REPRESENTATION DU LOCUTEUR PAR MODELES D'ANCRAGE POUR L'INDEXATION DE DOCUMENTS AUDIO

Mikaël Collet [(1)(2)], *Delphine Charlet* [(1)], *Frédéric Bimbot* [(2)]

(1) France Telecom R&D - TECH/SSTP - 2 av. Pierre Marzin - 22307 Lannion Cedex - FRANCE
{mikael.collet, delphine.charlet}@rd.francetelecom.com
(2) IRISA (CNRS & INRIA) - Campus de Beaulieu - 35042 Rennes Cedex - FRANCE
frederic.bimbot@irisa.fr

ABSTRACT

This paper presents a speaker indexing system of audio document entirely based on the anchor models approach. Evaluation is done on the audio database of the ESTER evaluation campaign for the rich transcription of French broadcast news. Results show that speaker indexing performance is improved when a speaker clustering process is performed and that a weighted measure of similarity, used in the speaker tracking process, can overcome some errors of the clustering process. The use of anchor models is particulary suitable for speaker indexing because the computational burden to search a speaker in an audio document is very low and performances are equivalent to those of a speaker indexing system using the classical speaker representation in the acoustic space (Gaussian model for speaker segmentation and clustering, Gaussian mixture model for speaker tracking).

1. INTRODUCTION

La représentation du locuteur par les modèles d'ancrage consiste à modéliser un locuteur relativement à un ensemble de locuteurs de référence appelés modèles d'ancrage. Ces modèles d'ancrage peuvent être vus comme une représentation du signal de parole (comme les coefficients cepstraux) particulièrement adaptée pour la caractérisation du locuteur. Cette représentation du locuteur est également appropriée pour la recherche de locuteurs cibles au sein de grandes bases de données audio [1]. En effet, dans le cadre de l'utilisation des modèles d'ancrage, la majorité des calculs est effectuée lors d'une phase de pré-traitement de la base de données (segmentation et regroupement en locuteurs), indépendamment des locuteurs à rechercher. Le coût de calcul pour rechercher des nouveaux locuteurs cibles est alors très faible.
Cet article, qui présente un système d'indexation en locuteurs de documents audio entièrement basé sur la représentation du locuteur par les modèles d'ancrage, est organisé de la manière suivante. Le paragraphe 2, présente le concept des modèles d'ancrage et les méthodes permettant de comparer des locuteurs dans l'espace des modèles d'ancrage. Ensuite les deux processus de pré-traitement de la base de données audio sont décrits : la segmentation en locuteurs (paragraphe 3) et le regroupement en locuteurs (paragraphe 4). Puis le paragraphe 5 présente différents processus de suivi de locuteurs détaillés dans des travaux récents [2] [3]. Enfin, au cours du paragraphe 6, les performances de ces processus de suivi de locuteurs sont évaluées sur la base de donnée d'émissions radiophoniques de la campagne d'évaluation ESTER.

2. REPRÉSENTATION DU LOCUTEUR PAR MODÈLES D'ANCRAGE

Le locuteur constitue une clé d'indexation très pertinente pour l'archivage de documents audio (archives radiophoniques, messageries vocales...). Cependant, les techniques actuelles pour la représentation du locuteur (modélisation par un mélange de gaussiennes - GMM) sont assez gourmandes en temps de calcul et en taille mémoire et ne sont pas adaptées à l'indexation de grandes bases de données audio.
C'est pourquoi cet article propose un système d'indexation en locuteurs utilisant les modèles d'ancrage, permettant ainsi de diminuer le nombre de paramètres pour la représentation du locuteur tout en conservant les performances d'une modélisation par mélange de gaussiennes.

2.1. Principe

Des recherches récentes [1] se sont orientées vers une représentation relative du locuteur. Cette modélisation consiste à projeter un énoncé d'un locuteur dans un espace de locuteurs de référence. Le locuteur n'est plus représenté de façon absolue mais relativement à un ensemble de locuteurs caractérisés par des modèles GMM. Ces modèles sont appelés modèles d'ancrage.
Le locuteur est caractérisé par un vecteur défini comme l'ensemble des rapports de vraisemblance des données du locuteur issues des différents modèles d'ancrage. Ce vecteur est appelé *Speaker Characterization Vector* (SCV) et dénoté $\widetilde{X}$.

$$\widetilde{X} = \begin{bmatrix} \widehat{s}(X|\overline{\lambda}_1) \\ \widehat{s}(X|\overline{\lambda}_2) \\ \vdots \\ \widehat{s}(X|\overline{\lambda}_E) \end{bmatrix} \quad (1)$$

où $\widehat{s}(X|\overline{\lambda}_e)$ est le logarithme du rapport de vraisemblance des données X (de N vecteurs acoustiques) pour le modèle GMM du locuteur de référence $\overline{\lambda}_e$ (modèles d'ancrage) relativement à un modèle indépendant du locuteur dit "modèle du monde" (ou UBM - Universal Background Model) :

$$\widehat{s}(X|\overline{\lambda}_e) = \frac{1}{N} \log \frac{p(X|\overline{\lambda}_e)}{p(X|\lambda_{UBM})} \quad (2)$$

2.2. Mesures de similarité dans l'espace des modèles d'ancrage

Les mesures de similarité précédemment utilisées pour comparer des SCV dans l'espace des modèles d'ancrage sont la mesure de similarité euclidienne et la mesure de similarité angulaire. Une nouvelle mesure de similarité basée sur le coefficient de corrélation est également proposée dans [4].
Soient X and Y deux segments de parole, $\widetilde{X}$ et $\widetilde{Y}$ leur SCV. La mesure de similarité de corrélation est définie par :

$$\rho(\widetilde{X}, \widetilde{Y}) = 1 - R(x, y) \tag{3}$$

où $R(x, y)$ est le coefficient de corrélation entre les composantes des deux SCV, les composantes étant considérées comme la réalisation de deux variables aléatoires x et y : $R(x, y) = \frac{C_{xy}}{\sigma_x \sigma_y}$
Cette nouvelle mesure de similarité s'avère expérimentalement plus robuste à la variabilité intra-locuteurs que les mesures de similarité euclidienne et angulaire [4].

2.3. Approche statistique dans l'espace des modèles d'ancrage

Le principe de l'approche statistique proposée dans [5] consiste à modéliser les différents énoncés d'un même locuteur par une distribution normale dans l'espace des modèles d'ancrage. Il s'agit d'un modèle statistique modélisant la variabilité intra-locuteur. En pratique, l'approche proposée dans [5] consiste à représenter un locuteur X ayant prononcé un ou plusieurs énoncés par une distribution notée $\hat{X}$ dans l'espace des SCV :

$$\hat{X} = \mathcal{N}(\mu_X, \Sigma_X) \tag{4}$$

où $\mathcal{N}$ est une distribution gaussienne de vecteur moyen μ_X et de matrice de covariance Σ_X. La matrice de covariance est la même pour tous les locuteurs et est égale à Σ_0. Le vecteur moyen du modèle du locuteur est adapté à partir d'un vecteur moyen μ_0 par une version simplifié de l'adaptation par maximum a posteriori (MAP). Les paramètres de la distribution a priori (μ_0 et Σ_0) sont estimés à partir d'un corpus de développement selon le processus décrit dans [5]. En vérification du locuteur, le score d'un segment de test Y pour un locuteur X est un log-rapport de vraisemblance entre le modèle du locuteur X et le modèle a priori $\mathcal{N}(\mu_0, \Sigma_0)$.

$$LLR(\widetilde{Y}|\hat{X}) = \log \frac{p(\widetilde{Y}|\mu_X, \Sigma_0)}{p(\widetilde{Y}|\mu_0, \Sigma_0)} \tag{5}$$

Une extension de cette approche statistique, proposée dans [5], consiste à estimer un modèle $\hat{Y}$ à partir du segment de test Y et à utiliser les données d'apprentissage $\widetilde{X}$ pour symétriser la mesure :

$$L(\widetilde{Y}, \widetilde{X}) = \frac{LLR(\widetilde{Y}|\hat{X}) + LLR(\widetilde{X}|\hat{Y})}{2} \tag{6}$$

3. SEGMENTATION EN LOCUTEURS

Le processus de segmentation en locuteur consiste à segmenter un document audio en segments homogènes de longueur raisonnable ayant été prononcés par un seul locuteur. Ce processus s'effectue sans aucune connaissance a priori sur les locuteurs présents dans le document. Le processus de segmentation en locuteurs utilisé dans cet article est basé sur une technique qui consiste à détecter des ruptures statistiques dans le signal audio, correspondant à des changements de locuteurs [6]. Dans notre cas, le processus de segmentation en locuteurs utilise la représentation du locuteur par les modèles d'ancrage et le rapport de vraisemblance généralisé calculé dans l'espace des coefficients acoustiques est remplacé par la mesure de similarité de corrélation définie par l'équation 3. Des expériences préliminaires ont montré que les mesures de similarité dans l'espace des modèles d'ancrage, notamment la mesure de corrélation, obtiennent de meilleures performances en segmentation en locuteurs que le log-rapport de vraisemblance dans l'espace des modèles d'ancrage (équation 6). Les travaux présentés dans [2] montrent également que la mesure de similarité de corrélation dans l'espace des modèles d'ancrage permet de mieux détecter les changement de locuteurs que le rapport de vraisemblance généralisé dans l'espace des coefficients acoustiques.

4. REGROUPEMENT EN LOCUTEURS

Le processus de regroupement en locuteurs consiste à regrouper au sein d'une même classe les segments supposés prononcés par le même locuteur. Ce processus s'effectue sans aucune connaissance a priori sur les locuteurs présents dans le document. L'algorithme de regroupement en locuteurs présenté dans cet article est basé sur une approche par *single-linkage* [7] et s'effectue en trois étapes :

1. Calcul d'une mesure de similarité de corrélation (équation 3) entre tous les segments issus du processus de segmentation en locuteurs.
2. Regroupement dans une même classe d'un segment et de son plus proche voisin au sens de la mesure de similarité.
3. Fusion des classes dont l'intersection n'est pas vide.

5. SUIVI DE LOCUTEURS

Le processus de suivi de locuteurs au sein d'un document audio consiste à rechercher les intervalles de temps au cours desquels un locuteur cible a parlé. La majorité des systèmes présentés dans la littérature utilisent un processus de vérification du locuteur où les énoncés de test sont les segments issus d'un module de segmentation en locuteurs. La représentation du locuteur par mélange de gaussiennes (GMM) dans l'espace des coefficients acoustiques [8] constitue l'état de l'art pour les processus de suivi de locuteurs. Cependant, le coût de calcul de recherche d'un locuteur cible utilisant l'approche GMM, équivalent au calcul d'une vraisemblance dans un espace multi-gaussien pour chaque trame acoustique du document audio, est très important. Ainsi, afin de réduire le coût de calcul, on utilise l'approche statistique dans l'espace des modèles d'ancrage qui obtient des performances en vérification du locuteur équivalente à l'approche GMM [5]. Cette approche est utilisée dans les trois processus de suivi de locuteurs présentés au cours des paragraphes suivants. Dans ce cas, le coût de calcul est équivalent au calcul d'une vraisemblance dans un espace monogaussien pour chaque segment du document audio. Dans la suite de cet article, nous considérons les notations suivantes : soit X un locuteur cible, Y_i, $i = 1, \ldots, N$ les segments issus du module de segmentation et C_{Y_i} une classe de $N_{C_{Y_i}}$ segments à laquelle appartient le segment Y_i.

5.1. Suivi sans regroupement en locuteurs

Le processus de suivi de locuteurs sans regroupement en locuteurs consiste à comparer les segments Y_i au locuteurs X en utilisant le log-rapport de vraisemblance défini par l'équation 6.

5.2. Suivi avec regroupement en locuteurs

Le processus de suivi de locuteurs avec regroupement en locuteurs utilise pour chaque segment issu du processus de segmentation en locuteurs une information d'appartenance à une classe de locuteurs. L'information d'appartenance à une classe de locuteurs est déterminée par le processus de regroupement en locuteurs présenté au paragraphe 4. Cette méthode, présentée dans [2], permet d'obtenir des mesures de similarité plus fiables, notamment pour les segments de parole courts. La mesure de similarité entre le locuteur cible X et un segment Y_i en fonction de la classe C_{Y_i} est définie par :

$$L_{C_{Y_i}}(\widetilde{X},\widetilde{Y}_i) = \frac{1}{N_{C_{Y_i}}} \sum_{j=1}^{N_{C_{Y_i}}} L(\widetilde{X},\widetilde{Y}_j) \qquad (7)$$

où les segments Y_j appartiennent à la classe C_{Y_i}.

5.3. Suivi avec regroupement en locuteurs et mesure de similarité pondérée

Des expériences présentées dans [3] ont montré que les erreurs du processus de regroupement en locuteurs engendrent de nouvelles erreurs de suivi de locuteurs. Une nouvelle mesure de similarité dite *pondérée* permet de limiter la contribution des segments Y_j dans le calcul de l'équation 7 lorsque la probabilité que Y_j ait été prononcé par le même locuteur que Y_i est faible. La réduction des erreurs de suivi dues aux erreurs de regroupement s'effectue en introduisant un coefficient de pondération basé sur une mesure de similarité entre les segments Y_i et Y_j, relié à la probabilité que les segments soient prononcés par un même locuteur :

$$L^p_{C_{Y_i}}(\widetilde{X},\widetilde{Y}_i) = \frac{1}{\sum_{j=1}^{N_{C_{Y_i}}} \gamma_{ij}} \sum_{j=1}^{N_{C_{Y_i}}} \gamma_{ij} L(\widetilde{X},\widetilde{Y}_j) \qquad (8)$$

En pratique, le coefficient de pondération γ_{ij} est défini comme la fonction d'erreur complémentaire de la mesure de similarité de corrélation $\rho(\widetilde{Y}_i,\widetilde{Y}_j)$.

6. EXPERIENCES ET RESULTATS

Le système d'indexation en locuteurs utilisant la représentation du locuteur par les modèles d'ancrage est évalué sur la tâche de suivi de locuteurs de la campagne d'évaluation de systèmes de transcription d'émissions radiophoniques (campagne ESTER [9]). Le corpus d'évaluation, les mesures d'évaluation, la configuration et les performances du système sont présentés au cours des paragraphes suivants.

6.1. Corpus d'évaluation

Le corpus utilisé pour ces expériences est un corpus d'émissions radiophoniques en français. Le corpus est divisé en un ensemble d'apprentissage, un ensemble de développement et un ensemble de test selon les spécifications de la campagne ESTER (voir [9] pour plus de détails). L'ensemble d'apprentissage contient 82h d'émissions radiophoniques enregistrées sur la période 1998-2003. L'ensemble de développement contient 10h d'émissions radiophoniques enregistrées en 2003 et l'ensemble de test contient 10h d'émissions radiophoniques enregistrées en 2004. Les expériences présentées dans cet article sont effectuées sur l'ensemble de développement avec une liste de 279 locuteurs cibles fourni par les organisateurs de la campagne ESTER.

6.2. Mesure d'évaluation

6.2.1. Segmentation et regroupement en locuteurs

Les performances des systèmes de segmentation et de regroupement en locuteurs sont évaluées en termes de pureté moyenne en locuteurs P_{Loc} et de pureté moyenne des classes P_{Cl} . Ces mesures d'évaluations proposées par [10] sont définies par les équations suivantes :

$$P_{Cl} = \frac{1}{N_0} \sum_{i=1}^{N} p_i^{Cl} n_i \text{ avec } p_i^{Cl} = \sum_{j=1}^{S} \frac{n_{ij}^2}{n_i^2} \qquad (9)$$

$$P_{Loc} = \frac{1}{N_0} \sum_{j=1}^{S} p_j^{Loc} n_j \text{ avec } p_j^{Loc} = \sum_{i=1}^{N} \frac{n_{ij}^2}{n_j^2} \qquad (10)$$

où N est le nombre de classes du document audio, S le nombre de locuteurs du document audio, N_0 le nombre de trames du document audio, n_i le nombre de trames de la classe i, n_j le nombre de trames du locuteur j, n_{ij} le nombre de trames dans la classe i prononcées par le locuteur j.

6.2.2. Suivi de locuteurs

Les performances du système de suivi de locuteurs sont évaluées en termes de précision (P) et de rappel (R) :

- $P = \frac{\text{Nombre de trames du locuteur cible détectées}}{\text{Nombre de trames détectées}}$
- $R = \frac{\text{Nombre de trames du locuteur cible détectées}}{\text{Nombres de trames du locuteur cible}}$

Les valeurs de Précision et de Rappel sont combinées en une seule valeur d'évaluation en utilisant la F-mesure, qui est définie par

$$F = \frac{2.P.R}{P+R} \qquad (11)$$

6.3. Configuration du système

Dans chacune des expériences, 13 MFCC avec leurs dérivées premières et secondes plus ΔE and $\Delta\Delta E$ sont utilisés. Les modèles d'ancrage sont des modèles statistiques GMM à 256 gaussiennes appris par adaptation MAP d'un modèle UBM indépendent du genre. L'espace des modèles d'ancrage est composé de tous les locuteurs, différents des locuteurs cibles, qui ont plus de 70 secondes de parole disponibles dans l'ensemble d'apprentissage, soit 316 locuteurs. Aucune compensation de canal n'est appliquée pour les processus de segmentation et de regroupement en locuteurs et un module de *feature warping* [11] est appliqué sur les données acoustiques pour le processus de suivi de locuteurs.

6.4. Résultats

6.4.1. Segmentation et regroupement en locuteurs

Le tableau 1 présente les performances des processus de segmentation et de regroupement en locuteurs en termes de pureté moyenne des classes et pureté moyenne en locuteurs. Les erreurs de regroupement en locuteurs font diminuer la pureté des classes, cependant la pureté en locuteur est considérablement augmentée.

	P_{Cl}	P_{Loc}
Segmentation en locuteurs	95.3	19.2
Regroupement en locuteurs	83.7	54.3

Table 1. Performances des processus de segmentation et regroupement en locuteurs en termes de pureté moyenne en locuteurs et pureté moyenne des classes

6.4.2. Suivi de locuteurs

Trois systèmes de suivi de locuteurs sont comparés et leurs performances sont représentées sur la figure 1 en termes de précision et de rappel : suivi sans regroupement en locuteurs, suivi avec regroupement en locuteurs, suivi avec regroupement en locuteurs et mesure de similarité pondérée. Cette figure montre que le regroupement en locuteurs permet d'augmenter significativement les performances du système de suivi de locuteur pour atteindre des performances équivalentes à celles obtenues par un système de suivi de locuteurs utilisant une modélisation du locuteur par mélange de gaussiennes dans l'espace acoustique [12].
Le tableau 2, qui indique les points de fonctionnement optimaux de chaque système (point de la courbe précision/rappel qui maximise la F-mesure), montre que l'amélioration des performances liée au regroupement en locuteurs se traduit par une augmentation du taux de rappel (79.5 % contre 67.5 %) tandis que la mesure de similarité pondérée améliore la précision en réduisant l'impact des erreurs du processus de regroupement (86.8 % contre 84.4 %).

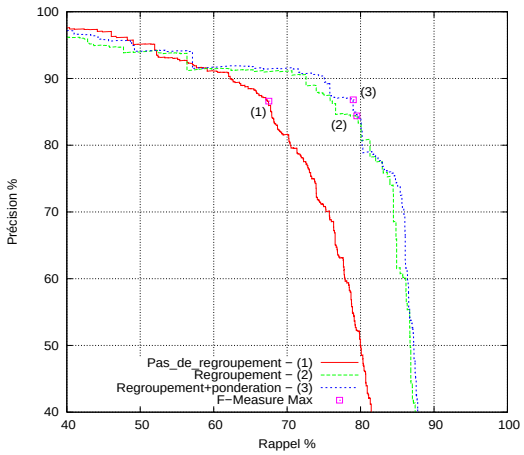

Fig. 1. Precision/Rappel pour chaque système de suivi de locuteurs : sans regroupement, avec regroupement, avec regroupement et mesure de similarité pondérée

Système	F_{max}	Précision	Rappel
Pas de regroupement (1)	75.8	86.6	67.5
Regroupement (2)	81.9	84.4	79.5
Regroupement + pondération (3)	82.7	86.8	79.0

Table 2. Points de fonctionnement des systèmes de suivi de locuteurs

7. CONCLUSION

Dans cet article, nous avons présenté un système d'indexation en locuteurs entièrement basé sur la représentation du locuteur par modèles d'ancrage qui consiste à représenter un locuteur dans un espace de locuteurs de référence. L'espace des locuteurs de référence est muni de méthodes permettant de comparer deux locuteurs utilisées dans des contextes différents : la mesure de similarité de corrélation [4] est utilisée pour les processus de segmentation et de regroupement en locuteurs tandis que l'approche statistique [5] est utilisée pour le processus de suivi de locuteurs.
Les évaluations montrent que l'utilisation d'un processus de regroupement pour le suivi de locuteurs améliore significativement le taux de rappel du système et que la mesure de similarité pondérée permet d'améliorer la précision en réduisant l'impact des erreurs de regroupement. Les performances du système d'indexation en locuteurs sont équivalentes aux performances d'un système d'indexation utilisant une représentation classique du locuteur dans l'espace des coefficients acoustiques [12]. Ce résultat est particulièrement intéressant car il montre que tout en conservant des performances équivalente à l'état de l'art [9], la représentation du locuteur par les modèles d'ancrage permet de diminuer le coût de calcul pour l'indexation de grandes bases de donnée audio.

8. REFERENCES

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

[2] M. Collet, D. Charlet, and F. Bimbot, "Speaker tracking by anchor models using speaker segment cluster information," in *ICASSP*, 2006.

[3] M. Collet, D. Charlet, and F. Bimbot, "A weighted measure of similarity for speaker tracking," in *Speaker Odyssey*, 2006.

[4] M. Collet, D. Charlet, and F. Bimbot, "A correlation metric for speaker tracking using anchor models," in *ICASSP*, 2005.

[5] M. Collet, Y. Mami, D. Charlet, and F. Bimbot, "Probabilistic anchor models approach for speaker verification," in *INTERSPEECH*, 2005.

[6] P. Delacourt, D. Kryze, and C. Wellekens, "Speaker-based segmentation for audio data indexing," in *ESCA ETRW Workshop*, 1999.

[7] L. Couvreur and J. M. Boite, "Speaker tracking in broadcast audio material in the framework of the thisl project," in *ESCA ETRW*, 1999.

[8] D.A. Reynolds, T.F. Quatieri, and R.B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital Signal Processing*, vol. 10, pp. 19–41, 2000.

[9] S. Galliano, E. Geoffrois, D. Mostefa, K. Choukri, J-F. Bonastre, and G. Gravier, "The ESTER phase 2 evaluation campaign for the rich transcription of french broadcast news," *EUROSPEECH*, 2005.

[10] A. Solomonoff, A. Mielke, M. Schmidt, and H. Gish, "Clustering speakers by their voices," in *ICASSP*, 1998.

[11] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," in *Speaker Odyssey*, 2001.

[12] D. Moraru, M. Ben, and G. Gravier, "Experiments on tracking and segmentation in radio broadcast news," in *INTERSPEECH*, 2005.