDP (Prince, 2012) considère qu'un i-vecteur **w** est généré par un modèle discriminant de la forme :

$$\mathbf{w} = \mathbf{Vy} + \mathbf{Ux} + \mathbf{z} \tag{5}$$

où **y** représente la composante locuteur, **x** la composante canal et **z** le résidu suivant une loi normale. Les matrices **U** et **V** imposent aux sous-espaces locuteur et canal une dimension inférieure a celle des **i**-vecteurs. Le score utilise pour la classification est un rapport de vraisemblance calculé entre l'hypothèse : H_1 : les deux i-vecteurs sont générés par un même locuteur et son hypothèse complémentaire. Dans cet article, la dimension des sous-espaces locuteur et canal est identique à celle des i-vecteurs.

3.2 Radial-NAP et distance de Mahalanobis

Le Radial-NAP est une méthode de compensation de la variabilité inter-sessions proposée par (Bousquet *et al.*, 2011). La matrice de covariance intra-classe $\mathbf{W}_i$ est estimée sur les données de développement comme suit :

$$\mathbf{W}_i = \frac{1}{n} \sum_{s=1}^{S} \sum_{i=1}^{n_s} (\mathbf{w}_i^s - \overline{\mathbf{w}_s})(\mathbf{w}_i^s - \overline{\mathbf{w}_s})^t \tag{6}$$

où s est le nombre de locuteurs dans les données de développement, n_s est le nombre de sessions appartenant a chaque locuteur et $\overline{\mathbf{w}_s}$ est leur moyenne.

$\mathbf{w}'$ est le projeté orthogonal de l'i-vecteur sur la matrice des k premiers vecteurs propres de la matrice $\mathbf{W}_i$. Un i-vecteur $\mathbf{w}''$ est alors modifie selon l'équation :

$$\mathbf{w}'' = \frac{w - w'}{||w - w'||} \tag{7}$$

Durant la phase de test, après application du Radial-NAP, la distance entre deux i-vecteurs $\mathbf{w}_1$ et $\mathbf{w}_2$ est calculée d'après l'équation 8.

$$score(\mathbf{w}_1, \mathbf{w}_2) = (\mathbf{w}_1 - \mathbf{w}_2)^t \mathbf{W}^{-1} (\mathbf{w}_1 - \mathbf{w}_2) \tag{8}$$

où $\mathbf{W}$ est la matrice de covariance intra-classe estimée sur les données de développement.

Remarque

Avant de procéder a la classification par ADLP ou avec la distance de Mahalanobis, les i-vecteurs sont centrés réduits et normés. En effet, (Bousquet *et al.*, 2011) et (Garcia-Romero et Espy-Wilson, 2011) ont montré que ce traitement permettait d'améliorer les performances des classifieurs. Les paramètres utilisés pour centrer réduire sont estimés sur les données de développement.

4 Systèmes de vérification du locuteur

Le nombre important de méthodes existant dans la littérature ne nous permet pas de comparer toutes les combinaisons possibles. Nous présentons ici cinq systèmes retenus car ils présentent des variations progressives du système état-de-l'art.

Système état-de-l'art : IV-ADLP dans cette configuration, les i-vecteurs sont extraits par Factor Analysis. La classification est effectuée au moyen d'une Analyse Discriminante Linéaire Probabiliste. Ce système correspond à l'état-de-l'art actuel (Garcia-Romero et Espy-Wilson, 2011; Senoussaoui *et al.*, 2011).

Système alternatif 1 : IV-RN-M dans cette configuration, les i-vecteurs sont extraits par Factor Analysis. Lors de la classification, un Radial-NAP de co-rang 350 est appliqué aux i-vecteurs et les scores de vérification sont obtenus par une distance de Mahalanobis.

Système alternatif 2 : ACP-MV-RN-M cette configuration est inspirée par les travaux de (Tipping et Bishop, 1999). Les i-vecteurs sont obtenus selon la méthode du Factor Analysis décrite par l'équation 2 mais dans laquelle la matrice $\mathbf{T}$ est remplacée par la matrice résultant d'une Analyse en Composante Principale. Cette configuration présente l'avantage de remplacer l'apprentissage itératif de la matrice $\mathbf{T}$ du Factor Analysis par un procédé plus rapide.

Système alternatif 3 : ACP-RN-M dans cette configuration, les i-vecteurs sont extraits par une Analyse en Composante Principale. Ils résultent de la projection orthogonale des super-vecteurs dans le sous-espace propre obtenu. Lors de la classification, un Radial-NAP de co-rang 350 est appliqué aux i-vecteurs et les scores de vérification sont obtenus par une distance de Mahalanobis.

Système alternatif 4 : ACP-EC-RN-M et remarque sur la compensation canal : La réduction de la variabilité due au canal en vérification du locuteur a fait l'objet de nombreuses recherches qui ont permis de développer des méthodes intervenant à différentes étapes de la chaîne de vérification du locuteur représentée par la Figure 1 (Pelecanos et Sridharan, 2001; Solomonoff *et al.*, 2005; Dehak *et al.*, 2011). Si la plupart des travaux récents compensent la variabilité canal dans l'espace des *i*-vecteurs (Bousquet *et al.*, 2011; Dehak *et al.*, 2011), la plupart des méthodes existantes peuvent être appliquées. A titre d'exemple, nous proposons un cinquième système utilisant la technique des EigenChannel. Dans cette configuration, les super-vecteurs sont obtenus selon méthode décrite dans (Matrouf *et al.*, 2007) en utilisant un sous-espace canal de dimension 50. Par la suite, les *i*-vecteurs sont extraits par une Analyse en Composante Principale. Il résultent de la projection orthogonale des super-vecteurs dans le sous-espace propre obtenu. Lors de la classification, un Radial-NAP de co-rang 350 est appliqué aux *i*-vecteurs et les scores de vérification sont obtenus par une distance de Mahalanobis.

4.1 Protocole expérimental

Les différents systèmes considérés dans cette étude ont été évalués sur la partie homme de la tâche principale de NIST-SRE08 [1]. Les paramètres acoustiques sont composés de 13 coefficients PLP ainsi que de leurs dérivées premières et secondes. Un unique modèle du monde à 512 distributions a été appris sur de données téléphone et microphone provenant de NIST04 et NIST05. Ces mêmes bases de données augmentées de NIST06 et SwitchBoard ont été utilisées pour l'apprentissage des différents paramètres des systèmes (matrice de Total Variabilitée, ACP, EigenChannel,Radial-NAP).

5 Performances des différents systèmes

Les performances des cinq systèmes considérés sont présentées dans le Tableau 1 en terme de taux d'égales erreurs (EER) pour les 8 conditions de l'évaluation homme NIST-SRE08.
Conformément a l'état-de-l'art, l'approche IV-ADLP obtient les meilleures performances dans 4 des 8 conditions. Il est cependant intéressant de noter que dans les deux conditions 1 et 3, ce même système est moins bon que tous les autres systèmes présentés.
Le système IV-RN-M surpasse le système IV-ADLP dans 3 des 8 conditions. Cependant, les taux d'égales erreurs obtenues dans les conditions incluant des données téléphoniques (conditions 4 à 8) laissent penser que la classification par Radial-NAP et distance de Mahalanobis est moins robuste a la variabilité du canal de transmission que l'Analyse Lineaire Discriminante Probabiliste.
Le système ACP-MV-RN-M utilisant la matrice calculée par ACP obtient les meilleures performances dans 4 des 8 conditions tout en étant relativement proche du système IV-ADLP dans les 4 autres conditions. D'après ces résultats il semble que l'utilisation de la matrice obtenue par Analyse en Composantes Principales sur les super-vecteurs peut être utilisée directement afin d'extraire les *i*-vecteurs.
Le système ACP-RN-M utilisant directement la projection orthogonale des super-vecteurs sur la matrice obtenue par PCA ne se distingue dans aucune des conditions. En revanche, il obtient des taux d'erreurs plus faibles que le système IV-ADLP dans 2 des 8 conditions. Ces résultats sont

1. `http://www.itl.nist.gov/iad/mig/tests/spk/2008/index.html`

d'autant plus intéressant que ce système est totalement déterministe.
De plus les résultats obtenus par le système ACP-EC-RN-M utilisant la technique des EigenChannels montrent qu'il est possible d'améliorer les performances de ce système en ajoutant une compensation de l'effet canal au sein de la chaîne de vérification.

	Systèmes				
Condition	IV-ADLP	IV-RN-M	ACP-MV-RN-M	ACP-RN-M	ACP-EC-RN-M
det 1	5.15	4.95	**3.77**	4.90	4.16
det 2	0.81	**0.62**	1.21	1.04	0.83
det 3	5.37	5.13	**3.91**	5.11	4.30
det 4	**3.73**	4.10	3.87	3.95	4.10
det 5	**3.14**	3.61	4.06	3.64	3.29
det 6	**4.12**	5.03	4.35	5.56	5.15
det 7	**1.37**	1.96	**1.37**	2.49	2.36
det 8	0.56	1.32	**0.44**	1.19	1.75

TABLE 1 – Performances lors de l'évaluation NIST-SRE08 (tests homme) en terme de taux d'égales erreurs (% EER)

6 Conclusions et perspectives

Nous avons montré qu'il est possible d'obtenir des performances proche de l'état-de-l'art en utilisant différentes combinaisons algorithmiques au sein d'un système de vérification du locuteur. La combinaison d'*i*-vecteurs extraits par Factor Analysis et d'une Analyse Linéaire Discriminante Probabiliste obtient les meilleures performances dans la plupart des conditions de test. Cependant, nous avons montré que d'autres systèmes obtenaient des performances comparables. L'Analyse en Composante Principale permet notamment d'approcher ces performances tout en offrant un cadre déterministe plus propice à l'analyse des différentes composantes de la chaîne de vérification du locuteur.
De plus, l'amélioration due à l'utilisation d'une compensation de l'effet canal avant ACP laisse penser qu'il est possible d'ajouter une telle compensation au système état-de-l'art. Cette possibilité sera explorée dans la suite de nos travaux. L'utilisation d'approches discriminantes pour la réduction des vecteurs de grande dimension devra elle aussi être investiguée.

Références

BOUSQUET, P.-M., MATROUF, D. et BONASTRE, J.-F. (2011). Intersession compensation and scoring methods in the i-vectors space for speaker recognition. *In International Conference on Speech Communication and Technology*.

DEHAK, N., KENNY, P., DEHAK, R., DUMOUCHEL, P. et OUELLET, P. (2011). Front-End Factor Analysis for Speaker Verification. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4):788–798.

GARCIA-ROMERO, D. et ESPY-WILSON, C. Y. (2011). Analysis of i-vector length normalization in speaker recognition systems. *In International Conference on Speech Communication and Technology*, pages 249–252.

HYVARINEN, A. et OJA, E. (2000). Independent component analysis : algorithms and applications. *Neural networks*, 13(4-5):411–430.

JOLLIFFE, I. (2002). Principal component analysis. *Encyclopedia of Statistics in Behavioral Science*.

KENNY, P. (2010). Bayesian speaker verification with heavy-tailed priors. *In Speaker and Language Recognition Workshop (IEEE Odyssey)*.

KENNY, P., BOULIANNE, G., OUELLET, P. et DUMOUCHEL, P. (2005). Factor analysis simplified. *In IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, volume 1.

KUHN, R., NGUYEN, P., JUNQUA, J.-C., GOLDWASSER, L., NIEDZIELSKI, N., FINCKE, S., FIELD, K. et CONTOLINI, M. (1998). Eigenvoices for speaker adaptation. *In Proceedings International Conference on Spoken Language Processing, ICSLP*, pages 1771–1774, Sydney (Australia).

LARCHER, A., BOUSQUET, P.-M., LEE, K.-A., MATROUF, D., LI, H. et BONASTRE, J.-F. (2012). I-Vectors in the context of phonetically-constrained short-utterances for speaker verification. *In IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*.

MATROUF, D., SCHEFFER, N., FAUVE, B. et BONASTRE, J.-F. (2007). A straightforward and efficient implementation of the factor analysis model for speaker verification. *In International Conference on Speech Communication and Technology*.

PELECANOS, J. et SRIDHARAN, S. (2001). Feature warping for robust speaker verification. *In Speaker and Language Recognition Workshop (IEEE Odyssey)*.

PRINCE, S. J. (2012). *Computer Vision : Models Learning and Inference*. Cambridge University Press. In press.

PRINCE, S. J. et ELDER, J. H. (2007). Probabilistic linear discriminant analysis for inferences about identity. *In International Conference on Computer Vision*, pages 1–8. IEEE.

SCHEFFER, N., LEI, Y. et FERRER, L. (2011). Factor analysis back ends for MLLR transforms in speaker recognition. *In International Conference on Speech Communication and Technology*, pages 257–260.

SENOUSSAOUI, M., KENNY, P., BRUMMER, N., de VILLIERS, E. et DUMOUCHEL, P. (2011). Mixture of PLDA models in I-vector space for gender independent speaker recognition. *In International Conference on Speech Communication and Technology*.

SOLOMONOFF, A., CAMPBELL, W. et BOARDMAN, I. (2005). Advances in channel compensation for svm speaker recognition. *In IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, volume 1, pages 629–632.

SRINIVASAN, B. V., ZOTKIN, D. N. et DURAISWAMI, R. (2011). A partial least squares framework for speaker recognition. *In IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, pages 5276–5279. IEEE.

TIPPING, M. E. et BISHOP, C. M. (1999). Probabilistic principal component analysis. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 61(3):611–622.

WAN, V. et CAMPBELL, W. M. (2000). Support Vector Machines for Speaker Verification and Identification. *In IEEE Signal Processing Society Workshop Neural Networks for Signal Processing*, volume 2, pages 775–784, Sydney (Australia).

YAMAN, S., PELECANOS, J. et OMAR, M. K. (2011). Boosting Speaker Recognition Performance with Compact Representations. *In International Conference on Speech Communication and Technology*, pages 381–384.