SUIVI DE LOCUTEUR PAR LA TECHNIQUE DES MODELES D'ANCRAGE

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
bimbot@irisa.fr

ABSTRACT

Ce papier présente l'application de la technique des modèles d'ancrage au suivi de locuteurs. Le système décrit est basé sur une procédure de segmentation en locuteur, suivi d'une procédure de détection des locuteurs cibles parmi les segments issus de la première procédure. La technique de représentation du locuteur par les modèles d'ancrages utilisée dans ce système de suivi de locuteur est présentée ainsi que les différentes métriques utilisées pour comparer deux locuteurs. Ensuite les modules de segmentation en locuteur et de détection des locuteurs cibles sont détaillés. Enfin le système est évalué sur la base de donnée d'émissions radiophoniques de la campagne d'évaluation ESTER. Dans le cadre de la segmentation en locuteur, les métriques utilisées par la technique des modèles d'ancrages apparaissent être plus performantes que les métriques statistiques de l'état de l'art.

1. INTRODUCTION

Depuis de récentes années, de nombreuses données audio (émissions radiophoniques) sont stockées dans de grandes bases de données. Dans ce contexte, la tâche de suivi de locuteurs qui consiste à chercher les zones de parole d'un locuteur cible, devient difficile. En réalité, la taille importante des archives audio augmente les temps de calcul d'un système de suivi de locuteurs et cela limite ses performances dans le cadre d'une application temps-réel. Dans la littérature, deux approches principales sont proposées. La première consiste à segmenter le signal audio et ensuite détecter le locuteur cible [1]. Dans la seconde approche, la segmentation et la détection sont effectués simultanément [2].

Le système de suivi de locuteurs proposé dans ce papier est basé sur la première approche et utilise la technique de modélisation du locuteur par les modèles d'ancrages [3]. Cette modélisation relative du locuteur est présentée dans la section 2 et permet de caractériser un locuteur par un vecteur indépendamment de la longueur de l'énoncé.

Les sections suivantes décrivent les deux modules constituant le système de suivi de locuteurs. Le premier, détaillé dans la section 3, sert à segmenter le signal audio en portions ayant été prononcées par un seul locuteur. Le second module permet de détecter les zones de parole d'un locuteur cible. La technique de détection décrite dans la section 4 compare un énoncé du locuteur cible avec tous les segments issus du module de segmentation en locuteur et décide si les segments ont été prononcés par le locuteur cible.

La dernière section de cet article présente les résultats des évaluations du système de suivi de locuteurs sur la base de donnée ESTER d'émissions radiophoniques françaises.

2. TECHNIQUE DES MODELES D'ANCRAGE

Les systèmes de suivi de locuteurs présentés dans la littérature sont basés sur une modélisation des locuteurs par GMM-UBM. Dans cet article, un système utilisant la modélisation des locuteurs par les modèles d'ancrage est proposé.

2.1. Principe

De récentes recherches [3][4] 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 références. Le locuteur n'est plus représenté de façon absolue mais relativement à un ensemble de locuteurs dont les modèles GMM-UBM sont appris. Ces modèles sont appelés modèles d'ancrages. Le locuteur est caractérisé par un vecteur défini comme l'ensemble des vraisemblances entre les données du locuteur et les 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 la log-vraisemblance moyenne des données X (de N vecteurs acoustiques) pour le modèle GMM du locuteur de référence $\overline{\lambda}_e$ relativement au modèle du monde (UBM) :

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

où λ_{UBM} est le modèle du monde qui a été utilisé pour initialiser l'apprentissage des modèles d'ancrages.

2.2. Métriques de comparaison des SCV

Les métriques pour comparer des SCV sont la métrique Euclidienne et la métrique Angulaire [4]. Une nouvelle métrique basée sur le coefficient de corrélation est également proposée par [5]. Soit X and Y deux segments de parole, $\widetilde{X}$ et $\widetilde{Y}$ leur SCV.

– Métrique Euclidienne :

$$d(\widetilde{X}, \widetilde{Y}) = \sqrt{|\widetilde{X} - \widetilde{Y}|^2} \quad (3)$$

avec

$$d(\widetilde{X},\widetilde{Y}) = 0 \Longleftrightarrow \widetilde{Y} = \widetilde{X} \quad (4)$$

– Métrique Angulaire :

$$\delta(\widetilde{X},\widetilde{Y}) = \arccos\left[\frac{\widetilde{X}\widetilde{Y}^T}{\sqrt{\widetilde{X}\widetilde{X}^T.\widetilde{Y}\widetilde{Y}^T}}\right] \quad (5)$$

avec

$$\delta(\widetilde{X},\widetilde{Y}) = 0 \Longleftrightarrow \widetilde{Y} = a\widetilde{X} \quad \forall a \in \Re\,,\, a \neq 0 \quad (6)$$

– Métrique de Corrélation :

$$\rho(\widetilde{X},\widetilde{Y}) = 1 - R(x,y) \quad (7)$$

où $R(x,y)$ est le coefficient de corrélation entre les composantes des deux SCV (les composantes sont considérées comme la réalisation de deux variables aléatoires x et y) :

$$R(x,y) = \frac{C_{xy}}{\sigma_x \sigma_y} \quad (8)$$

avec

$$\rho(\widetilde{X},\widetilde{Y}) = 0 \Longleftrightarrow \widetilde{Y} = a\widetilde{X} + b \quad \forall (a,b) \in \Re^2\,,\, a \neq 0 \quad (9)$$

3. SEGMENTATION EN LOCUTEUR

La première étape du système consiste à segmenter le document audio en segments homogènes de longueur raisonnable ayant été prononcés par un seul locuteur. Cette tâche de segmentation en locuteur s'effectue sans aucune donnée a priori sur les locuteurs à détecter lors de l'étape suivante.

3.1. Calcul de la courbe des distances

Dans un contexte d'indexation de donnée audio, [6] décrit un système de segmentation en locuteur basé sur le calcul d'une courbe des distances. Cette technique consiste à calculer une distance entre deux segments consécutifs X et Y. La fenêtre, composée des deux segments de 2.4 s chacun, est décalée toutes les 160 ms sur l'ensemble du signal audio et à chaque décalage, une distance entre ces deux segments est calculée.
Les systèmes de segmentation en locuteur de l'état de l'art utilisent des distances statistiques pour comparer deux segments de parole comme par exemple le rapport de vraisemblance généralisé (GLR pour *Generalized Likelihood Ratio*).

$$GLR = \frac{L(XY, \hat{\mu}_{XY}; \hat{\Sigma}_{XY})}{L(X, \hat{\mu}_X; \hat{\Sigma}_X) L(Y, \hat{\mu}_Y; \hat{\Sigma}_Y)} \quad (10)$$

où $L(X, \hat{\mu}_X; \hat{\Sigma}_X)$ représente la vraisemblance de la séquence acoustique X pour le processus multi-gaussien $\mathcal{N}(\mu_X; \Sigma_X)$ et XY est la concaténation des énoncés X et Y.
La distance GLR est calculée en prenant le logarithme de l'expression précédente :

$$d_{GLR} = -log(GLR) \quad (11)$$

Cette mesure est efficace pour détecter des ruptures statistiques dans le signal mais toutes les ruptures statistiques ne correspondent pas à un changement de locuteur.
C'est pourquoi, la technique des modèles d'ancrage, qui est une technique de modélisation orientée locuteur, apparait être une solution adaptée à la détection des changements de locuteurs.
Dans ce cas, les segments X et Y ne sont plus modélisés par un processus multi-gaussien dans l'espace des coefficients acoustiques, mais par leurs SCV $\widetilde{X}$ et $\widetilde{Y}$ dans l'espace des modèles d'ancrages et la distance dérivée du rapport de vraisemblance généralisé est remplacée par la métrique de corrélation $\rho(\widetilde{X},\widetilde{Y})$ définie dans la section 2.2.

3.2. Détection des maximums significatifs

La seconde étape du processus de segmentation en locuteur est la détection des maximums significatifs de la courbe des distances. La méthode de détection de ruptures proposée par [7] est utilisée dans ce système.
Cette méthode calcule un critère de rupture qui est défini comme étant la dévinellation de la courbe des distances à un instant t, par rapport à un niveau de palier local. Si ce critère de rupture est supérieur à un seuil, alors un changement de locuteur est détecté. Ce seuil permet de fixer le nombre de segments à la sortie du système de segmentation, mais le seuil optimal peut être différent d'un document à un autre.
Afin de tenir compte de ce problème, le seuil est déterminé a posteriori de façon à obtenir la même longueur moyenne de segment pour chaque document audio.

4. DETECTION DES LOCUTEURS CIBLES

La détection des locuteurs cibles consiste à rechercher parmi les segments issus du module de segmentation en locuteur, ceux ayant été prononcés par les locuteurs cibles. Le principe est de calculer une mesure de similarité entre le locuteur cible et un segment, puis de décider en fonction d'un seuil si le segment a été prononcé par le locuteur cible.
Dans le cadre de la technique des modèles d'ancrages, les locuteurs cibles et l'ensemble des segments sont projetés dans l'espace de référence et la métrique de corrélation est utilisée pour comparer les SCV des locuteurs cibles avec ceux des segments (des résultats experimentaux présentés dans [5] ont montré que la métrique de corrélation est plus performante que les métriques euclidienne et angulaire). Afin d'optimiser les performances de détection, une technique de sélection des modèles d'ancrages décrite dans le paragraphe suivant est utilisée.

4.1. Sélection des modèles d'ancrages

Dans le cadre de l'indexation de document audio, [8] propose de sélectionner E locuteurs d'ancrages parmi l'ensemble des locuteurs constituant l'espace de référence. Les locuteurs d'ancrages les plus proches du document audio (au sens de la vraisemblance des données du document par rapport au modèle des locuteurs d'ancrages) sont sélectionnés.
Pour la tâche de détection des locuteurs cibles, cette méthode de sélection des locuteurs d'ancrages est appliquée à chaque locuteur cible. Ainsi, pour chaque locuteur cible, un sous-espace d'ancrage est créé en choisissant les 50 locuteurs les plus proche du locuteur cible parmi 596 locuteurs de références (parmi lesquels se trouvent tous les locuteurs cibles). Le nouveau SCV d'un locuteur X dans le sous-espace de référence d'un locuteur Y sera appelé $\widetilde{X}_Y$. Le nombre de locuteurs sélectionnés a été optimisé par des

expériences préliminaires et ces dernières ont montré que $E = 50$ donne les meilleures performances.

4.2. Symétrisation de la métrique

Cette sélection des modèles d'ancrages par rapport au locuteur cible entraîne une disimilarité entre le locuteur cible et les segments de test. Afin d'obtenir une mesure symétrique, une extension de la métrique de corrélation est proposée.
Soit X un segment, Y le locuteur cible, $\widetilde{X}$ et $\widetilde{Y}$ leurs SCV dans l'espace de référence.

– Métrique non symétrique :

$$\rho(\widetilde{X}, \widetilde{Y}) = \rho(\widetilde{X}_Y, \widetilde{Y}_Y) \tag{12}$$

– Métrique symétrique :

$$\rho_{sym}(\widetilde{X}, \widetilde{Y}) = \frac{\rho(\widetilde{X}_X, \widetilde{Y}_X) + \rho(\widetilde{X}_Y, \widetilde{Y}_Y)}{2} \tag{13}$$

où $\widetilde{X}_X$ et $\widetilde{Y}_X$ sont les SCV de X et Y dans le sous-espace de référence de X, et $\widetilde{X}_Y$ et $\widetilde{Y}_Y$ sont les SCV de X et Y dans le sous-espace de référence de Y. Les sous-espaces de X et Y sont construits comme décrit dans le paragraphe 4.1.

5. EXPERIENCES ET RESULTATS

Le système de suivi de locuteur par les modèles d'ancrages est évalué sur la base de donnée d'émissions radiophoniques ESTER [9]. Les différentes techniques de segmentation en locuteur sont comparées et la métrique de corrélation ainsi que son extension symétrique sont évaluées sur la tâche de détection des locuteurs cibles. Le corpus d'évaluation, les mesures d'évaluations et la configuration du système sont présentés avant la description des résultats.

5.1. 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 (80 H), un ensemble de développement (10 H) et un ensemble de test (10 H). Les ensemble de d'apprentissage et de développement contiennent des émissions radiophoniques de France Inter, France Info, Radio France Internationnale et RTM (radio marocaine en français), tandis que l'ensemble de test contient des émissions des mêmes radio plus France Culture et France Classique. Pour chaque expérience réalisée, 279 locuteurs cibles d'une liste établie par les organisateurs de la campagne d'évaluation, appris sur le corpus d'apprenstissage, ont été utilisés.

5.2. Mesure d'évaluation

Les performances du système de suivi de locuteur sont évaluées en termes de Precision/Recall où Precision (PR) et Recall (RC) sont définis par :

– $PR = \frac{\text{Durée du locuteur cible détectée}}{\text{Durée détectée}}$

– $RC = \frac{\text{Durée du locuteur cible détectée}}{\text{Durée du locuteur cible}}$

Les valeurs de Precision et de Recall sont combinées en une seule valeur d'évaluation en utilisant la $F - measure$ [10], qui est définie par

$$F = \frac{2.PR.RC}{PR + RC} \tag{14}$$

5.3. Configuration du système

Dans chacunes des expériences, 13 MFCC avec leurs dérivées premières et secondes plus ΔE and $\Delta\Delta E$ sont utilisés et la CMS est appliquée. Les 596 modèles d'ancrages sont des modèles statistiques GMM à 256 gaussiennes appris sur le corpus d'apprenstissage de la campagne ESTER par adaptation MAP d'un modèle UBM indépendent du genre. La quantité des données d'apprentissage de chaque modèle d'ancrage varie entre 70 s et 200 s. Les modèles (GMM ou SCV) de chacun des locuteurs à suivre sont également appris sur le corpus d'apprentissage et la quantité d'apprentissage est bornée à 200 s.

5.4. Résultats

5.4.1. Segmentation en locuteur

La figure 1 montre les courbes Precision/Recall pour un système de détection des locuteurs cibles par GMM-UBM en fonction du type de segmentation en locuteur pour une même longueur moyenne des segments égale à 12 s. Chaque point de ces courbes correspond à un seuil de décision différent. Le point de fonctionnement de chaque système est celui qui maximise la $F-measure$ et est marqué par un '+' sur la figure 1. Il s'avère que la technique de segmentation utilisant les modèles d'ancrages ($F - measure = 77.8$) améliore nettement les performances du système de détection par rapport à un système de segmentation classique où la distance GLR est utilisée ($F - measure = 73$).

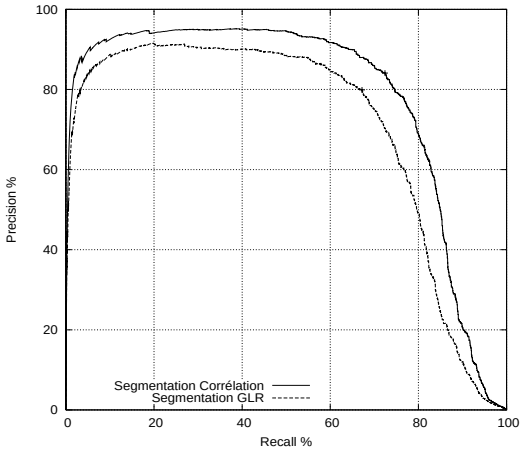

Fig. 1. Precision/Rappel pour la segmentation par GLR et par la métrique de corrélation

5.4.2. Détection des locuteurs cibles

Les performances des métriques de corrélation non symétrique et symétrique en détection des locuteurs cibles sont reportées sur la figure 2 ainsi que dans le tableau 1. Ces résultats sont obtenus pour une même segmentation en locuteur et ils montrent que la métrique symétrique améliore les performances du système.

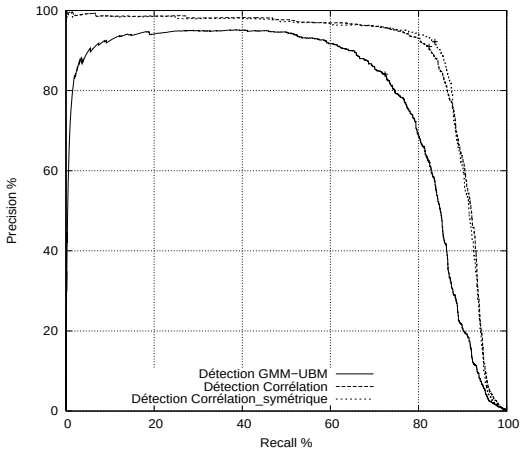

Fig. 2. Precision/Rappel en fonction du système de détection pour une segmentation en locuteur par la métrique de corrélation

Métrique	F_{max}	PR_{max}	RC_{max}
Corrélation	86.5	91.0	82.4
Corrélation symétrique	87.8	92.2	83.7

Table 1. Points de fonctionnement des systèmes de détection

5.4.3. Résultats sur le corpus de test

A partir des résultats obtenus sur l'ensemble de développement, le système de suivi de locuteurs obtenant les meilleures performances est appliqué sur l'ensemble de test.

- Segmentation en locuteur par la métrique de corrélation avec une longueur moyenne des segmnets de 12 s.
- Détection des locuteurs cibles par la métrique de corrélation symétrique.

Corpus	F_{max}	PR_{max}	RC_{max}
DEV	87.8	92.2	83.7
TEST	56.6	61.4	52.5

Table 2. Performances du système de suivi de locuteurs

Les performances obtenues sur l'ensemble de test sont décevantes par rapport à celle obtenues sur l'ensemble de développement. Cette importante différence de performances peut être expliquée par la date de diffusion des émissions radiophoniques qui varient de décembre 1998 à juillet 2003 pour l'ensemble d'apprenstissage, d'avril 2003 à juillet 2003 pour l'ensemble de développement et d'octobre 2004 à décembre 2004 pour l'ensemble de test. L'écart de date important entre l'ensemble de test et l'ensemble d'apprenstissage peut être un facteur de dégradation des performances lié à la dérive temporelle de la voix des locuteurs ou bien à des changements de l'environnement acoustique lors de la diffusion des émissions radiophoniques .
Ce facteur n'explique pas à lui seul l'écart et d'autres facteurs, comme par exemple la probabilité a priori des locuteurs cibles qui est beaucoup plus faible sur l'ensemble de test ou la présence de nouvelle radio (ex : France Culture), peuvent aussi intervenir.

6. CONCLUSION

Cet article a présenté un système suivi de locuteur basé sur la technique de modélisation relative du locuteur dans un espace de modèle d'ancrages. Une nouvelle méthode de segmentation en locuteur utilisant la métrique de corrélation a été proposée et apparait être plus performante que les méthodes statistiques classiques. D'autres part, le module de détection de locuteur cible intègre une extension de la métrique de corrélation par symétrisation qui permets d'obtenir une légère amélioration des performances de détection.
Les résulats obtenus sur le corpus de développement de la base de donnée ESTER ont été comparés à ceux obtenus sur le corpus de test de cette même base de donnée et une forte dégradation des performances est observée entre ces deux corpus.
Les travaux futurs vont donc s'orienter vers une analyse approfondie des facteurs expliquant ce manque de robustesse pour proposer des solutions afin d'améliorer les performances.

7. REFERENCES

[1] Lie Lu and Hong-Jiang Zhang, "Speaker change detection and tracking in real-time news broadcasting analysis," in *ACM International Conference on Multimedia*, 2002, pp. 602–610.

[2] I-M. Chagnolleau, A-E. Rosenberg, and S. Parthasarathy, "Detection of target speakers in audio databases," in *ICASSP'99*, 1999.

[3] 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.

[4] Yassine Mami and Delphine Charlet, "Speaker identification by location in an optimal space of anchor models," in *ICSLP*, 2002, vol. 2, p. 1333.

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

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

[7] M. Seck, R. Blouet, and F. Bimbot, "The irisa/elisa speaker detection and tracking systems for the nist'99 evaluation campaign," *Digital Signal Processing*, vol. 10, pp. 154–171, 2000.

[8] Yuya Akita and Tatsuya Kawahara, "Unsupervised speaker indexing using anchor models and automatic trancription of discussions," in *EUROSPEECH 2003*, 2003, pp. 2985–2988.

[9] G. Gravier, J-F. Bonastre, S. Galliano, E. Geoffrois, K. Mc Tait, and K. Choukri, "The ester evaluation campaign of rich transcription of french broadcast news," in *Language Evaluation and Resources Conference*, 2004, www.afcp-parole.org/ester.

[10] T. Kemp, M. Schmidt, M. Westphal, and A. Waibel, "Strategies for automatic segmentation of audio data," in *ICASSP 2000*, 2000.