Reconnaissance du locuteur : robustesse face à la variabilité canal

Driss Matrouf

Pub. Inst. Stat. Univ. Paris
56, fasc. 2-3, 2012, 87-98

RECONNAISSANCE DU LOCUTEUR : ROBUSTESSE FACE À LA VARIABILITÉ CANAL

PAR DRISS MATROUF

CERI, Université d'Avignon et des Pays de Vaucluse

Résumé. La classification des formes dans le cadre vocal inclut plusieurs applications, telles que la reconnaissance du locuteur, la reconnaissance de la langue, la reconnaissance des émotions, la reconnaissance de la parole, etc. L'information acoustique utilisée dans ces domaines est généralement basée sur la représentation cepstrale * à court terme. Les vecteurs cepstraux contiennent non seulement l'information utile pour la reconnaissance, mais aussi d'autres types d'informations, telle que la variabilité session. Ces différents types d'informations sont difficilement séparables dans le domaine des vecteurs cepstraux. Récemment, dans le contexte des systèmes de reconnaissance fondés sur les GMM (mélange de gaussiennes), une nouvelle approche utilisant l'Analyse Factorielle a été proposée pour décomposer le modèle d'une forme donnée, en deux composantes : l'information utile et la variabilité session. Dans des travaux précédents nous avons appliqué ce paradigme avec succès à trois applications, la reconnaissance du locuteur, la reconnaissance de la langue et la reconnaissance du genre vidéo. Dans cet article nous allons expliquer les fondements de ce nouveau paradigme, les résultats expérimentaux porteront sur la reconnaissance du locuteur. Nous allons montrer qu'avec cette approche une amélioration supérieure à 50% est observée.

Abstract. Audio pattern classification includes a number of tasks, such as speaker recognition, language recognition, emotion recognition, speech recognition, etc. The feature being used in all these tasks is generally based on a short-term cepstral representation. The cepstral vectors contain at the same time useful information and session variability, which are difficult to separate in this domain. Recently, in the context of GMM-based recognizers, a novel approach using a Factor Analysis paradigm has been proposed for decomposing the target model into a useful information component and a session variability component. In previous work, we successfully apply this paradigm to three automatic audio processing applications, speaker verification, language recognition, and video genre recognition. In this paper we will focus on speaker recognition, We will show that this approach allows for a relative error reduction of over 50%.

1. Introduction. Malgré les efforts déployés dans les domaines de la paramétrisation et de la modélisation en vu de la reconnaissance du locuteur, les Systèmes de Vérification du Locuteur (SVL) échouent face au problème de changement des

Mots-clefs : Varaibilité session, Reconnaissance du locuteur, GMM-UBM, Analyse Factorielle

*.

Il s'agit de la transformée de Fourier inverse du logarithme de la densité spectrale

conditions acoustiques, qui peuvent varier grandement d'une session à l'autre. D'une manière générale, les termes décalage de session (*session mismatch*) ou variabilité session (*session variability*) sont utilisés pour désigner ce phénomène. C'est une des grandes causes de dégradation des performances des SVL. Le terme « variabilité session » englobe un grand nombre de phénomènes acoustiques : le canal de transmission, le bruit environnant, (brouhaha, voitures, téléviseur,...), la géométrie de l'endroit où se passe l'enregistrement (hall, bureau, ville,...), la position du microphone par rapport à la bouche, et enfin les variabilités introduites par le locuteur lui-même.

Ces phénomènes introduisent des décalages acoustiques ayant un grand pouvoir de nuisance sur les SVL. En effet, la plupart des erreurs de vérification sont dues, non pas à la ressemblance entre les voix des différents locuteurs, mais surtout à la variabilité intrinsèque de phrases (ou de sessions) appartenant à un même locuteur. Il n'est pas facile de trouver des solutions pour chaque type de variabilité acoustique, la plupart des solutions proposées dans la littérature traitent le problème de la variabilité de la session dans sa globalité. Les solutions proposées agissent à différents niveaux du système de reconnaissance du locuteur : dans l'espace des paramètres acoustiques, dans l'espace des modèles acoustiques ou encore dans l'espace des scores de reconnaissance. Nous allons passer en revue ces différentes approches. Nous allons décrire avec plus de détails une solution proposée récemment et qui a permis une grande robustesse face à la variabilité session : l'analyse factorielle (FA : *Factor Analysis*).

2. Analyse Factorielle pour la modélisation de session. Les SVL « état de l'art » sont fondés sur l'approche appelée « GMM-UBM ». L'UBM (Univeral Background Model) est un modèle statistique modélisant les voix d'une manière générale, il est aussi appelé modèle du monde. Ce modèle est un GMM (Gaussian Mixture Model) modélisant la génération de vecteurs cepstraux provenant d'une multitude de locuteurs (plusieurs centaines). L'estimation des paramètres de ce modèle est réalisée en utilisant l'algorithme EM (Expectation and Maximization), connu pour la résolution de problèmes d'estimation à partir de données incomplètes [2]. À partir du modèle du monde, les modèles de locuteurs sont dérivés en utilisant l'approche MAP (*Maximum A Posteriori*) [3]. L'approche MAP permet d'estimer le GMM modélisant un locuteur avec relativement peu de données. Dans le cadre de la vérification du locuteur seules les moyennes sont réestimées, les poids et les variances restent inchangés. D'une manière plus formelle, notons $\mathbf{m}$, la concaténation des moyennes du modèle du monde, appelé, dorénavant, super-vecteur. Le super-vecteur d'un locuteur s peut s'écrire comme suit,

$$\mathbf{m}_{(s)} = \mathbf{m} + \mathbf{D}\mathbf{y}_s, \tag{2.1}$$

où $\mathbf{m}_s$ est le super-vecteur moyenne correspondant au locuteur s. $\mathbf{D}$ est une matrice diagonale $MD \times MD$ (M étant le nombre de gaussiennes dans les GMMs et D la dimension de l'espace des paramètres acoustiques), $\mathbf{y}_s$ est un vecteur de dimension MD estimé sur les données du locuteur s. $\mathbf{D}$ satisfait l'équation $\mathbf{I} = \tau \mathbf{D}^t \mathbf{\Sigma}^{-1} \mathbf{D}$

où τ est un facteur appelé *relevance factor*, utilisé dans l'approche MAP. $\mathbf{\Sigma}$ est la matrice dont la diagonale est formée par la concaténation des variances des diffrentes gaissiennes du GMM du monde, sa taille est $MD \times MD$ ($\mathbf{DD}^t$ représente la matrice de covariance *a priori* de $\mathbf{y}_s$).

Imaginons que pour le même locuteur s, il existe deux sessions h_1 et h_2, les équations d'adaptation MAP pour les deux sessions deviennent :

$$\mathbf{m}_{(s,h1)} = \mathbf{m} + \mathbf{D}\mathbf{y}_{s,h_1}, \mathbf{m}_{(s,h2)} = \mathbf{m} + \mathbf{D}\mathbf{y}_{s,h_2}, \tag{2.2}$$

on constate que le modèle résultant est différent selon la session utilisée pour l'apprentissage. En effet, les modèle résultants contiennent non seulement le locuteur s mais aussi les sessions h_i, ce qui n'est, bien sûr, pas souhaitable. En effet, le score de vérification, pour le même locuteur cible, aura une valeur différente selon la session utilisée durant l'apprentissage.

S'il était possible de séparer les deux composantes locuteur et session dans les équations 2.2 alors le problème serait résolu. C'est ce qui a était proposé dans l'approche **Factor Analysis** (FA). Dans ce cadre, le modèle de locuteur s'écrit comme étant la somme de trois composantes : une composante générale indépendante du locuteur et de la session, une composante dépendante seulement du locuteur et une composante dépendante seulement de la session. Dans le cadre du FA, le modèle du locuteur s dans la session h s'écrit :

$$\mathbf{m}_{(h,s)} = \mathbf{m} + \mathbf{D}\mathbf{y}_s + \mathbf{U}\mathbf{x}_{(h,s)}, \tag{2.3}$$

où $\mathbf{m}_{(h,s)}$ est le super-vecteur correspondant au locuteur s dans la session h. $\mathbf{D}$ joue le même rôle que dans l'équation 2.2. Ce qui est nouveau ici, c'est le dernier terme $\mathbf{U}\mathbf{x}_{(h,s)}$: $\mathbf{U}$ est une matrice de rang faible R ($R << MD$). Ses vecteurs colonnes forment une base d'un sous-espace dans lequel la variabilité session est la plus forte. $\mathbf{x}_{(h,s)}$ est un vecteur de dimension R contenant les composantes relatives à la session dans ce sous-espace. L'hypothèse faite est que la variabilité session peut être localisée dans un sous-espace de faible dimension.

Le succès du FA repose essentiellement sur l'estimation de la matrice $\mathbf{U}$. Cette estimation nécessite une grande quantité de données provenant d'un grand nombre de locuteurs, avec plusieurs sessions par locuteur. Dans le paradigme FA, cette dernière contrainte est capitale, car c'est en utilisant plusieurs sessions par locuteur qu'on arrive à isoler et modéliser l'effet session.

Le modèle correspondant à l'équation 2.3 a rencontré un grand succès pour plusieurs raisons. La première est que les hypothèses faites se sont avérées, a posteriori, correspondre à la réalité ; la seconde est liée à la qualité d'estimation des paramètres. En effet le modèle de l'équation 2.3 permet un équilibre entre la quantité de paramètres à estimer et les quantités de données d'apprentissage disponibles : $\mathbf{x}_{(h,s)}$ est un vecteur de petite dimension estimé sur la session ; $\mathbf{y}_s$ est un vecteur de dimension plus grande MD estimé sur toutes les sessions appartenant au locuteur s ; enfin $\mathbf{U}$ une matrice avec beaucoup de paramètres estimée sur un très grand nombre de sessions appartenant à un grand nombre de locuteurs.

3. Une implémentation facile du FA. Dans cette section, nous définissons $\mathbf{A}$ comme la matrice de dimension $MD \times K$ formée par concaténation verticale de M matrices de dimensions $D \times K$. Notons $\{\mathbf{A}\}_{[g]}$ la g^{eme} sous-matrice dans $\mathbf{A}$ (g correspondra à l'indice d'une gaussienne dans l'UBM).

3.1. *Statistiques.* Afin d'estimer les différentes composantes du modèle FA et notamment la matrice $\mathbf{U}$, il est nécessaire de calculer pour chaque gaussienne dans l'UBM les statistiques d'ordre zéro et d'ordre un. Comme dans l'approche MAP, toutes les statistiques sont calculées en utilisant le modèle du monde (UBM). Soient $\mathbf{N}_s$ et $\mathbf{N}_{(h,s)}$ deux vecteurs de dimension M contenant les statistiques d'ordre 0 relatives respectivement au locuteur s et à la session (h, s) :

$$\mathbf{N}_s[g] = \sum_{t \in s} \gamma_g(t), \quad \mathbf{N}_{(h,s)}[g] = \sum_{t \in (h,s)} \gamma_g(t), \tag{3.1}$$

où $\gamma_g(t)$ est la probabilité *a posteriori* de la gaussienne g étant donnée la trame à l'instant t. Dans l'équation $\sum_{t \in s}$ signifie la somme sur toutes trames appartenant au locuteur s, et $\sum_{t \in (h,s)}$ signifie la somme sur toutes les trames appartenant à la session h du locuteur s.

Soient $\mathbf{X}_s$ et $\mathbf{X}_{(h,s)}$ deux vecteurs (de dimension $M \times D$) contenant les statistiques du premier ordre relatif au locuteur et à la session :

$$\{\mathbf{X}_s\}_{[g]} = \sum_{t \in s} \gamma_g(t) \cdot x_t, \quad \{\mathbf{X}_{(h,s)}\}_{[g]} = \sum_{t \in (h,s)} \gamma_g(t) \cdot x_t, \tag{3.2}$$

où x_t désigne la trame à l'instant t.

3.2. *Estimation des variables latentes.* Ici nous supposons que la matrice $\mathbf{U}$ est connue, et nous nous intéressons à l'estimation de la composante locuteur $\mathbf{y}_s$ et de la composante session $\mathbf{x}_{(h,s)}$. Nous commençons par appliquer une opération de centrage des statistiques du premier ordre :

- on enlève des statistiques relatives au locuteur les composantes sessions
- on enlève des statistiques relatives à la session la composante locuteur

voici les équations :

$$\{\overline{\mathbf{X}}_s\}_{[g]} = \{\mathbf{X}_s\}_{[g]} - \sum_{h \in s} \mathbf{N}_{(h,s)}[g] \cdot \{\mathbf{m} + \mathbf{U}\mathbf{x}_{(h,s)}\}_{[g]}, \tag{3.3}$$

$$\{\overline{\mathbf{X}}_{(h,s)}\}_{[g]} = \{\mathbf{X}_{(h,s)}\}_{[g]} - \{\mathbf{m} + \mathbf{D}\mathbf{y}_s\}_{[g]} \cdot \sum_{h \in s} \mathbf{N}_{(h,s)}[g]. \tag{3.4}$$

Soient $\mathbf{L}_{(h,s)}$ la matrice de dimension $R \times R$, et $\mathbf{B}_{(h,s)}$ le vecteur de dimension R, définies comme suit :

$$\mathbf{L}_{(h,s)} = \mathbf{I} + \sum_{g \in \text{UBM}} \mathbf{N}_{(h,s)}[g] \cdot \{\mathbf{U}\}^t_{[g]} \cdot \mathbf{\Sigma}^{-1}_{[g]} \cdot \{\mathbf{U}\}_{[g]}, \tag{3.5}$$

$$\mathbf{B}_{(h,s)} = \sum_{g \in \text{UBM}} \{\mathbf{U}\}^t_{[g]} \cdot \mathbf{\Sigma}^{-1}_{g} \cdot \{\overline{\mathbf{X}}_{(h,s)}\}_{[g]}, \tag{3.6}$$

où $\mathbf{\Sigma}_g$ est la matrice de covariance de la g^{eme} gaussienne dans l'UBM. En utilisant $\mathbf{L}_{(h,s)}$ et $\mathbf{B}_{(h,s)}$, $\mathbf{x}_{(h,s)}$ et $\mathbf{y}_s$ sont obtenus en utilisant les équations suivantes :

$$\mathbf{x}_{(h,s)} = \mathbf{L}_{(h,s)}^{-1} \cdot \mathbf{B}_{(h,s)}, \tag{3.7}$$

$$\{\mathbf{y}_s\}_{[g]} = \frac{\tau}{(\tau + \mathbf{N}_s[g])} \cdot \mathbf{D}_g \cdot \mathbf{\Sigma}_g^{-1} \cdot \{\overline{\mathbf{X}}_s\}_{[g]}, \tag{3.8}$$

où $\mathbf{D}_g = \frac{\mathbf{\Sigma}_g^{1/2}}{\sqrt{\tau}}$, τ est le *relevance factor* du MAP (14.0 dans nos expériences).

3.3. *Estimation de la matrice intersession (U).* La matrice $\mathbf{U}$ peut être obtenue ligne par ligne. Soit $\{\mathbf{U}\}_{[g]}^i$ la i^{eme} ligne de $\{\mathbf{U}\}_{[g]}$:

$$\mathbf{U}_{[g]}^i = \mathbf{L}\mathbf{U}_g^{-1} \cdot \mathbf{R}\mathbf{U}_g^i, \tag{3.9}$$

où $\mathbf{R}\mathbf{U}_g^i$ et $\mathbf{L}\mathbf{U}_g$ sont donnés par :

$$\mathbf{L}\mathbf{U}_g = \sum_s \sum_{h \in s} (\mathbf{L}_{(h,s)}^{-1} + \mathbf{x}_{(h,s)} \mathbf{x}_{(h,s)}^T) \cdot \mathbf{N}_{(h,s)}[g], \tag{3.10}$$

$$\mathbf{R}\mathbf{U}_g^i = \sum_s \sum_{h \in s} \{\overline{\mathbf{X}}_{(h,s)}\}_{[g]}[i] \cdot \mathbf{x}_{(h,s)}. \tag{3.11}$$

Les équations précédentes conduisent à l'Algorithme 1, permettant d'estimer la matrice de variabilité inter-sessions ($\mathbf{U}$). Cet algorithme utilise exclusivement les équations décrites ci-dessus ; la fonction de vraisemblance standard est utilisée comme critère de convergence. Nos expériences ont montré que cet algorithme nécessite une dizaine d'itérations pour converger. Dans la pratique l'algorithme n'est pas sensible à l'initialisation. Avec différentes initialisations on aboutit aux même performances. Une fois la matrice $\mathbf{U}$ estimée, les modèles de locuteurs peuvent être estimés en utilisant le même algorithme (Algorithme 1), sans réestimer la matrice $\mathbf{U}$. Dans le cas d'une seule session par locuteur, les expériences ont montré qu'une seule itération est suffisante pour converger.

3.4. *Score de vérification.* Dans les paragraphes précédents nous avons montré comment estimer un modèle de locuteur (GMM) en mettant en évidence les deux composantes session et locuteur. Dans les systèmes standards fondés sur l'approche UBM-GMM, en phase de test, on utilise d'un côté des données de test et de l'autre côté un modèle de locuteur. Bien entendu l'effet de session est présent dans les deux phases test et apprentissage. La procédure de vérification doit intégrer le fait que l'effet de session est présent non seulement en phase d'apprentissage mais aussi en phase de test.

Le *Log-Likelihood Ratio* (LLR) est généralement utilisé pour estimer le score de vérification. Soit $\mathcal{Y} = \{y_1 \dots y_T\}$ une suite de trames correspondant à la séquence de test, et $\mathbf{s}_{tar}$ l'identité annoncée.

$$Score_{s_{tar}}(\mathcal{Y}) = \frac{1}{T}(LLK_{s_{tar}}(\mathcal{Y}) - LLK_w(\mathcal{Y})), \tag{3.12}$$

Algorithme 1 : Algorithme d'estimation de $\mathbf{U}$

Pour chaque locuteur s et session h : $\mathbf{y}_s \leftarrow 0$, $\mathbf{x}_{(h,s)} \leftarrow 0$ $\mathbf{U} \leftarrow$ *random* ($\mathbf{U}$ est initialisée aléatoirement);
Estimer les statistiques : $\mathbf{N}_h$, $\mathbf{N}_{(h,s)}$, $\mathbf{X}_s$, $\mathbf{X}_{(h,s)}$ (eq. 3.1 et 3.2);
for $i = 1$ *to nb_iterations* **do**
  for *all s et h* **do**
    Centrer les statistiques : $\overline{\mathbf{X}}_s$, $\overline{\mathbf{X}}_{(h,s)}$ (eq. 3.4);
    Estimer $\mathbf{L}^{-1}_{(h,s)}$ et $\mathbf{B}_{(h,s)}$ (eq. 3.6);
    Estimer $\mathbf{x}_{(h,s)}$ et $\mathbf{y}_s$ (eq. 3.8);
  end
  Estimer la matrice $\mathbf{U}$ (eq. 3.9 et 3.11) ;
end

où $LLK_{s_{tar}}(\mathcal{Y})$ est le logarithme de la vraisemblance de $(\mathcal{Y})$ par rapport au modèle de locuteur s_{tar}, et $LLK_w(\mathcal{Y})$ est le logarithme de la vraisemblance de $(\mathcal{Y})$ par rapport au modèle du monde. T étant le nombre de trames dans la séquence de trames $\mathcal{Y}$. Afin de prendre en compte l'effet session dans le test et l'apprentissage, $LLK_{s_{tar}}(\mathcal{Y})$ peut être estimée de la manière suivante :

$$LLK_{s_{tar}}(\mathcal{Y}) = argmax_{\mathbf{x}} f(\mathcal{Y}|\mathbf{m} + \mathbf{D}\mathbf{y}_{s_{tar}} + \mathbf{U}\mathbf{x})\mathcal{N}(\mathbf{x}|O, I). \tag{3.13}$$

$\mathbf{m}+\mathbf{D}\mathbf{y}_{s_{tar}}$ dans l'équation 3.13 représente le super-vecteur moyenne du locuteur s_{tar} sans l'effet de session. La décomposition est obtenue en utilisant l'Algorithme 1 (en fixant $\mathbf{U}$). f est la fonction de vraisemblance. $\mathcal{N}(x|O, I)$ est la distribution *a priori* de l'effet du canal. $\mathbf{U}\mathbf{x}$ représente le décalage de session pouvant exister entre le test et l'apprentissage. $LLK_w(\mathcal{Y})$ est obtenue de la même manière que pour le modèle de locuteur s_{tar} avec $\mathbf{y} = 0$:

$$LLK_w(\mathcal{Y}) = argmax_{\mathbf{x}} f(\mathcal{Y}|\mathbf{m} + \mathbf{U}\mathbf{x})\mathcal{N}(\mathbf{x}|O, I). \tag{3.14}$$

Le problème dans les équations 3.13 et 3.14 est l'hypothèse que l'effet session du test dépend du locuteur cible. Ceci est bien évidemment incorrect. De plus, elle suppose que l'effet session du locuteur cible est correctement éliminé avant d'estimer la vraisemblance en utilisant les équations 3.13 et 3.14. Les expériences menées dans [4] et [5] montrent qu'en utilisant cette approche, l'essentiel du gain n'est obtenu qu'après normalisation des scores en utilisant, par exemple, la ZT-norm. Généralement dans le domaine de la reconnaissance du locuteur on utilise des techniques de normalisation des score telles Z-norm, T-norm ou ZT-norm. Il s'agit de transformations appliquées aux scores permettant de reduire l'effet session[14].

Dans la section suivante nous présentons une approche de test de vérification hybride [1] : la compensation de l'effet session dans le test est faite dans le domaine des paramètres acoustiques, par contre, la compensation de l'effet session durant l'apprentissage est effectuée dans l'espace des modèles.

3.5. *Test : compensation hybride de la session.* Soit $\mathbf{s}_{tar}$ et $\mathbf{s}_{test}$ les locuteurs cible et test respectivement. Soit $\mathcal{Y} = \{y_1 \dots y_T\}$ une séquence de vecteurs acoustiques appartenant au locuteur de test $\mathbf{s}_{test}$. La tâche de vérification consiste à déterminer si $\mathcal{Y}$ a été prononcée par le locuteur $\mathbf{s}_{tar}$. En utilisant la décomposition FA pour les locuteurs d'apprentissage et de test, on peut écrire :

$$\mathbf{m}_{(\mathbf{h}_{tar},\mathbf{s}_{tar})} = \mathbf{m} + \mathbf{D}\mathbf{y}_{\mathbf{s}_{tar}} + \mathbf{U}\mathbf{x}_{\mathbf{h}_{tar}}, \tag{3.15}$$

$$\mathbf{m}_{(\mathbf{h}_{test},\mathbf{s}_{test})} = \mathbf{m} + \mathbf{D}\mathbf{y}_{\mathbf{s}_{test}} + \mathbf{U}\mathbf{x}_{\mathbf{h}_{test}}. \tag{3.16}$$

En éliminant les effets sessions dans les équations 3.15 et 3.16, on obtient :

$$\mathbf{m}_{\mathbf{s}_{tar}} = \mathbf{m} + \mathbf{D}\mathbf{y}_{\mathbf{s}_{tar}};\ \mathbf{m}_{\mathbf{s}_{test}} = \mathbf{m} + \mathbf{D}\mathbf{y}_{\mathbf{s}_{test}}. \tag{3.17}$$

Ici, contrairement à [4] et [5], les locuteurs de test et d'apprentissage sont traités d'une manière tout-à-fait indépendante. Le score de vérification est généralement estimé comme étant l'espérance du log de rapport de vraisemblance :

$$LLK(\mathcal{Y}|\mathbf{m}_{(\mathbf{h}_{tar},\mathbf{s}_{tar})}) - LLK(\mathcal{Y}|\mathbf{m}), \tag{3.18}$$

où $LLK(\cdot|\cdot)$ désigne la moyenne du log-vraisemblance sur toutes les trames de test. Deux approches de compensation sont possibles. La première serait de compenser l'effet session au niveau des trames acoustiques, autrement dit, la compensation est vu comme faisant partie de la paramétrisation. L'approche que nous avons adoptée est plutôt hybride : l'effet session côté locuteur cible est enlevé dans l'espace des modèles, par contre, celui de test est enlevé dans l'espace des trames (l'espace des paramètres acoustiques). La compensation de l'effet session est effectuée trame par trame, en utilisant la formule suivante (la même formule a été utilisée avec succès dans [6]) :

$$\hat{x}_t = x_t - \sum_{g=1}^{M} \gamma_g(t) \cdot \{\mathbf{U} \cdot \mathbf{x}_{\mathbf{h}_{test}}\}_{[g]}. \tag{3.19}$$

Le score de vérification devient :

$$LLK(\hat{\mathcal{Y}}|\mathbf{m}_{\mathbf{s}_{tar}}) - LLK(\hat{\mathcal{Y}}|\mathbf{m}), \tag{3.20}$$

où $\hat{\mathcal{Y}}$ est la suite de trames obtenues en compensant l'effet session dans les trames de $\mathcal{Y}$ (en utilisant l'équation 3.19).

3.6. *Vérification et modélisation par SVM.* En utilisant les équations 3.17, les super-vecteurs de moyennes ne contiennent plus que l'information concernant les locuteurs en question, les composantes session étant écartées. Dans [7], les auteurs ont proposé un noyau probabiliste permettant de calculer une distance entre GMMs, bien adapté pour un classifieur fondé sur les SVM. Soit $\mathcal{X}_{\mathbf{s}}$ et $\mathcal{X}_{\mathbf{s}'}$ deux séquences de vecteurs acoustiques correspondants à deux locuteurs $\mathbf{s}$ et $\mathbf{s}'$, la formulation du noyau proposé est donnée par :

$$K(\mathcal{X}_{\mathbf{s}}, \mathcal{X}_{\mathbf{s}'}) = \sum_{g=1}^{M} \left(\sqrt{\alpha_g}\Sigma_g^{-\frac{1}{2}}\mathbf{m}_s^g\right)^t \left(\sqrt{\alpha_g}\Sigma_g^{-\frac{1}{2}}\mathbf{m}_{s'}^g\right). \tag{3.21}$$

α_g, $\mathbf{m}_g$ et Σ_g sont le poids, la moyenne et la matrice de covariance de la gaussienne g dans les GMMs. Ce noyau est valide dans le cas où les GMMs de locuteurs ne diffèrent par rapport au modèle du monde que par leurs moyennes. $\mathbf{m}_s$ est extrait du modèle de locuteur suivant les équations 3.17, *i.e.* $\mathbf{m}_s = \mathbf{m} + \mathbf{D}\mathbf{y}_s$. $\mathbf{m}_s$ étant le super-vecteur obtenu par concaténation des $\mathbf{m}_s^g$.

En utilisant un classifieur fondé sur les SVM, le test de vérification se fait complètement dans le domaine des modèles. Ce qui simplifie le processus, car la décomposition FA est naturellement faite dans le domaine des modèles, et pas dans le domaine des paramètres acoustiques.

Une approche similaire à FA, appelé NAP (*Nuisance Attribute Projection*) a été proposée afin de réduire l'effet du canal [8]. NAP est une approche dédiée et adaptée à l'utilisation dans le cadre d'un système SVM-UBM.

4. Protocole expérimental. Toutes les expériences ont été réalisées en utilisant la plate-forme ALIZE et LIA_SpkDet toolkit[9][10]. Les expériences présentées dans cette section ont été réalisées sur la base de données NIST SRE 2005, considérée comme corpus de développement ; et sur la base de données NIST SRE 2006, considérée comme corpus de validation. Les expériences concernent seulement des locuteurs masculins. Le protocole NIST SRE 2005 défini 274 locuteurs clients et 9012 tests, dont seulement 741 sont des tests clients. Le protocole NIST SRE 2006 défini 354 locuteurs clients et 9720 tests, dont seulement 741 sont des tests clients[1]. Les résultats sont donnés en terme de taux d'égale erreur (EER : *Equal-Error-Rate*) et en terme de DCF minimum[2]. Dans ces conditions, les durées des segments d'apprentissage et de test sont de l'ordre de 2.5 minutes (conversations téléphoniques, seulement 30% des trames sont utilisées). La matrice engendrant le sous-espace des variabilités sessions ($\mathbf{U}$) a été entraînée en utilisant la base de données NIST 2004 (124 locuteurs et 2938 sessions). Pour les normalisations de scores et les listes noires (utilisées pour entraîner les SVM), nous avons utilisé la base de donnée NIST 2004.

Les trames sont composées de 19 coefficients LFCC et de leurs dérivées de premières et secondes ordre (nous retenons 50 coefficients incluant l'énergie). Seules les trames paroles sont retenues (un séparateur parole/silence simple est utilisé). Sur les trames retenues, une normalisation est appliquée, de façon à obtenir pour chaque session, une distribution avec une moyenne 0 et une variance 1 sur chaque coefficient. Le modèle du monde possède 512 gaussiennes.

4.1. *Apprentissage des SVM.* La boite à outils LIA_SpkDet contient désormais la librairie LIBSVM [11] permettant d'entraîner et d'utiliser les SVM. C'est ce qui

1. Le protocole 2005 correspond à la condition de base appelée *det7* et le protocole 2006 correspond la condition de base appelée *det3*

2. DCF est une mesure de performance intégrant certaines connaissances, telles que le coût d'une mauvaise décision et la probabilité *a priori* qu'un segment provienne d'un imposteur

est utilisé pour réaliser les expériences de ce chapitre.

4.2. *Le système de base GMM-UBM.* Nous comparerons tous les résultats avec ceux du système de base GMM-UBM [12]. Le modèle du monde utilisé est le même que celui qui a été utilisé dans la campagne NIST-SRE-2006 [3]. Il a été entraîné en utilisant la base de données Fisher database [4], 10 million de trames ont été utilisées. Les modèles de locuteurs ont été obtenus par adaptation Baysienne des moyennes.

5. Résultats expérimentaux. Le tableau Tab. 1 montre les résultats obtenus par le système de base (UBM-GMM), seule la normalisation T-norm est présentée (les autres normalisations, Z et ZT-norm, n'ont pas apporté de gain).

	SRE-05		SRE-06	
	DCFmin (× 100)	EER (%)	DCFmin (× 100)	EER (%)
sans_norm	3.83	7.15	3.88	6.79
T-norm	3.05	8.52	2.9	5.7

TABLE 1
Les résultats du système de base (GMM-UBM) sur les protocoles NIST SRE 2005 et 2006. DCFmin (× 100), EER(%).

Le tableau Tab. 2 donne les résultats (sans normalisation des scores) selon le rang de la matrice de projection **U**. Nous observons que le meilleur rang est 40. Il est imporant de noter que notre implémentation du FA, donne de très bons résultats sans avoir besoin de procéder à des normalisation des scores [5].

	Subspace rank					
	0	20	40	60	80	100
DCFmin (x100)	3.83	2.05	**1.83**	1.93	1.95	1.99
EER (%)	7.15	5.1	4.42	**4.22**	4.31	4.23

TABLE 2
Résultats selon le rang de la matrice **U**. *Protocole NIST SRE 2005. DCFmin(×100), EER(%)*

Le tableau Tab. 3 montre qu'un gain supplémentaire (du même ordre que celui habituellement observé lors de l'application de techniques de normalisation) est obtenu lorsqu'on applique des techniques de normalisation de scores. Nous remarquons que la T-norm apporte un gain seulement dans le protocole 2006. La Z-norm permet, dans le protocole 2005, une baisse de DCF-min de 1.83 à 1.64. Dans le protocole 2006, la ZT-norm permet une réduction de DFC-min de 1.61 à 1.18. Les comportements après normalisations sont très différents entre les protocoles 2005 et 2006, cependant nous observons que dans les deux cas, la ZT-norm apporte un gain significatif et stable.

3. NIST 2006, SRE evaluation plan, `www.nist.gov/speech/tests/spk/2006/sre-06_evalplan-v9.pdf`

4. Fisher English Training Speech Part 1, LDC, number : LDC2004S13

5. les publications dans ce domaine, montrent que les gain ne sont obtenus qu'après normalisation des score (T-norm, ZT-norm)

	SRE-05		SRE-06	
	DCFmin (× 100)	EER (%)	DCFmin (× 100)	EER (%)
No-norm	1.83	4.42	1.61	2.97
T-norm	1.84	4.72	1.29	2.83
ZT-norm	1.72	4.62	**1.18**	**2.15**
Z-norm	**1.64**	**4.21**	1.46	2.33

TABLE 3

Le rang est fixé à 40. Résultats de l'approche GMM-UBM-FA avec différentes normalisations. DCFmin(x100), EER(%)

Le tableau Tab. 4, présente les résultats du système SVM-UBM. Avec le protocole 2005, Le SVM donnent des performances similaires à celles obtenues par le système GMM-UBM. Par contre, avec le protocole 2006, le SVM est plus performant. Il est important de noter qu'avec le SVM, la Z-norm ne permet aucun gain, le gain est essentiellement apporté par la T-norm.

	SRE-05		SRE-06	
	DCFmin (× 100)	EER (%)	DCFmin (× 100)	EER (%)
Baseline (40)	1.97	4.83	1.40	2.83
Z-norm	1.92	5.36	1.54	2.70
T-norm	1.61	**4.42**	**1.03**	2.29
ZT-norm	**1.58**	4.51	1.06	**2.16**

TABLE 4

Résultats obtenus par le SVM-UBM (rang=40) sur les protocoles NIST SRE 2005 et 2006. DCFmin(×100), EER(%). La T-norm apporte le gain le plus significatif.

6. Conclusion. La majorité des techniques de compensation échouent à atteindre leur but pour des raisons différentes. Par exemple, la technique *feature mapping*, la technique *speaker model synthesis* [13], et la *H-norm* sont sous-optimales car elles considèrent que les sources de variabilités sessions sont un ensemble discret et fini, ce qui simplifie grandement le problème. Cependant, quand les caractéristiques de session sont très différentes de celles considérées, l'application de la normalisation devient inappropriée, ce qui peut causer des erreurs de vérification.

Une seconde explication d'échec d'un certain nombre d'approches réside dans le fait qu'elles tentent de compenser les variabilités de session sans les modéliser. Nous pouvons citer dans cette catégorie, le *feature warping*, la T-Norm et la Z-Norm. Ces méthodes n'utilisent aucune connaissance *a priori* concernant les variabilités indésirables, mais utilisent certaines connaissances sur les effets causés par celles-ci.

Le FA présenté dans ce chapitre, modélise explicitement la variabilité session. Cette approche ne suppose pas que l'effet session est discret ou fini, au contraire, il est supposé infini et continu. Cette approche est fondée sur une hypothèse forte : la variabilité session est située dans un sous-espace de faible dimension. Quand cette hypothèse est vérifiée, les gains de performances apportés sont très importants (réduction du taux d'erreurs de plus de 50%).

Enfin, dans ce chapitre, nous avons montré que la contrainte selon laquelle on doit disposer de plusieurs sessions par locuteur pour l'apprentissage de la matrice **U** (génératrice du sous-espace de la variabilité session), n'est pas indispensable pour observer des gains significatifs par l'application du FA.

RÉFÉRENCES

[1] Matrouf, D. and Scheffer, N. and Fauve, B. and Bonastre, J-F. (2007) *A Straightforward and Efficient Implementation of the Factor Analysis Model for Speaker Verification*, INTERSPEECH Conference, Antewerp, Belgium, 2007.

[2] Dempster, A. P. and Laird, N. M. and Rubin, D. B. (1977). *Maximum-likelihood from incomplete data via the (EM) algorithm*, Journal of Acoustical Society of America (JASA), Volume 39, Pages 1-38, 1977.

[3] Reynolds, D. A. (1995). *Speaker identification and verification using gaussian mixture speaker models*, sc,Volume 17(1-2), Pages 91-108, 1995.

[4] Kenny, P. and Boulianne, G. and Ouellet, P. and Dumouchel, P. (2005). *Factor Analysis Simplified*, icassp-05, Volume 1, 2005.

[5] Vogt, R. and Baker, B. and Sridharan, S. (2005) *Modelling Session Variability in Text-Independent Speaker Verification*, eurospeech-05.

[6] Vair, C. and Colibro, D. and Laface, P. *Channel Factors Compensation in Model and Feature Domain for Speaker Recognition*, odyssey-06, Jun, 2006.

[7] Campbell, WM and Campbell, JP and Reynolds, DA and Singer, E. and Torres-Carrasquillo, PA *Support vector machines for speaker and language recognition*, Computer Speech and Language, Volume 20, Number (2-3), Pages 210-229, publisher Elsevier, 2006.

[8] A. Solomonoff, W. M. Campbell, and I. Boardman *Advances in channel compensation for SVM speaker recognition*, Proceedings of ICASSP, 2005.

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

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

[11] Chih-Chung Chang and Chih-Jen Lin *LIBSVM : a library for support vector machines*, Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`, 2001.

[12] Reynolds, D. A. and Quatieri, T. F. and Dunn, R. B. *Speaker verification using adapted Gaussian mixture models*, Digital Signal Processing, Volume 10, Number 1, Pages 19-41, editor J. Schroeder, J. Campbell, year 2000.

[13] Teunen, R. and Shahshahani, B. and Heck, L. *A model-based transformational approach to robust speaker recognition*, International Conference on Spoken Language Processing, Pages 495 - 498, 2000

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

Centre d'Enseignement et de Recherche en Informatique (CERI)
Agroparc - BP 91228
339, chemin des Meinajariès
84911 Avignon cedex 9
driss.matrouf@univ-avignon.fr