SETIT 2005
3rd International Conference: **S**ciences of **E**lectronic, **T**echnologies of **I**nformation and **T**elecommunications
March 27-31, 2005 – TUNISIA

Recherche des Points de Rupture dans les Discours Multi-Locuteurs - Application en Indexation par Locuteurs -

H. Sayoud*, S. Ouamour* et M. Guerti**

USTHB, Institut d'Electronique, BP 32 Bab Ezzouar, Alger, Algérie
sayoud@ifrance.com
ouamour@ifrance.com

***ENP, Institut d'Electronique Av. H-badi ,El-Harrach, Alger, Algérie*

Résumé : L'indexation par locuteurs d'un document audio, qui consiste, en fait, à reconnaître la séquence de locuteurs engagés dans la conversation, tient une place essentielle en reconnaissance du locuteur. Il s'agit de savoir *Qui parle ? et Quand ?* afin de saisir la cohérence du dialogue. Une des tâches les plus importantes en indexation par locuteurs est la segmentation de la parole en zones délimitant l'identité des différents locuteurs parlants.

Dans cette étude, nous avons utilisé les mesures statistiques du 2ème ordre (SOSM) afin de segmenter ces zones et ceci en détectant les points de changement des locuteurs appelés aussi zones de transition ou points de rupture. Les tests ont été effectués sur des documents audio multi-locuteurs de « Broadcast News ». Les résultats ainsi obtenus, sont satisfaisants.

Mots clés : Indexation de documents audio, segmentation en locuteurs, points de rupture.

1 Introduction

Actuellement, avec la multiplication des chaînes de télévision et de radio, des milliers d'heures d'émission sont stockées chaque année par des instituts d'archivage (Delacourt, 2000). Parmi cette masse immense de données, l'accessibilité directe à des zones contenant la parole d'un président, par exemple, apparaît une tâche lourde car elle demande beaucoup de temps pour écouter l'intégralité du document audio, pour ensuite en extraire ces zones. En revanche, cette tâche peut être aisée si on connaissait les intervalles de temps de ces zones d'intérêts. Ce problème nous a incité à développer un système automatique pour la segmentation des documents audio en détectant les points de changement des locuteurs basé sur les mesures statistiques du 2ème ordre. Les tests qui ont été effectués sur des discours multi-locuteurs de HUB-4 ont donné des résultats intéressants.

2 Méthode Statistique utilisée

Nous avons élaboré un système automatique pour le suivi automatique du locuteur, basé sur les mesures Statistiques de vraisemblance du type SOSM (Second Order Statistical Measures) (Bimbot & al, 1995). Cette méthode a été introduite par F. Bimbot et al. en 1995.

2.1 Propriétés du modèle gaussien

Soit une suite de M vecteurs résultant de l'analyse acoustique de dimension p d'un signal de parole prononcé par le locuteur $\boldsymbol{x}$. Les coefficients composant ces vecteurs sont obtenus soit par bancs de filtres, par prédiction linéaire ou par cepstre.

Sous l'hypothèse d'un modèle Gaussien du locuteur (Bimbot & al, 1995) (Bonastre & al, 1997), la suite des vecteurs $\overline{x}$ peut être résumée par son vecteur moyenne et sa matrice de covariance *X*, tels que :

$$\overline{x} = \frac{1}{M}\sum_{t=1}^{M} x_t \qquad (1.a)$$

et

$$X = \frac{1}{M}\sum_{t=1}^{M}(x_t - \bar{x})(x_t - \bar{x})^T \quad (1.b)$$

De même, pour un autre locuteur ***y***, la suite de N vecteurs peut être modélisée par $\bar{y}$ et Y, avec :

$$\bar{y} = \frac{1}{N}\sum_{t=1}^{N} y_t \quad (2.a)$$

et

$$Y = \frac{1}{N}\sum_{t=1}^{N}(y_t - \bar{y})(y_t - \bar{y})^T \quad (2.b)$$

Les vecteurs moyenne $\bar{x}$ et $\bar{y}$ sont de dimension p, tandis que les covariances X et Y sont des matrices symétriques de dimension p*p.

Ainsi, le locuteur ***x*** (respectivement ***y***) sera représenté par $\bar{x}$, X et M, (respectivement $\bar{y}$, Y et N).

2.2 Notion de mesure de similarité

La mesure de similarité μ (***x***, ***y***) entre les locuteurs ***x*** et ***y*** peut être exprimée comme la fonction **Φ** suivante.

$$\mu(\boldsymbol{x},\boldsymbol{y}) = \boldsymbol{\Phi}(\bar{x}, X, M, \bar{y}, Y, N) \quad (3)$$

Elle est non-négative, c'est à dire :

$$\forall\, \boldsymbol{x},\ \forall\, \boldsymbol{y},\ 0 \le \mu(\boldsymbol{x},\boldsymbol{y}), \quad (4)$$

et elle satisfait la propriété (5):

$$\forall\, \boldsymbol{x},\ \mu(\boldsymbol{x},\boldsymbol{x}) = 0. \quad (5)$$

$$\forall\, \boldsymbol{x},\ \forall\, \boldsymbol{y},\ \mu(\boldsymbol{x},\boldsymbol{y}) = \mu(\boldsymbol{y},\boldsymbol{x}). \quad (6)$$

2.3 La mesure de vraisemblance gaussienne

2.3.1 Définition

En supposant que tous les vecteurs acoustiques extraits du signal de parole prononcé par le locuteur x sont distribués selon une distribution gaussienne (Bimbot & al, 1995), la vraisemblance d'un vecteur acoustique y_t seul prononcé par le locuteur ***y*** est donnée par la fonction G(y_t /***x***) suivante.

$$G(y_t / x) = \frac{1}{(2\Pi)^{p/2}(\det X)^{1/2}} \times \exp\left(-\frac{1}{2}(y_t - \bar{x})^T X^{-1}(y_t - \bar{x})\right) \quad (7)$$

Et si nous supposons que tous les vecteurs sont indépendamment observables, la moyenne du log-vraisemblance de $\{y_t\}_{1\le t\le N}$ peut être décrite par :

$$\begin{aligned}\bar{G}_x(y_1^N) &= \frac{1}{N}\log G(y_1 \ldots y_N / x)\\ &= \frac{1}{N}\sum_{t=1}^{N}\log G(y_t / x) =\\ &= -\frac{1}{2}\left[p\log 2\Pi + \log(\det X)\right] +\\ &-\frac{1}{2}\left[\frac{1}{N}\sum_{t=1}^{N}(y_t - \bar{x})^T X^{-1}(y_t - \bar{x})\right]\end{aligned} \quad (8)$$

Par ailleurs, en remplaçant $y_t - \bar{x}$ par $y_t - \bar{y} + \bar{y} - \bar{x}$ par et en utilisant la propriété mathématique (9)

$$\frac{1}{N}\sum_{t=1}^{N}(y_t - \bar{y})^T X^{-1}(y_t - \bar{y}) = tr(YX^{-1}) \quad (9)$$

Nous aurons alors

$$\begin{aligned}&\bar{G}_x(y_1^N) \ +\\ &+\frac{p}{2}\log 2\Pi =\\ &-\frac{1}{2}\left[\log(\det X) + tr(YX^{-1}) + (\bar{y} - \bar{x})^T X^{-1}(\bar{y} - \bar{x})\right]\end{aligned} \quad (10)$$

et aussi l'expression

$\frac{2}{p}\bar{G}_x(y_1^N) + \log 2\Pi + \frac{1}{p}\log(\det Y) + 1$ sera égale à

$$\frac{1}{p}\left[\log\left(\frac{\det Y}{\det X}\right) - tr(YX^{-1}) - (\bar{y} - \bar{x})^T X^{-1}(\bar{y} - \bar{x})\right] + 1 \quad (11)$$

Donc, si nous définissons la mesure de vraisemblance gaussienne μ_G comme : $\mu_G(\boldsymbol{x},\boldsymbol{y})$=

$$= \frac{1}{p}\left[tr(YX^{-1}) - \log\left(\frac{\det Y}{\det X}\right) + (\bar{y} - \bar{x})^T X^{-1}(\bar{y} - \bar{x})\right] - 1 \quad (12)$$

$$= \frac{1}{p}\left[tr(\Gamma) - \log(\det\Gamma) + \delta^T X^{-1}\delta\right] - 1 \qquad (13)$$

$$= a - \log g + \frac{1}{p}\delta^T X^{-1}\delta - 1 \qquad (14)$$

alors nous aurons

$$\mathbf{Argmax}_x \{\overline{G}_x(\overline{y}_1^N)\} = \mathbf{Argmin}_x \{\mu_G(\boldsymbol{x}, \boldsymbol{y})\} \qquad (15)$$

Sachant que

$$a(\lambda_1, \lambda_2, ..., \lambda_p) = \frac{1}{p}\sum_{i=1}^{p}\lambda_i \qquad (16)$$

représente la moyenne arithmétique,

les λ_i représentent les valeurs propres de Γ,

avec $\Gamma = X^{-\frac{1}{2}} Y X^{\frac{1}{2}}$, (17)

X représente la matrice de covariance représentant $\boldsymbol{x}$,

Y représente la matrice de covariance représentant $\boldsymbol{y}$,

$$g(\lambda_1, \lambda_2, ..., \lambda_p) = \left(\prod_{i=1}^{p}\lambda_i\right)^{1/p} \qquad (18)$$

représente la moyenne géométrique,

et $\delta = \overline{y} - \overline{x}$ (19)

La mesure gaussienne de vraisemblance μ_G est non symétrique. C'est-à-dire que, si on considère son terme dual $\mu_G(\boldsymbol{y}, \boldsymbol{x})$,

$$\mu_G(\boldsymbol{y}, \boldsymbol{x}) = \frac{1}{h} + \log g + \frac{1}{p}\delta^T Y^{-1}\delta - 1 \qquad (20)$$

alors on remarque que $\mu_G(\boldsymbol{x}, \boldsymbol{y}) \neq \mu_G(\boldsymbol{y}, \boldsymbol{x})$

sachant que

$$h(\lambda_1, \lambda_2, ..., \lambda_p) = \left(\frac{1}{p}\sum_{i=1}^{p}\left(\frac{1}{\lambda_1}\right)_i\right)^{-1} \qquad (21)$$

représente la moyenne harmonique

La mesure de vraisemblance gaussienne par covariance μ_{GC} peut donc être dérivée de la mesure précédente de la vraisemblance par l'équation (22) suivante, en supposant que la variabilité inter-locuteur de la moyenne est nulle.

Soit :

$$\mu_{GC}(\boldsymbol{x}, \boldsymbol{y}) = a - \log g - 1 \qquad (22)$$

Cette mesure est aussi non symétrique :

$$\mu_{GC}(\boldsymbol{y}, \boldsymbol{x}) = \frac{1}{h} + \log g - 1 \neq \mu_{GC}(\boldsymbol{x}, \boldsymbol{y}) \qquad (23)$$

2.3.2 La symétrisation

La mesure vue précédemment a la propriété d'être non symétrique, autrement dit, les rôles joués par les données de l'apprentissage et les données du test ne sont pas interchangeables. Cependant, notre intuition logique serait que la mesure de similarité devrait être symétrique.

L'asymétrie des mesures μ_G et μ_{Gc} peut être expliquée par le fait que ces mesures sont basées sur des essais statistiques qui supposent que la référence (le modèle du locuteur x) est exacte, tandis que le modèle test (le locuteur y) est une estimation. Mais en pratique, les deux modèles de référence et de test sont des estimations (Bimbot & al, 1995).

De plus, il est clair que la fiabilité d'un modèle de référence est dépendante du nombre de données qui a été employé pour estimer ses paramètres. Ceci est confirmé expérimentalement par les différences en performance qui peuvent être observées dans les tests d'identification du locuteur entre μ (x, y) et μ (y, x), surtout si M et N (le nombre de vecteurs de référence et celui de test) sont disproportionnés.

Une simple possibilité pour la symétrisation de la mesure μ (x, y), est de construire la moyenne :

μ $(\boldsymbol{x}, \boldsymbol{y})$, est de construire la moyenne :

$\mu_{[0.5]}(\boldsymbol{x}, \boldsymbol{y})$ entre la mesure et son terme dual :

$$\mu_{[0.5]}(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{2}\mu(\boldsymbol{x}, \boldsymbol{y}) + \frac{1}{2}\mu(\boldsymbol{y}, \boldsymbol{x}) = \mu_{[0.5]}(\boldsymbol{y}, \boldsymbol{x}) \qquad (28)$$

La mesure de vraisemblance gaussienne symétrisée de cette façon, devient :

$\mu_{G[0.5]}(\boldsymbol{x}, \boldsymbol{y}) =$

$$= \frac{1}{2}\left[a + \frac{1}{h} + \frac{1}{p}\delta^T\left[X^{-1} + Y^{-1}\right]\delta\right] - 1, \qquad (29)$$

Tandis que la mesure de vraisemblance à covariance devient :

$$\mu_{GC[0.5]}(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{2}\left[a + \frac{1}{h}\right] - 1, \qquad (30)$$

Cette procédure de symétrisation peut améliorer nettement les performances de classification, en comparaison avec les deux termes non symétriques pris individuellement.

3 Détection des points de changement

L'algorithme utilisé pour la détection des points de changement des locuteurs se résume par les étapes suivants :

a- Pré-traitement du signal de parole.

b- Sélection d'une zone de parole de 4 secondes correspondant à 2 fenêtres adjacentes de 2 secondes chacune et centrées autour de l'instant i.

c- Calcul de la mesure statistique de similarité di (telle que définie dans le chapitre 2) entre ces 2 fenêtres adjacentes.

d- Stockage de cette valeur, notée di, en mémoire (dans ce cas $d = \mu_{GC}$).

e- Incrémentation du temps i (i = i+1).

f- Refaire les étapes b, c,…, e, jusqu'à la fin du signal de parole.

g- Lissage de la courbe di.

h- Recherche des maxima de di.

i- Mémorisation des valeurs des temps imax correspondant à ces maxima.

Finalement, les zones de changement des locuteurs (zones de rupture) seront localisées aux instants imax (voir figure 1).

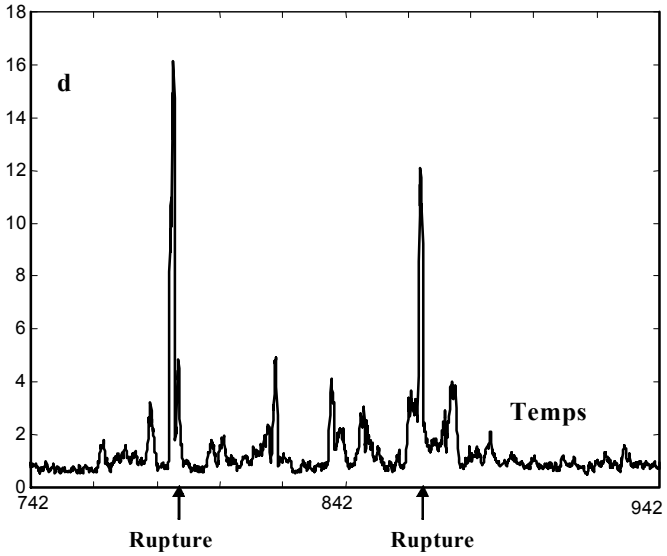

Figure 1: Détection des ruptures aux maxima de la mesure statistique d (dans ce cas $d = \mu_{Gc}$.).

Cette procédure de détection de rupture s'appelle la segmentation.

L'indexation par locuteurs consiste alors à identifier le locuteurs parlant pendant chaque zone homogène du discours.

L'identification se fait par la règle du plus proche voisin par rapport à un ensemble de locuteurs de référence, en utilisant la mesure statistique de similarité μ_{Gc}.

Le clustering (accompagnant l'indexation) (Delacourt, 2000) consiste à regrouper les segments adjacents d'un même locuteur en un seul segment.

4 Résultats et Interprétations

Les tests expérimentaux ont été faits sur un enregistrement d'un journal télévisé de la CNN, de durée 30 minutes et contenant 19 locuteurs. Cet enregistrement est extrait de la base de données Broadcast News (HUB-4) développée par le NIST.
Afin d'estimer la précision de notre méthode, nous avons adopté 3 taux d'erreur qui sont :

- Le taux de fausses alarmes noté TFA qui est égal au rapport du nombre de fausses ruptures détectés sur le nombre total de ruptures détectées.

- Le taux de détections manquées noté TDM qui est égal au rapport du nombre de ruptures non détectées sur le nombre total de ruptures détectées.

- Le taux d'erreur d'indexation noté TEI représente le rapport entre le nombre de fenêtres à confusion (mal indexées) sur le nombre total de fenêtres.

Tableau 1 : Taux d'erreur TFA.

Intervalle de précision	TFA avant indexation	TFA après indexation
+/- 0.5 seconde	44,03 %	11,11 %
+/- 1 seconde	42.12 %	6,20 %

Tableau 2 : Taux d'erreur TDM.

Intervalle de précision	TDM avant indexation	TDM après indexation
+/- 0.5 seconde	17%	16.2%
+/- 1 seconde	11.7%	10.3%

Nous remarquons que le TFA après indexation et clustering est plus faible que le TDM (tableaux 1 et 2). Nous remarquons aussi que le clustering a permis de réduire nettement les taux d'erreurs : soit de 42% jusqu'à 6.2% pour le TFA et de 11.7% jusqu'à 10.3% pour le TDM dans le cas d'une résolution d'une seconde. Ceci montre l'intérêt de l'indexation pour l'amélioration de la précision en détection de ruptures (changement des locuteurs).
Le taux d'erreur d'indexation après clustering global a été trouvé égal à 2.38%. Ce taux d'erreur est relativement très faible et correspond à un taux de bonne indexation de 97.6%.

5 Conclusion Générale

Durant ce travail de recherche, nous nous sommes intéressés à détecter les points de changement (ruptures) des locuteurs dans un discours multi-locuteurs. Pour ce faire, nous avons employé des mesures de similarité statistiques pour la segmentation et l'indexation du signal de parole.

Les expériences effectuées sur la base de données HUB-4 (comprenant 19 locuteurs) ont montré l'efficacité des mesures statistiques dans ce type de tâche : Ainsi, dans le cas d'une résolution d'une seconde, le TFA vaut seulement 6.2% contre un TDM de 10.3% ; ce qui explique que 90% des ruptures seront détectées. Concernant le procédé d'indexation, la précision (d'étiquetage) s'élève à 97,6%, ce qui représente un taux satisfaisant dans le domaine du suivi de locuteur.
Ce travail confirme alors l'efficacité de l'approche statistique du 2ème ordre en détection des points de changement des locuteurs.

Références bibliographiques

(Bimbot & al, 1995) F. Bimbot, I. Magrin-Chagnolleau, et L. Mathan *"Second-Order Statistical measures for text-independent Broadcaster Identification"*. Speech Communication, Volume. 17, Number, 1-2, August 1995, pp. 177-192.

(Bonastre & al, 1997) J.F. Bonastre et L. Besacier *"Traitement Indépendant de Sous-bandes Fréquentielles par des méthodes Statistiques du Second Ordre pour la Reconnaissance du Locuteur"*. Actes du 4ème Congrès Français d'Acoustique, pp 357-360, Marseille 14-18 April 1997.

(Bonastre & al, 2000) Bonastre, J.F., Delacourt, P., Fredouille, C., Merlin, T., Wellekens, C.J. 2000. *"A speaker tracking system based on speaker turn detection for NIST evaluation"*. In: IEEE International Conference on Acoustics Speech and Signal Processing, 2000.

(Bonastre & al, 2000²) J.F. Bonastre et Al, *" Modèle de Markov évolutif pour les tâches de suivi de locuteurs "*. JEP'2000, Aussois, France, pp 69-72, 19-23 juin 2000.

(Delacourt, 2000) P. Delacourt, *"Indexation de données audio: segmentation et regroupement par locuteurs"*. PhD thesis , Ecole Normale Supérieure des Télécommunications, Paris, France, 2000.

(Nishida & al, 1998) Nishida, M., Ariki, Y. *"Real time speaker indexing based on subspace method: applications to TV news articles and debate"*. In: International Conference on Spoken Language Processing. Vol. 4, pp. 1347-1350. 1998.

(Nishida & al, 1999) Nishida, M., Ariki, Y. *"Speaker indexing for news articles debates and drama in broadcasted TV programs"*. In: IEEE International Conference on Multimedia Computing and Systems. pp. 466-471, 1999.

(Stiefelhagen & a1, 1999) Stiefelhagen, R., Yang, J., & Waibel, *"A. Modeling focus of attention for meeting indexing"*. Proc. of ACM Multiemdia'99, 1999.

(Woodland & al, 1997) Woodland, P.C., Gales, M.J.F., Pye, D., Young, S.J. *"The Development of the 1996 HTK broadcast news transcription system"*. In: DARPA Speech Recognition Workshop. pp. 97-99, 1997.