Identification des locuteurs par regroupement hiérarchique ascendant et modèles d'ancrage

Yassine Mami, Delphine Charlet

France Télécom R&D
DIH/IPS, 2 av. Pierre Marzin 22307 Lannion Cedex - FRANCE
Tél.: ++33 (0)2 96 05 27 11 - Fax: ++33 (0)2 96 05 35 30
Mél: {yassine.mami, delphine.charlet}@rd.francetelecom.com

ABSTRACT

The process of speaker recognition is generally based on modeling the characteristics of each speaker. An interesting method for modeling consists in representing a new speaker, not in an absolute manner, but relatively to a set of well trained speakers. Each speaker is represented by its location in an optimal space of eigen or virtual voices. We hope that the relative position of a speaker in this space of virtual speakers is invariant whatever the conditions of sound recording and the content of sentences are. This paper describes a representation space built by clustering speakers and how we can locate a speaker by using anchor models. The paper also presents experimental results and compares with GMM. We show that clustering gives an optimal space. If we have a few amount of training data, we also show that our system gives better performances.

1. INTRODUCTION

Depuis les premiers travaux dédiés à la reconnaissance automatique du locuteur, de nombreuses approches ont été proposées dans la littérature - approche vectorielle, connexionniste, prédictive, statistique, etc. De ce large panel, seule l'approche statistique demeure au premier plan des systèmes de la reconnaissance automatique des locuteurs des récentes années. En effet, la modélisation par un mélange de gaussiennes (GMM) [3] fournit les meilleurs taux de reconnaissance en mode indépendant du texte et constitue l'état de l'art en la matière. Le champ d'application est très large, allant des applications domestiques aux applications militaires, en passant par des applications judiciaires. Malheureusement, les performances se dégradent considérablement si les données d'apprentissage sont insuffisantes. Or dans la plupart de ces applications, la phase d'enrôlement doit être très brève (de l'ordre de quelques secondes de parole). Il s'agit d'estimer avec très peu de données un modèle suffisamment robuste du locuteur pour permettre la reconnaissance du locuteur même lorsque les conditions de prise de son ou le contenu phonétique de la phrase de test ne sont pas les mêmes qu'à l'apprentissage. Pour tenter de remédier à ce problème, une perspective intéressante de modélisation consiste à représenter un nouveau locuteur, non plus de façon absolue, mais relativement à un ensemble de locuteurs bien appris [1] [2]. Chaque locuteur est représenté par sa localisation dans un espace optimal de voisins.
Dans cette optique, des nouveaux systèmes de représentation et de modélisation des locuteurs utilisant les voix propres (eigenvoices) sont apparus. En général, ces systèmes se divisent en trois modules. Dans le premier, un espace de représentation est construit tandis que le deuxième est dédié à la localisation des nouveaux locuteurs dans cet espace. Selon la technique de construction de l'espace et de localisation, plusieures variantes peuvent être explorées. Dans le dernier module, on effectue un test de reconnaissance.
Dans ce papier, on s'est proposé d'étudier un système où l'espace de représentation est généré par regroupement hiérarchique ascendant (ou clustering) et la localisation se fait par la méthode des modèles d'ancrage (Anchor Models) [7].

2. PRINCIPE D'IDENTIFICATION DES LOCUTEURS PAR LOCALISATION DANS UN ESPACE OPTIMAL DES VOISINS

La représentation des locuteurs par localisation est une nouvelle technique de reconnaissance et d'adaptation des locuteurs. L'une des voies les plus intéressantes est celle des voix propres [1]. Il s'agit de représenter un locuteur relativement à d'autres locuteurs. La motivation principale repose sur le fait que la dimension de ces problèmes est très grande par rapport à celle qu'on peut estimer de façon fiable. Ainsi, plutôt que d'estimer les nombreux paramètres d'un modèle absolu du locuteur, on cherche à estimer des paramètres moins nombreux d'un modèle relatif à d'autres locuteurs estimés de façon absolue avec suffisamment de données. Par conséquent, on ne modélise plus de façon absolue mais relativement à des locuteurs de référence. Ces derniers peuvent être "virtuels" et sont supposés les plus représentatifs de l'espace de l'ensemble des locuteurs. Ainsi, on associe à chaque locuteur λ un vecteur caractéristique $\{w_e\}$ $(e = 1, \ldots, E)$ où w_e est la caractéristique propre de chaque locuteur par rapport au locuteur virtuel $\bar{\lambda}_e$ et E est le nombre de locuteurs virtuels. Le système d'identification de locuteurs par voix virtuelles se déroule en trois phases :

1. Construction d'un espace de locuteurs virtuels. Cette étape consiste à créer une base de locuteurs virtuels (ou un nouvel espace de représentation) : on dispose des modèles GMM d'un ensemble de locuteurs bien appris qui serviront à la construction de l'espace représentatif, leurs paramètres étant rangés dans une matrice représentant l'espace initial ; ensuite, soit on construit un espace propre (voix propres ou orthogonales) soit on réalise un regroupement hiérarchique des locuteurs.
Les voix propres s'obtiennent en appliquant des algorithmes de réduction de dimensionnalité de l'espace, telle que la PCA (Analyse en Composantes Principales), la PPCA (Probabilistic PCA) ou la LDA (Analyse Linéaire

Discriminante). Il s'agit de réajuster le nuage de points de l'espace et rechercher les axes principaux les plus discriminants.
Le regroupement hiérarchique, quant à lui, est une alternative aux voix propres. Son principe consiste à agréger deux à deux les locuteurs les plus proches (cf. 3.1).

2. Placement des locuteurs à reconnaître. La deuxième étape consiste à localiser des nouveaux locuteurs dans cet espace de représentation : cette localisation est réalisée soit par une simple projection (qui n'a de sens que si l'espace est orthogonal), soit en maximisant la vraisemblance du problème (MLED) [1] ou encore en utilisant la technique des modèles d'ancrage [7]. Chaque locuteur est associé à un vecteur de caractéristiques propres $w = [w_1, \ldots, w_E]^T$, qui représentent ses coordonnées dans l'espace des locuteurs virtuels. Dans le cas où l'espace est construit par voix propres, chaque locuteur λ est approximé par une somme pondérée de l'ensemble des voix propres $\bar{\lambda}_e$ c'est à dire :

$$\lambda = \sum_{e=1}^{E} w_e \bar{\lambda}_e \quad (1)$$

Dans une autre approche, les locuteurs peuvent aussi être placés par MLED ou par les modèles d'ancrage. Le principe de ces derniers consiste à représenter et caractériser un signal de parole d'un locuteur par rapport à un ensemble de locuteurs (ou modèles d'ancrage). Un signal de parole d'un locuteur est représenté par un vecteur caractéristique de E composantes, chacune d'elles est un score de vraisemblance des données x sachant les modèles des locuteurs virtuels $\bar{\lambda}_e$:

$$w = \left[\; p(x|\bar{\lambda}_1) \quad p(x|\bar{\lambda}_2) \quad \ldots \quad p(x|\bar{\lambda}_E) \; \right]^T \quad (2)$$

3. Identification des locuteurs. La représentation intuitive d'un locuteur par sa localisation dans l'espace de représentation présume que plus des locuteurs sont similaires plus leurs points de projection sont proches et la distance entre eux est petite. Donc, pour exploiter la notion du voisinage et évaluer la proximité dans l'espace, on utilise une métrique entre les coordonnées des locuteurs. Soit un locuteur inconnu X représenté par $\lambda_X = \{w_e^X\}_{e=1,\ldots,E}$, le locuteur reconnu : $\hat{R}$ est celui dont le modèle de référence $\lambda_R = \{w_e^R\}_{e=1,\ldots,E}$ donne la plus petite distance :

$$\hat{R} = \arg\min_R d(\lambda_X, \lambda_R) \quad (3)$$

Les coefficients de localisation des locuteurs peuvent aussi donner lieu à une meilleure estimation des modèles [1].

3. SYSTÈME DE RECONNAISSANCE ÉTUDIÉ

On s'est proposé d'étudier un système de reconnaissance où l'espace représentatif est généré par regroupement hiérarchique et les locuteurs sont placés par la technique des Modèles d'ancrage. Ainsi, notre système d'identification de locuteurs par voix virtuelles se déroule en trois phases :

1. Construction de l'espace représentatif par regroupement et génération des E locuteurs virtuels.
2. Placement des locuteurs à reconnaître relativement aux modèles d'ancrage. Chaque locuteur λ_i est représenté de la façon suivante :
   $$\lambda_i = \{p(x_i|\bar{\lambda}_e)\}$$
   où $p(x_i|\bar{\lambda}_e)$ représente la vraisemblance des données x_i du locuteurs λ_i sachant les modèles des locuteurs virtuels.
3. Test de reconnaissance : placement de chaque nouveau locuteur dans l'espace représentatif. Le modèle de référence le plus proche est donc l'identité présumée du locuteur.

Notons que la première phase est un traitement off-line. Les paragraphes suivants décrivent chaque étape du système.

3.1. Regroupement hiérarchique ascendant des locuteurs

La réduction de l'espace par un algorithme de réduction de dimensionnalité constitue une des approches les plus simples et les plus intuitives. En revanche, les sous-espaces recherchés ne sont pas significatifs s'ils n'ont pas été générés par un grand nombre de données. Les matrices de covariance, d'inter-classes et d'intra-classes risquent d'être très mal estimées. Indépendamment de ces problèmes d'estimation, il est intéressant de construire un espace de locuteurs virtuels pour lesquels on dispose de trames de parole associées à chacun d'eux. L'intérêt majeur est qu'on peut travailler directement sur les trames en appliquant des métriques sur les coefficients acoustiques ou bien en les projetant dans un autre espace acoustique.
Le regroupement hiérarchique répond à ce type de problème : il s'agit de créer, à chaque étape, une partition obtenue en agrégeant deux à deux les locuteurs les plus proches. Le regroupement est généralement utilisé, d'une part, pour segmenter un signal de parole de façon à ce que chaque cluster final ne contienne que les trames d'un seul locuteur et, d'autre part, pour regrouper les trames d'un même locuteur dans un cluster [5] [6]. Dans le cadre de ce travail, le regroupement hiérarchique est utilisé uniquement pour regrouper deux à deux les locuteurs les plus proches. Le but est de trouver un ensemble qui soit le plus représentatif de tous les locuteurs. L'algorithme se déroule de la façon suivante :

1. Initialiser le tableau de distances $\mathcal{D}$ entre les locuteurs.
2. Rechercher les deux plus proches locuteurs et les regrouper en un nouvel élément.
3. Construire un nouveau tableau de distances et aller à (2). Réitérer ensuite le processus jusqu'à n'avoir que E locuteurs virtuels.

Calcul des distances : pour calculer la distance $d(i,j)$ entre un locuteur i et un locuteur j, on évalue la vraisemblance $p(x_i|\lambda_j)$ de l'ensemble des trames acoustiques x_i par rapport au modèle λ_j du locuteur j.
Sachant que chaque locuteur est modélisé par une somme pondérée de gaussiennes [3], la distance $d(i,j)$ s'écrit (après normalisation) :

$$d(i,j) = \frac{1}{N_i} \log \frac{p(x_i|\lambda_{UBM})}{p(x_i|\lambda_j)} + \frac{1}{N_j} \log \frac{p(x_j|\lambda_{UBM})}{p(x_j|\lambda_i)} \quad (4)$$

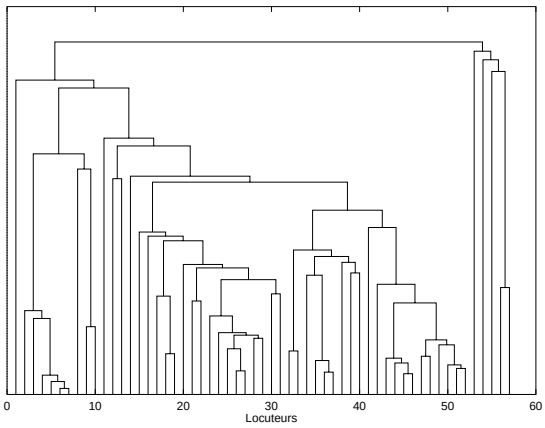

Figure 1: Exemple d'un regroupement hiérarchique de 57 locuteurs

où N_i, N_j représentent, respectivement, le nombre de trames acoustiques du locuteur i et du locuteur j. λ_{UBM} est le modèle du monde (Universal Background Model) à partir duquel les modèles des locuteurs ont été appris.
Cette distance, appelée rapport de vraisemblance croisée, est symétrique $d(i,j) = d(j,i)$; en revanche $d(i,i) \neq 0$ et l'inégalité triangulaire n'est pas forcément vérifiée. Notons que plus la distance $d(i,j)$ est petite, plus les locuteurs i et j sont proches.

Construction du dendrogramme : après avoir estimé la matrice des distances $\mathcal{D}$, on recherche les deux locuteurs les plus proches qui correspondent à la plus petite distance. Le locuteur hybride est modélisé à partir des données des deux locuteurs et son modèle GMM est estimé. Si i, j, k sont trois locuteurs, et si i et j sont regroupés en un seul élément noté "ij", la distance de ce groupement au locuteur k s'écrit :

$$d(k,ij) = \frac{1}{N_k} \log \frac{p(x_k|\lambda_{UBM})}{p(x_k|\lambda_{ij})} + \frac{1}{N_i + N_j} \log \frac{p(x_{ij}|\lambda_{UBM})}{p(x_{ij}|\lambda_k)} \quad (5)$$

où x_{ij} est la concaténation des données x_i et x_j. λ_{ij} est le modèle du locuteur hybride "ij".
Le processus est réitéré jusqu'à un niveau de partition donné. La figure 1 montre un exemple de regroupement hiérarchique appliqué sur 57 locuteurs.

3.2. Les modèles d'ancrage

Le regroupement hiérarchique des locuteurs ne fournit pas un espace orthogonal. La localisation des locuteurs dans cet espace ne peut pas se faire par une simple projection. Les coefficients caractéristiques sont estimés par la MLED ou par les modèles d'ancrage. Ces derniers ont été utilisés pour la détection et l'indexation des locuteurs. Il s'agit de représenter un locuteur par rapport à un ensemble des modèles d'ancrage. Dans notre étude, ces modèles correspondent à l'espace des E locuteurs virtuels construit par regroupement. Chaque locuteur est modélisé de la façon suivante :

$$w = \left[\begin{array}{cccc} \tilde{p}(x|\bar{\lambda}_1) & \tilde{p}(x|\bar{\lambda}_2) & \ldots & \tilde{p}(x|\bar{\lambda}_E) \end{array} \right]^T \quad (6)$$

où $\tilde{p}(x|\bar{\lambda}_e)$ est la vraisemblance normalisée des données x (de N trames acoustiques) sachant le modèle GMM du locuteur i :

$$\tilde{p}(x|\bar{\lambda}_e) = \frac{1}{N} \log \frac{p(x|\bar{\lambda}_e)}{p(x|\lambda_{UBM})} \quad (7)$$

3.3. Identification des locuteurs

Après avoir représenté tous les locuteurs par des points dans l'espace, on évalue la proximité entre eux. Le modèle de référence le plus proche du locuteur inconnu constitue le locuteur reconnu. Il est donc nécessaire de définir une distance entre deux points ou deux locuteurs.
Soit R un locuteur à reconnaître et T un locuteur de test représentés respectivement par les vecteurs $[r_1, \ldots, r_E]^T$ et $[t_1, \ldots, t_E]^T$. On utilise par exemple les métriques suivantes :
la distance Hamming : $d_1(R,T) = \sum_{i=1}^{E} |r_i - t_i|$;
l'angle entre les deux vecteurs des coordonnées : $\delta(R,T) = \arccos \left[\frac{r^T t}{\sqrt{r^T r \cdot t^T t}} \right]$.

4. EVALUATION

4.1. Contexte expérimental

Cette partie présente les expériences réalisées dans le but d'évaluer l'identification du locuteur par localisation, en mode indépendant du texte. L'espace des paramètres acoustiques est composé de 27 coefficients. A chaque trame, on associe un vecteur de représentation acoustique composé de l'énergie temporelle de la trame et des 8 premiers MFCC. A cela, on rajoute leurs dérivées premières et secondes. Les locuteurs sont modélisés par 16 gaussiennes. L'apprentissage des modèles GMM des locuteurs est un apprentissage incrémental [4]. Il s'agit d'un algorithme itératif qui adapte un modèle de référence λ_{UBM}, indépendant du locuteur, aux données d'apprentissage du locuteur considéré. Il correspond à un apprentissage bayésien avec un choix particulier des paramètres a priori. La base de donnée de parole utilisée est une base interne à France Télécom R&D. Elle comporte 107 locuteurs. Cette base est divisée en deux sous-ensembles :

- L'ensemble $\mathcal{E}_1$ composé de 57 locuteurs utilisés pour le regroupement.
- L'ensemble $\mathcal{E}_2$ composé de 50 locuteurs à identifier (33 femmes et 17 hommes).

Pour chaque locuteur de l'ensemble $\mathcal{E}_2$, on dispose de 25 phrases réservées à l'apprentissage des modèles enregistrées au cours d'un seul appel et de 125 phrases réservées au test enregistrées au cours de 25 appels durant plusieurs mois. Le modèle λ_{UBM} est appris à partir d'un ensemble distinct de 527 locuteurs (approximativement 5 heures de parole) issu de la même base de données. La durée moyenne des phrases est de l'ordre de 8 secondes, soit environ 500 trames, incluant des silences plus ou moins longs. On effectue un test par phrase, soit plus de 6000 tests. L'intervalle de confiance à 95% est de $\pm 1.25\%$. Les phrases de cette base sont lues et extraites du journal $Le\ Monde$. Les conditions de prise de son varient d'une phrase à une autre, mais la qualité générale des enregistrements est de type Réseau Téléphonique Commuté (RTC).

4.2. Résultats et discussion

Influence de la quantité d'apprentissage : sur la figure 2, on a représenté les variations des taux d'identification des 50 locuteurs de l'ensemble $\mathcal{E}_2$ obtenus avec les deux

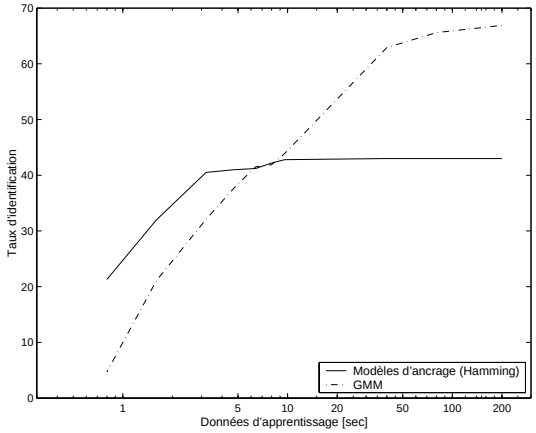

Figure 2: Variations des taux d'identification en fonction des données d'apprentissage

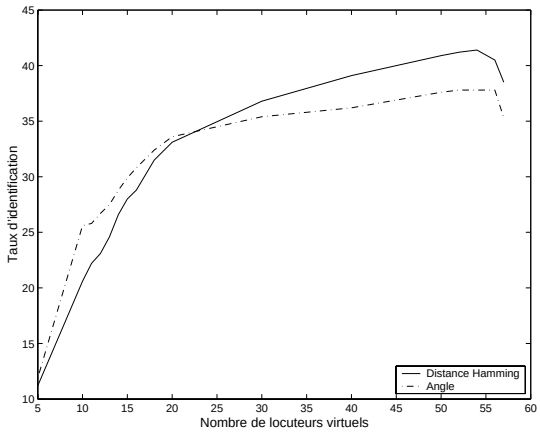

Figure 3: Variations des taux d'identification en fonction du nombre de locuteurs virtuels

approches GMM et modèles d'ancrage sur regroupement hiérarchique, en fonction de la quantité d'apprentissage. Cette figure donne un aperçu des variations des taux d'identification que l'on peut attendre en fonction de la quantité de données d'apprentissage. On distingue deux régions de cette courbe : si les données d'apprentissage sont inférieures à 6 secondes, l'identification par regroupement hiérarchique et modèles d'ancrage permet d'obtenir des performances meilleures qu'une modélisation GMM. En revanche, si la quantité de données est suffisamment grande, les taux d'identification GMM grimpent très rapidement pour atteindre les 66.9% contre 43% avec 200 secondes de données. Les taux d'identification par modèles d'ancrage de cette expérience ont été obtenus avec la distance Hamming pour un regroupement hiérarchique de 54 locuteurs virtuels.

Influence du nombre de locuteurs virtuels : la figure 3 représente les variations des taux d'identification des 50 locuteurs de l'ensemble $\mathcal{E}_2$ obtenus par regroupement hiérarchique et modèles d'ancrage. La proximité entre les locuteurs est évaluée par la distance Hamming et l'angle entre leurs vecteurs de coordonnées. La quantité de données qui a servi à placer les locuteurs à identifier est au voisinage des 6 secondes de parole. Dans les cas où les locuteurs sont regroupés en un grand nombre de locuteurs virtuels (supérieur à 25), la distance Hamming conduit à de meilleures performances d'identification. Dans le cas contraire, l'angle entre les coordonnées des locuteurs semble être la bonne métrique discriminante comme le montre la figure 3. La figure 3 montre aussi une progression significative des taux d'identification pour un espace de moins de 20 locuteurs virtuels. A ce point d'inflexion, les taux d'identification continuent à grimper jusqu'à une valeur optimale (54 locuteurs virtuels) significativement meilleure que le taux obtenu en conservant tous les locuteurs de l'espace. On constate ensuite une dégradation des performances. Ainsi, le regroupement hiérarchique a permis de déterminer un espace optimal pour la reconnaissance. Compte-tenu de la métrique qui donne à tous les modèles d'ancrage la même importance, il apparaît intéressant de regrouper certains modèles très proches pour obtenir un pavage plus homogène de l'espace.

5. CONCLUSION

Dans ce papier, nous avons présenté un système d'identification des locuteurs par regroupement hiérarchique ascendant et modèles d'ancrage. Le regroupement hiérarchique permet d'obtenir un espace optimal des voisins. Il offre la possibilité d'associer des trames de parole à chaque voix virtuelle et d'avoir plusieurs niveaux de précision pour chacune d'elle. Dans notre étude, la légère amélioration des taux de reconnaissance, pour une partition donnée par rapport à la partition initiale, nous incite à continuer dans cette voie et augmenter le nombre des locuteurs nécessaires pour la construction de l'espace. Les modèles d'ancrage conduisent à une bonne localisation des locuteurs. Les taux d'identification sont nettement meilleurs que ceux des GMM lorsque nous disposons de peu de données d'apprentissage.

BIBLIOGRAPHIE

[1] R. Kuhn, P. Nguyen, J-C. Junqua, L. Goldwasser, N. Niedzielski, S. Fincke, K. Field, and M. Contolini. Eigenvoices for speaker adaptation. In *ICSLP98*, 1998.

[2] T. Merlin, J.-F. Bonastre, and C. Fredouille. Non directly acoustic process for costless speaker recognition and indexation. In *COST-254 International Workshop on Intelligent Communication Technologies and Applications, with emphasis on Mobile Communications*, 1999.

[3] D.A. Reynolds. Speaker identification and verification using gaussian mixture speaker models. *Speech Communication*, 17(1):91–108, 1995.

[4] D.A. Reynolds. Comparison of background normalization methods for text-independent speaker verification systems. In *Eurospeech'97*, pages 963–966, 1997.

[5] D.A. Reynolds, E. Singer, B.A. Carlson, G.C. O'Leary, J.J. McLaughlin, and M.A. Zissman. Blind clustering of speech utterances based on speaker and language characteristics. In *ICASSP98*, 1998.

[6] A. Solomonoff, A. Mielke, M. Schmidt, and H. Gish. Clustering speakers by their voices. In *ICASSP98*, pages 757–760, 1998.

[7] D.E. Sturim, D.A. Reynolds, E. Singer, and J.P. Campbell. Speaker indexing in large audio databases using anchor models. In *ICASSP2001*, pages 429–432, 2001.