# Représentation compacte des locuteurs par distribution sur les modèles d'ancrage

*Yassine Mami et Delphine Charlet*

France Télécom R&D
2 avenue Pierre Marzin 22307 Lannion - FRANCE
{yassine.mami, delphine.charlet}@rd.francetelecom.com

## ABSTRACT

Speaker representation by location in a reference space is a new technique of speaker recognition and adaptation. It consists in representing a speaker not absolutely but rather relatively, by comparing him to a set of well trained speaker models. The main motivation is that the dimension (number of parameters) of the absolute speaker models is very large compared to the amount of free parameters that can be reliably estimated with few training data. In this paper, we recall the concept of relative location for speaker recognition. Then, we introduce a statistical approach for speaker location to cope with the weaknessess of the classical relative approach. In-depth evaluations on a telephone database show that this concept of relative location is a promising way, as it leads to recognition rates similar to those obtained with the GMM-UBM approach with more compact models.

## 1. INTRODUCTION

La représentation par localisation dans un espace de locuteurs de référence est une nouvelle technique de reconnaissance du locuteur. Il s'agit de représenter un locuteur, non plus de façon absolue, mais relativement à un ensemble de locuteurs dont les modèles sont bien appris [4]. Dans cet article, nous rappelons le principe de reconnaissance de locuteurs par placement dans un espace optimisé. Cette approche "géométrique" accorde une place symétrique à l'apprentissage et au test, ce qui peut être un défaut important puisque la quantité de données pour l'apprentissage et pour le test peuvent être radicalement différentes. Pour pallier ce problème, nous introduisons une asymétrie entre l'apprentissage et le test. Pour cela, nous présentons une nouvelle représentation des locuteurs basée sur une distribution de distances. L'idée est de représenter un locuteur par une densité de probabilité sur les distances qui modélise sa position par rapport à un ensemble de modèles de locuteurs de référence.

## 2. RECONNAISSANCE DE LOCUTEURS PAR LOCALISATION

Les techniques d'adaptation rapide, basées sur le principe de la représentation relative des locuteurs, ont été développées initialement dans le cadre de la reconnaissance automatique de la parole. Ces techniques reposent sur le principe d'utiliser des connaissances a priori obtenues à partir d'un ensemble de "locuteurs de référence". Les principales techniques sont : RMP (*Regression-Based Model Prediction*), *Speaker Clustering*, RSW (*Reference Speaker Weighting*) et les voix propres (ou *eigenvoices*). En reconnaissance du locuteur, la représentation relative a été appliquée dans [6], [2], [7] et plus récemment dans [4]. Ces nouvelles approches ont donné naissance à la notion d'espace de locuteurs (*speaker space*) où un modèle de locuteur est représenté généralement par une combinaison linaire des modèles de référence ce qui réduit considérablement le nombre de paramètres. Notre travail s'inscrit dans le domaine de la représentation relative en reconnaissance du locuteur [4]. Ces systèmes de représentation et de modélisation des locuteurs exploitent la position d'un locuteur par rapport à un ensemble de locuteurs de référence. Chaque locuteur peut être représenté dans cet espace et son modèle $\lambda$ approximé par la relation :

$$\lambda \approx \sum_{e=1}^{E} w_e \overline{\lambda}_e \tag{1}$$

où les $\overline{\lambda}_e$ représentent les vecteurs propres de l'espace représentatif ou voix propres, $E$ est la dimension de l'espace c'est-à-dire le nombre de locuteurs ou de voix propres.
Le calcul des $\overline{\lambda}_e$ correspond à la recherche du meilleur espace de représentation. Dans [4], les locuteurs de référence sont déterminés par clustering ou par sélection. Ensuite, on localise chaque locuteur $\lambda$ dans cet espace et on lui associe un vecteur caractéristique :

$$w = \{w_e\}_{e=1,\ldots,E}$$

Dans cet article, les locuteurs sont localisés par les modèles d'ancrage [7] [4]. Il s'agit de caractériser et de représenter un signal de parole (du locuteur $\lambda$) par rapport un ensemble de modèles de locuteurs bien appris (les modèles de référence $\overline{\lambda}_e$). Le principe repose sur le calcul d'un score de vraisemblance dans chaque direction de l'espace de référence, c'est-à-dire qu'on évalue la vraisemblance des données $X$ par rapport à chaque locuteur de référence. L'ensemble des scores constitue ainsi le vecteur des coordonnées $w$ du locuteur $\lambda$, soit :

$$w = \left[ \begin{array}{cccc} \tilde{p}(X|\bar{\lambda}_1) & \tilde{p}(X|\bar{\lambda}_2) & \ldots & \tilde{p}(X|\bar{\lambda}_E) \end{array} \right]^T \tag{2}$$

où $\widetilde{p}(X|\overline{\lambda}_e)$ est un score de vraisemblance normalisée des données $X$ (de $N_t$ trames acoustiques) sachant le modèle GMM du locuteur de référence $\overline{\lambda}_e$. Il correspond à la vraisemblance normalisée par un modèle universel :

$$\widetilde{p}(X|\overline{\lambda}_e) = \frac{1}{N_t} \log \frac{p(X|\overline{\lambda}_e)}{p(X|\lambda_{UBM})} \tag{3}$$

où $\lambda_{UBM}$ est le modèle du monde (*Universal Background Model*).

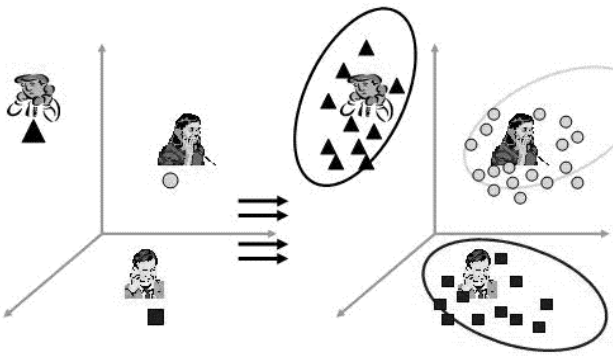

**FIG. 1:** Représentation relative des locuteurs

En plaçant plusieurs locuteurs dans l'espace, on peut évaluer la proximité spatiale et dire que tel ou tel locuteur est proche ou loin d'un autre. La similarité entre locuteurs est évaluée par des distances appliquées sur les vecteurs des coordonnées. Cependant, ces distances doivent être utilisées dans des espaces orthogonaux. Dans le cas ou l'espace de locuteurs de référence ne l'est pas, on peut réajuster les axes et faire une rotation pour retrouver l'orthogonalité et appliquer correctement les distances. En pratique, cela se traduit par l'application d'une ACP ou d'une ALD sur les vecteurs des coordonnées $w$ des locuteurs [5]. Cette orthogonalisation améliore significativement les performances des systèmes de reconnaissance mais elles restent inférieures à celles du GMM-UBM. Ceci est probablement dû au fait que cette approche est une simple approche géométrique. Son principal inconvénient est qu'elle accorde une place symétrique à l'apprentissage et au test, alors qu'en pratique, il existe souvent une asymétrie entre les occurrences d'apprentissage et de test. Dans la prochaine section, nous proposons une nouvelle représentation des locuteurs basée sur une distribution de distances. L'idée est de modéliser un locuteur par une distribution sur les "distances" mesurées dans l'espace des modèles d'ancrage.

## 3. REPRÉSENTATION PAR DISTRIBUTION SUR LES MODÈLES D'ANCRAGE

### 3.1. Principe

Le but de cette nouvelle représentation est de représenter un locuteur par une densité de probabilité qui modélise ses distances à un ensemble de locuteurs de référence [3]. En d'autres termes, au lieu de localiser un locuteur par un seul point dans l'espace de représentation, il est localisé par une distribution : c'est une référence statistique au lieu d'être géométrique (voir figure 1). Dans un tel système, nous conservons la représentation compacte des modèles d'ancrage et nous introduisons une densité de probabilité. Ce qui permettra d'une part, d'utiliser des informations a priori pour la modélisation et d'autre part, d'appliquer une mesure statistique entre l'occurrence de test et les modèles des locuteurs à reconnaître (au lieu d'une mesure géométrique). En pratique, cela consiste à représenter un locuteur $\lambda$, pour lequel on dispose d'un ou de plusieurs segments de parole, par :

$$\lambda = \mathcal{N}(\mu, \Sigma) \tag{4}$$

où $\mathcal{N}$ est une distribution gaussienne de moyenne $\mu$ et de covariance $\Sigma$. Ces paramètres sont estimés dans l'espace des coordonnées (ou l'espace des distances) à un ensemble de locuteurs de référence (figure 1).

### 3.2. Estimation des paramètres du modèle de locuteur

Soit un locuteur $\lambda$ pour lequel on dispose de $N$ segments de parole. Ces segments sont représentés dans l'espace des distances par $N$ vecteurs de coordonnées, soit :

$$W = \left( \begin{array}{ccc} w_1^\lambda & \cdots & w_N^\lambda \end{array} \right)$$

où $W$ représente l'ensemble des $N$ vecteurs de coordonnées (donnés par l'équation 2).
Ainsi, dans cette représentation, les données utilisées pour l'estimation de la densité gaussienne modélisant le locuteur ne représentent plus des vecteurs acoustiques mais des vecteurs de coordonnées de dimension $E$.

### 3.3. Estimation par maximum a posteriori

Une première possibilité pour l'estimation des paramètres de $\mathcal{N}(\mu, \Sigma)$ consiste à les estimer par maximum de vraisemblance, soit pour chaque locuteur :

$$\mu_i = \frac{1}{N} \sum_{j=1}^{N} W_{ij} \tag{5}$$

et

$$\Sigma_{ii'} = \frac{1}{N} \sum_{j=1}^{N} (W_{ij} - \mu_i)(W_{i'j} - \mu_{i'}) \tag{6}$$

où $i, i' = 1, \cdots, E$ et $W_{ij}$ est la distance du segment de parole $j$ (du locuteur $\lambda$) par rapport au locuteur de référence $\overline{\lambda}_i$.
Cependant, l'estimateur de vraisemblance n'est efficace que lorsque nous disposons de beaucoup de données ($N >>$). Dans le cas contraire (notamment dans des systèmes réels), le nombre de segments de parole disponibles n'est pas suffisant et l'estimation par maximum de vraisemblance n'est pas fiable. Une possibilité pour remédier à ce problème est d'introduire de l'information a priori. Les paramètres du locuteur seront adaptés à partir des paramètres initiaux par MAP (maximum a posteriori) [1] et la formule de ré-estimation des moyennes est donnée pour chaque locuteur par l'équation suivante :

$$\widetilde{\mu} = \frac{\widetilde{N}_0 \mu_0 + N\mu}{\widetilde{N}_0 + N} \tag{7}$$

où $\widetilde{N}_0$ est un paramètre de contrôle qui permet de donner un poids à l'information a priori.
On définit un paramètre de contrôle normalisé $\alpha$ :

$$\alpha = \frac{\widetilde{N}_0}{\widetilde{N}_0 + N} \tag{8}$$

La formule de ré-estimation des moyennes devient :

$$\widetilde{\mu} = \alpha\mu_0 + (1 - \alpha)\mu \tag{9}$$

La matrice de covariance $\widetilde{\Sigma}$ du locuteur $\lambda$ peut être aussi estimée par MAP. Cependant, il est plus robuste et plus simple de choisir une matrice de covariance commune à tous les locuteurs. Elle correspond à la matrice intra-classes des données initiales c'est-à-dire :

$$\widetilde{\Sigma} = \Sigma_0 \tag{10}$$

### 3.4. Estimation du modèle a priori

**Cas du mono-gaussian :** La matrice de toutes les données initiales $I$ correspond à l'expression suivante :

$$I = \left( \begin{array}{cccccccc} w_1^{loc_1} & \cdots & w_{N_1}^{loc_1} & w_1^{loc_2} & \cdots & w_{N_2}^{loc_2} & \cdots \end{array} \right)$$

où les vecteurs de coordonnées $w$ sont de dimension $E$ et $N_1, N_2, \cdots, N_S$ sont le nombre de segments du locuteur 1, locuteur 2, etc.
On obtient donc un nuage de points et on estime sa densité de probabilité par maximum de vraisemblance (chaque vecteur de coordonnées devient une réalisation de la gaussienne). Ce nuage est caractérisé par son vecteur moyennes $\mu_0$ et par sa matrice de covariance intra-classes $\Sigma_0$. Cette densité de probabilité $\mathcal{N}(\mu_0, \Sigma_0)$ est considérée comme le modèle a priori pour maximiser la probabilité a posteriori des nouvelles données. L'estimation par maximum de vraisemblance du vecteur des moyennes donne :

$$\mu_{0i} = \frac{1}{N_0} \sum_{j=1}^{N_0} I_{ij} \tag{11}$$

où $N_0 = \sum_{s=1}^{S} N_s$ est le nombre des segments de parole des locuteurs de développement (utilisés pour l'estimation de la distribution a priori).
En ce qui concerne la matrice de covariance $\Sigma_0$ , elle peut être estimée également par maximum de vraisemblance ou par MAP. Dans cette étude, nous utilisons une matrice de covariance intra-classes des données initiales :

$$\Sigma_{0\,ii'} = \frac{1}{N_0} \sum_{s=1}^{S} \sum_{j \in C_s} (I_{ij} - \overline{I}_{is})(I_{i'j} - \overline{I}_{i's}) \tag{12}$$

où $S$ est le nombre total des locuteurs de développement. Chaque classe est caractérisée par un ensemble $C_s$ de $N_s$ segments de parole provenant d'un même locuteur, de moyenne :

$$\overline{I}_{is} = \frac{1}{N_s} \sum_{j \in I_s} I_{ij} \tag{13}$$

et $i, i' = 1, \cdots, E$.

**Choix de la distribution a priori parmi plusieurs gaussiennes :** Cette démarche peut être étendue au cas multi-gaussien pour la densité initiale. En effet, le nuage de points des données initiales peut être modélisé par plusieurs gaussiennes $M$ (de moyenne $\mu_0^m$ et de covariance $\Sigma_0$, $m = 1, \cdots, M$).
L'obtention de plusieurs densités gaussiennes peut se faire de plusieurs façons : En utilisant une connaissance annexe (e.g. sexe des locuteurs, pour apprendre deux gaussiennes, une pour les femmes et une pour les hommes), ou bien en utilisant directement les données. Par exemple, on peut apprendre les $M$ gaussiennes du nuage par éclatement. Ensuite, la phase d'adaptation consiste à déterminer la meilleure gaussienne de l'ensemble a priori et l'adapter aux nouvelles données du locuteur $\lambda$. Ainsi, la formule de ré-estimation des moyennes, pour un locuteur $\lambda$, est toujours donnée par l'équation 9 où $\mu_0$ correspond, cette fois-ci, à la moyenne de la gaussienne qui donne le meilleur score de vraisemblance, soit :

$$\mu_0 = \arg \max_{\mu_0^m} p(W|\mu_0^m, \Sigma_0) \tag{14}$$

Précisons qu'il s'agit toujours d'une densité mono-gaussienne pour représenter un locuteur mais adaptée à partir d'une gaussienne initiale choisie dans un ensemble de gaussiennes.

### 3.5. Application à l'identification et la vérification du locuteur

Quelle que soit l'application ou la tâche visée, le module de décision est basé sur les deux processus classiques d'identification et/ou de vérification de locuteur.
En identification du locuteur, la phase de test consiste à évaluer une mesure de vraisemblance entre les coordonnées du segment de test $w_X$ et l'ensemble des modèles de locuteurs à identifier. Ainsi le locuteur reconnu correspond à (en supposant l'équi-probabilité des locuteurs) :

$$\widehat{\lambda} = \arg \max_{\lambda} p(w_X|\widetilde{\mu}, \widetilde{\Sigma}) \tag{15}$$

En vérification du locuteur, le score obtenu sur le modèle du locuteur prétendu est normalisé par le score obtenu sur le modèle du monde. Ce dernier correspond à la distribution a priori $\mathcal{N}(\mu_0, \Sigma_0)$ :

$$score = \frac{p(w_X|\widetilde{\mu}, \widetilde{\Sigma})}{p(w_X|\mu_0, \Sigma_0)} \tag{16}$$

où $X = \{\text{abonnés, imposteurs}\}$.

## 4. EVALUATION

### 4.1. Contexte expérimental

La base de données de parole utilisée est une base téléphonique de France Télécom R&D. Elle comporte 357 locuteurs. Cette base est divisée en trois sous-ensembles :
- L'ensemble $\mathcal{E}_1$ composé de 50 locuteurs (33 femmes et 17 hommes) à reconnaître utilisés comme corpus de test.
- L'ensemble $\mathcal{E}_2$ composé de 57 locuteurs (23 femmes et 34 hommes) utilisés comme corpus de développement (pour estimer la distribution a priori).
- L'ensemble $\mathcal{E}_3$ composé de 250 (127 femmes et 123 hommes) locuteurs utilisés pour estimer les modèles d'ancrage.

Les locuteurs des ensembles $\mathcal{E}_1$ et $\mathcal{E}_2$ ont suivi le même protocole de collecte. Pour chaque locuteur de $\mathcal{E}_1$ et $\mathcal{E}_2$, on dispose de 5 appels de 25 phrases réservées à l'apprentissage des modèles (soit 125 phrases d'apprentissage) et de 25 appels de 5 phrases réservées au test enregistrées durant plusieurs mois (soit 125 phrases de test). La durée moyenne des phrases est de l'ordre de 4 secondes. Les phrases de cette base sont lues et extraites du journal $Le\ Monde$. Les locuteurs de l'ensemble $\mathcal{E}_3$ disposent de moins de données mais ils sont suffisamment nombreux pour construire un espace de locuteurs. Chaque locuteur de l'ensemble $\mathcal{E}_3$ dispose d'une quinzaine de phrases, relativement courtes. Elles peuvent être des phrases lues et extraites du journal $Le\ Monde$, des suites de chiffres, des réponses à des questions, etc. Les conditions de prise de son de l'ensemble des locuteurs varient d'une phrase à une autre, mais la qualité générale des enregistrements est de type Réseau Téléphonique Commuté (RTC). Par ailleurs, l'espace des paramètres acoustiques est composé de 42 coefficients. A chaque trame, on associe un vecteur de représentation acoustique composé de l'énergie temporelle de la trame et des 13 premiers MFCC après soustraction cepstrale. A cela, on rajoute leurs dérivées premières et secondes. Les locuteurs sont modélisés par 256 gaussiennes. L'apprentissage des modèles GMM des locuteurs est un apprentissage incrémental à partir du modèle

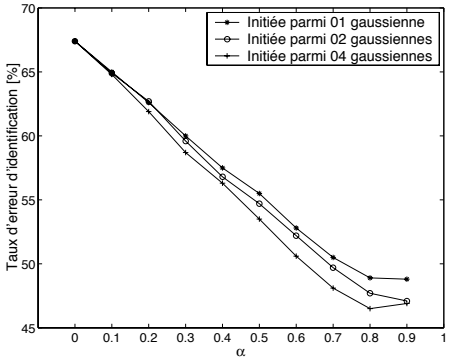


**FIG. 2:** Performances d'identification : choix de la distribution initiale parmi plusieurs gaussiennes (pour 04 secondes d'apprentissage)

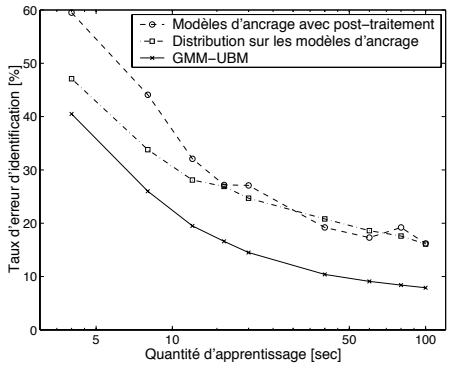


**FIG. 3:** Performances d'identification par distribution sur les modèles d'ancrage (avec choix de la distribution a priori parmi 04 gaussiennes et $\alpha = 0.9$)

du monde UBM (il correspond à un apprentissage MAP avec un choix particulier de l'a priori).

### *4.2. Evaluation*

Nous avons évalué cette nouvelle représentation en identification et en vérification du locuteur sur les 50 locuteurs de l'ensemble $\mathcal{E}_1$ (tests croisés homme/femme). Les locuteurs initiaux correspondent aux 57 locuteurs du corpus de développement $\mathcal{E}_2$. Les locuteurs sont représentés par des vecteurs de distances de dimension 250. Sur la figure 2, nous avons représenté les variations des taux d'erreur en fonction des valeurs de $\alpha$ (cf. équation 8) et dans le cas où nous avons une distribution a priori choisie parmi 1, 2 ou 4 gaussiennes. La figure 2 montre que l'introduction des connaissances a priori permet d'apporter une nette amélioration par rapport à une distribution sans a priori ($\alpha = 0$ soit $\widetilde{\mu} = \mu$). Dans le cas où $\alpha = 1$ (soit $\widetilde{\mu} = \widehat{\mu}_0$), tous les locuteurs auraient quasiment le même modèle (choisi parmi l'ensemble des distributions initiales).

En identification du locuteur, la figure 3 montre que la distribution sur les modèles d'ancrage apporte une amélioration sur l'approche géométrique [5]. En revanche en vérification du locuteur, la figure 4 montre que si on dispose de très peu de données d'apprentissage, la nouvelle approche donne des résultats similaires voir meilleurs que ceux du GMM-UBM et avec des modèles de locuteurs plus compactes.

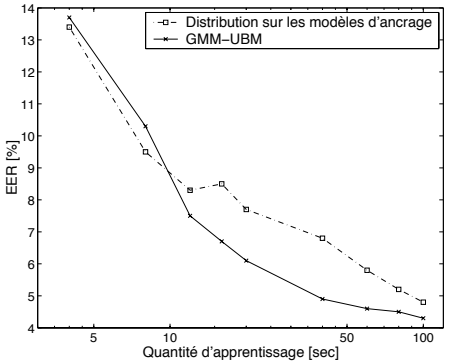


**FIG. 4:** Performances de vérification par distribution sur les modèles d'ancrage (avec choix de la distribution a priori parmi 04 gaussiennes et $\alpha = 0.9$)

## 5. CONCLUSION

Dans cet article, nous avons proposé une nouvelle représentation des locuteurs basée sur une distribution des distances par rapport à des modèles de locuteurs de référence. Nous avons appliqué cette représentation en identification et en vérification du locuteur. Les évaluations de cette nouvelle technique ont montré que les performances de vérification de locuteur sont comparables à celles des GMM lorsque nous disposons de peu de données d'apprentissage. Par ailleurs, le choix de la distribution a priori parmi plusieurs gaussiennes améliore les performances mais le gain est peu significatif au-delà de 4 gaussiennes.
De nombreuses perspectives permettent de prolonger ce travail notamment l'utilisation de cette représentation compacte dans les tâches particulières requises par un système d'indexation, comme la segmentation en locuteurs ou le regroupement en locuteurs.

## RÉFÉRENCES

[1] J.-L. Gauvain and C.-H. Lee. Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains. *IEEE Transactions on Speech and Audio Processing*, 2(2) :291–298, 1994.

[2] R. Kuhn, J-C. Junqua, P. Nguyen, and N. Niedzielski. Rapid speaker adaptation in eigenvoice space. *IEEE Transactions on Speech and Audio Processing*, 8(6) :695–707, November 2000.

[3] Yassine Mami. *Reconnaissance de locuteurs par localisation dans un espace de locuteurs de référence*. PhD thesis, ENST Paris, 2003.

[4] Yassine Mami and Delphine Charlet. Speaker identification by location in an optimal space of anchor models. In *ICSLP*, volume 2, pages 1333–1336, 2002.

[5] Yassine Mami and Delphine Charlet. Speaker identification by anchor models with PCA/LDA post-processing. In *ICASSP*, volume 1, pages 180–183, 2003.

[6] T. Merlin, J.-F. Bonastre, and C. Fredouille. Non directly acoustic process for costless speaker recognition and indexation. In *COST-254*, 1999.

[7] D.E. Sturim, D.A. Reynolds, E. Singer, and J.P. Campbell. Speaker indexing in large audio databases using anchor models. In *ICASSP*, pages 429–432, 2001.