SETIT 2007
4th International Conference: **S**ciences of **E**lectronic,
Technologies of **I**nformation and **T**elecommunications
March 25-29, 2007 – TUNISIA

"ISI" Une Nouvelle Technique pour la Segmentation Automatique par Locuteurs.

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

*USTHB, FEI, BP 32 Bab Ezzouar, Alger. # E.N.P. 10, Avenue Hacen Badi, El Harrach, Alger

Emails: ouamour@ifrance.com mhaniag@yahoo.fr sayoud@ifrance.com

Abstract : In this paper we propose a new algorithm called ISI or "Interlaced Speech Indexing", developed and implemented for the task of speaker detection and tracking. It consists in finding the identity of a well-defined speaker and the moments of his interventions inside an audio document, in order to access rapidly, directly and easily to his speech. Speaker Tracking can broadly be divided into two problems: Locating the points of speaker change (Segmentation of the document) and looking for the target speaker in each segment using a verification system in order to extract his global speech in the document: Speaker Detection. For the segmentation task, we developed a method based on an Interlaced Equidistant Segmentation (IES) associated with the ISI algorithm. This approach uses a speaker identification method based on symmetric statistical measures. As statistical measures, we chose the "µGc" one, which is based on the covariance matrix. However, the experiments showed that this method needs, at least, a speech length of 2 seconds, which means that the best segmentation resolution will be 2 seconds. By combining these measures with our new Indexing technique (ISI), we demonstrate that the average segmentation error is reduced to only 0.5 second, which is more accurate and more interesting for real-time applications.
Results indicate that the association IES-ISI provides a high resolution and a high tracking performance: the tracking score (percentage of correct micro-segments) is 95% on TIMIT database and 92.4% on Hub4 database.

Key words: Speaker Tracking, Speech Processing.

INTRODUCTION

Le domaine du suivi ou de détection du locuteur s'intéresse à chercher dans un document audio, toutes les interventions d'un locuteur particulier (cible). Cependant, avec l'évolution de la technologie de l'information et la multiplication des moyens de communication (satellite, internet, etc), il existe des milliers de chaînes de TV et radio qui transmettent une quantité immense d'informations. Parmi ce nombre énorme d'informations, trouver les paroles d'un locuteur particulier et les moments de son intervention dans le document audio nécessite que ces documents soient bien archivés et facilement accessibles. Pour cela, plusieurs techniques existent utilisant différents mots clés (mot, thème ou sujet, etc) suivant l'application souhaitée. Toutefois, ces techniques restent non efficaces pour la tâche de suivi ou détection du locuteur, où l'identité du locuteur reste le mot clé le plus convenable dans de telles applications.
Pour ce faire, le système de suivi doit posséder dans son dictionnaire de références un modèle du locuteur à suivre. Donc, la tâche du suivi du locuteur peut être vue comme une tâche de vérification appliquée tout le long du document audio contenant plusieurs interventions appartenant à des locuteurs inconnus : *Détection du locuteur*. Les moments du Début/Fin pour chaque intervention du locuteur suivi doivent aussi être retrouvés durant le processus. A la fin de cette opération, les différentes paroles du locuteur cible seront rassemblées pour l'obtention de l'intervention globale de ce dernier.
Nous avons alors développé pour cette tâche un nouveau système basée sur les mesures statistiques et un nouveau algorithme d'indexation entrelacée de la parole (ISI). Cet algorithme est facile à implanter, simple d'utilisation et efficace car il a amélioré significativement les résultats. En effet, cette association a amélioré les scores du suivi du locuteur ainsi que la précision de segmentation.

1. Bref overview des techniques de suivi

Dans cette section, nous donnons une brève description des techniques existant en suivi audio visuel.
Ainsi, dans ce vaste domaine du suivi du locuteur, Meignier [1] a employé un modèle de Markov progressif (le modèle change à chaque détection d'un nouveau locuteur) sur un sous-ensemble de la base de données Switchboard. Par ailleurs, Magrin-Chagnolleau [2] a utilisé un modèle GMM pour le suivi d'un locuteur cible sur la base de données Broadcast News. Tandis que Cettolo [3] a utilisé une approche presque similaire sur un corpus italien de Broadcast News. Une nouvelle règle de suivi présentée par Johnson [4] est basée, par contre, sur une stratégie de clustering par agglomération et par division (sur Hub-4).
Généralement, ces méthodes utilisent trois types de segmentation : utilisation d'un SAD, détection des changements des caractéristiques de la parole ou bien une identification de la nature des segments.
Dans notre travail de recherche, nous avons approché le problème avec une nouvelle méthode:
Nous utilisons d'abord une segmentation équidistante entrelacée, puis une technique de détection de locuteur pour l'étiquetage et finalement, en cas de confusion ou d'erreur de transition, une nouvelle règle de correction et de clustering est propose pour assurer cette tâche: tout ce procédé est appelé ISI.

2. Mesures de similarité statistiques

La méthode de comparaison des segments (par locuteurs) est basée sur les mesures de similarité suivantes [5]:

Soient $\{x_t\}_{1\le t\le M}$ (respectivement $\{y_t\}_{1\le t\le N}$) une séquence de M (respectivement N) vecteurs résultant de l'analyse acoustique *P*-dimensionelle d'un signal de parole prononcé par un locuteur $\boldsymbol{x}$ (respectiv. $\boldsymbol{y}$). Ces vecteurs sont résumés par le vecteur moyenne $\bar{x}$ (respectiv. $\bar{y}$) et la covariance X (respectiv. Y)

La mesure de similarité $\mu_{GC}(\boldsymbol{x},\boldsymbol{y})$ entre la prononciation de test $\{y_t\}_{1\le t\le M}$ du locuteur $\boldsymbol{y}$ et le model du locuteur $\boldsymbol{x}$ est définit par:

$$\mu_{GC}(\boldsymbol{x},\boldsymbol{y}) = \frac{1}{P}\left[-\log(\frac{\det(Y)}{\det(X)}) + tr(YX^{-1})\right] - 1 \quad (1)$$

Une mesure symétrique peut être construite en combinant $\mu_{GC}(\boldsymbol{x},\boldsymbol{y})$ avec son terme dual $\mu_{GC}(\boldsymbol{y},\boldsymbol{x})$:

$$\mu_{GC0.5}(\boldsymbol{x},\boldsymbol{y}) = \frac{\mu_{GC}(\boldsymbol{x},\boldsymbol{y}) + \mu_{GC}(\boldsymbol{y},\boldsymbol{x})}{2} \quad (2)$$

Une autre possibilité de symétrisation est donnée par :

$$\mu_{GC\beta}(\boldsymbol{x},\boldsymbol{y}) = \frac{M\mu_{GC}(\boldsymbol{x},\boldsymbol{y}) + N\mu_{GC}(\boldsymbol{y},\boldsymbol{x})}{M+N} \quad (3)$$

Cette symétrisation peut améliorer la performance de classification, comparée aux autres mesures individuelles.

3. suivi et détection du locuteur

Le suivi du locuteur consiste à retrouver dans un flux audio toutes les paroles prononcées par une personne particulière appelée cible et les moments de ses différentes interventions [6]. Dans notre cas, notre système de suivi est base sur la segmentation entrelacée du locuteur.

A. Segmentation Entrelacée

Dans notre application, nous avons divisé le signal de parole en deux groupes uniformes de segments, chaque segment à une durée de 2 secondes. Le second segment est retardé par rapport au premier de 1 seconde, c'est à dire que les segments sont recouverts à 50%.Ces deux groupes de segments, appelés respectivement: séquences paire et impaire forment la segmentation entrelacée.
Dans cette étude, nous supposons que le signal de parole est composé de 2n+1 segments numérotés, représentant une séquence impaire (1, 3, 5, 7,…, 2n+1) et une séquence paire (2, 4, 6, 8,…, 2n). Chaque segment est analysé comme suit : le signal de parole est décomposé en fenêtres de 32 ms avec un recouvrement de 50%. Puis, pour chaque fenêtre, une FFT est calculée pour le calcul des coefficients du banc de filtres et la matrice de covariance.

B. L'étiquetage

Une fois la matrice de covariance calculée, quelques mesures de distance sont utilisées pour trouver la référence la plus proche dans chaque segment (figure 1).

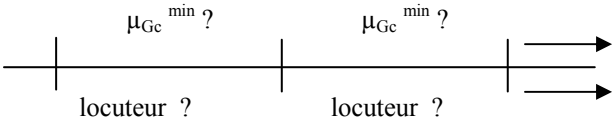

Fig. 1. Calcul de la distance minimale.

Une fois que la distance minimale entre les caractéristiques du segment et les caractéristiques de référence (ex. correspondant au locuteur L_j) est calculée, le segment est étiqueté par l'identité de cette référence (locuteur L_j). Ainsi, ce processus est appliqué jusqu'au dernier segment du flux de parole. Finalement, nous obtenons deux séquences étiquetées résultant de l'étiquetage pair et impair (figure 2), auquel on appliquera l'algorithme ISI.

C. ISI: Indexation entrelacée de la parole

L'algorithme ISI est une technique nouvelle dans laquelle deux segmentations (une décalée par rapport à l'autre) et une règle de correction logique est utilisée pour trouver les meilleures étiquettes du locuteur (figure 2).
En effet, ayant deux séquences d'indexation

différentes, nous essayons de donner un compromis d'étiquetage raisonnable entre les deux étiquetages précédents.
Ainsi nous divisons chaque segment en deux autres segments similaires (d'une seconde chacun), appelés sub-segments, de telle sorte qu'on obtienne "2n" labels pairs (dénotés par $L'^{1/2'}_{even}$) et "2n+2" labels impairs (dénotés par $L'^{1/2'}_{odd}$) pour les sub-segments (pairs et impairs).
Cependant, notre intuition serait que les sub-labels pairs et impairs sur le même segment devraient être les mêmes, par conséquent, nous tenterons de comparer $L'^{1/2'}_{even}(j)$ avec $L'^{1/2'}_{odd}(j)$ pour chaque sub-segment j (j=2, 3, … 2n+1). Ici, deux cas sont possibles:

- **si** $L'^{1/2'}_{even}(j) = L'^{1/2'}_{odd}(j)$

  alors le label est correct:

  nouveau label = $L'^{1/2'}(j) = L'^{1/2'}_{even}(j) = L'^{1/2'}_{odd}(j)$ (4)

  où $L'^{1/2'}$ représente un sub-label.

- **si** $L'^{1/2'}_{even}(j) \neq L'^{1/2'}_{odd}(j)$

  alors il y a une confusion:

  nouveau label = $L'^{1/2'}(j) = Cf$ (5)

  où Cf signifie une confusion dans l'étiquetage.

Dans le cas de confusion, nous dérivons un nouvel algorithme de correction qu'on a appelé "correction ISI".

Algorithme de correction ISI:
Dans le cas de confusion, nous divisons les sub-segments correspondants (de 1 s) en deux autres sub-segments de 0.5 seconde chacun, appelés micro-segments. Leurs labels, appelés micro-labels, sont dénotés par $L'^{1/4'}$.
L'algorithme de correction est donc donné comme suit:

- **si** $\{ L'^{1/4'}(j) = Cf$ **et** $L'^{1/4'}(j+1) = Cf$ **et** $L'^{1/4'}(j-1) \neq Cf \}$

  alors $L'^{1/4'}(j) = L'^{1/4'}(j-1)$ (6)

  Ceci est appelé correction gauche (voir le micro-segment j_0 dans la figure 2),

- **si** $\{ L'^{1/4'}(j) = Cf$ **et** $L'^{1/4'}(j-1) = Cf$ **et** $L'^{1/4'}(j+1) \neq Cf \}$

  alors $L'^{1/4'}(j) = L'^{1/4'}(j+1)$ (7)

  $L'^{1/4'}$ *dénote un micro-label pour un micro-segment de 0.5s. Ceci est appelé correction droite (voir fig. 2).*

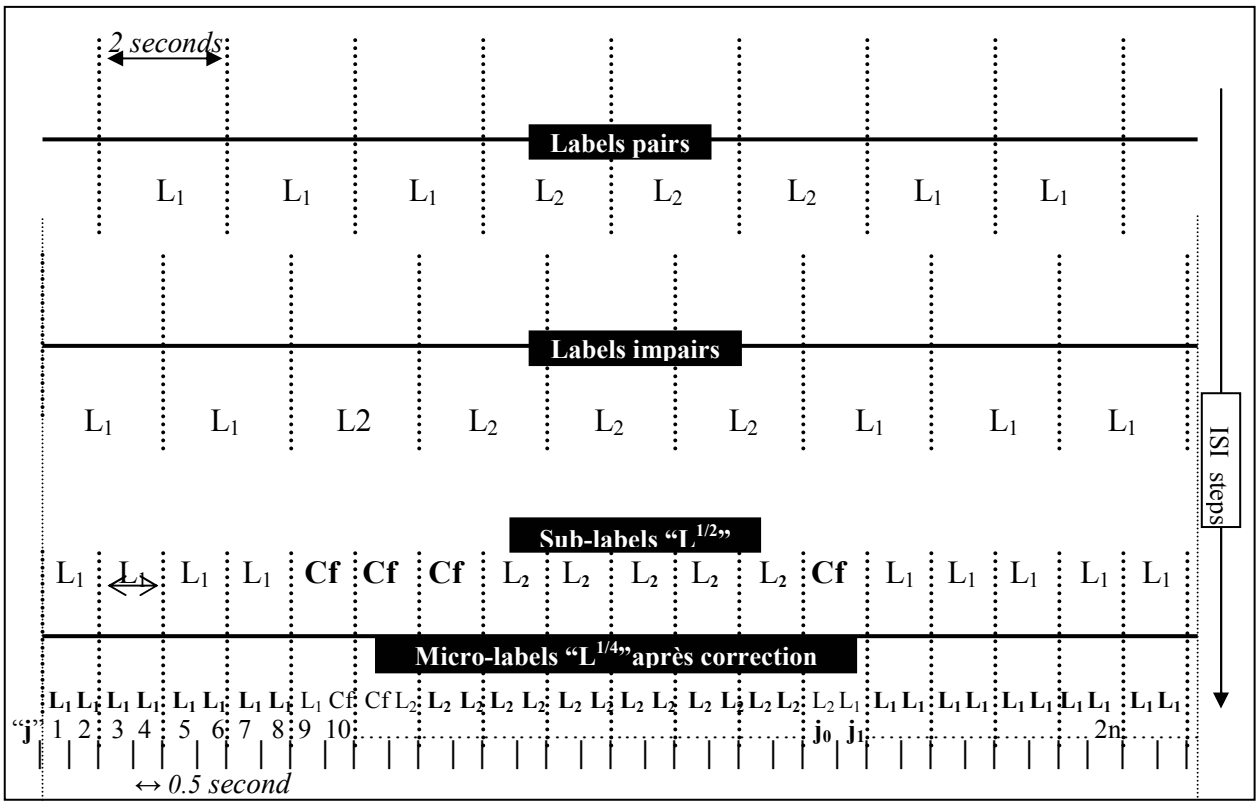

Fig. 2. Les étapes de l'algorithme ISI avec une itération. L_j représente le locuteur "j" et Cf signifie une confusion.

La correction ISI peut être utilise plusieurs fois (plusieurs itérations) pour raffiner progressivement la précision d'indexation. Dans notre application, nous avons utilisé cet algorithme avec 2 et 4 itérations.

Les expériences ont indiqué que la correction ISI permet de trouver la meilleure décision d'étiquetage, à partir des deux séquences d'étiquetage entrelacées, en réduisant efficacement l'erreur de suivi. De plus, la résolution de segmentation (résolution de L(j)), qui était de 2 secondes, est réduite à 0.5 seconde uniquement (résolution of $L^{1/4}(j)$), ainsi la performance apportée par la technique ISI est observable dans la précision du suivi et la résolution de la segmentation.

4. Resultats et discussion

A. Experiences sur TIMIT

La première base de données de test consiste en plusieurs prononciations de TIMIT prononcées par des locuteurs différents et concaténées en fichiers de durée allant de 30 à 130s chaque, de sorte que chaque fichier contienne une séquence de 2, 3, 5 ou 10 locuteurs différents. Dans le but d'étudier la robustesse de notre méthode, une partie de la base de données est mixée avec du bruit et de la musique. Tous les résultats sont résumés dans le tableau 1.
Sur ce tableau on remarque que l'erreur de suivi augmente quand le nombre de locuteurs augmente aussi: 5.3% avec 2 locuteurs contre 7.3% avec 3 locuteurs, etc...
Concernant les différents bruits ajoutés, on voit que le bruit humain ne perturbe pas significativement le suivi: une dégradation de seulement 4% à 12 dB. Par contre les autres types de bruit comme le bruit de fond et le bruit de bureau causent réellement une dégradation significative des performances du suivi.
Concernant le meilleur résultat obtenu, on arrive à atteindre une erreur de 5% pour le suivi de 2 locuteurs.

B. Experiences sur Hub4 Broadcast News

La deuxième base de données utilisée dans nos expériences est extraite de HUB-4 96-Broadcast-News et consiste en des news naturels enregistrés de la CNN. De plus, nous avons testé différents algorithmes de correction de post-suivi dans le but de les comparer (voir figure 3).

Cette figure 3 représente l'erreur d'indexation obtenue avec différentes mesures et différentes durées de segments. Nous remarquons que la meilleure mesure est la $\mu_{Gc\beta}$ donnant la meilleure performance de suivi. Par exemple pour une durée de segment de 3s la μ_{Gc1} donne une erreur de 10.4%, la $\mu_{Gc0.5}$ donne une erreur de 8.6% et la $\mu_{Gc\beta}$ donne la plus petite erreur: 7.7% (pourcentage de segments correctement étiquettes).

Sous un autre aspect, cette figure représente l'erreur de suivi obtenue avec ou sans correction ISI et pour différents nombres d'itérations. Nous pouvons remarquer que l'erreur obtenue après correction ISI est plus faible que celle obtenue sans correction ISI. Par exemple, si la durée des segments est de 3 secondes, l'erreur de suivi sans correction ISI est environ 9% mais elle décroît à 7.7% quand la correction ISI à 2 itérations est appliqué et décroît jusqu'à 7.6% seulement lorsque une correction ISI à 4 itérations est appliquée.

Table 1: Erreur d'indexation sur TIMIT (discussions entre plusieurs locuteurs :2, 3, 5 ou 10 locuteurs).

		Erreur d'indexation % pour des discussions entre:			
		2 locuteurs	3 locuteurs	5 locuteurs	10 locuteurs
▪ **Parole propre**	Avec detection de silence	7,2	8,1	7,9	10,3
	Sans detection de silence	5,3	7,3	5,9	8,0
▪ **Music + parole**	Sans detection de silence	4,8	6,6	7,5	9,1
▪ **Parole bruitée à 12 dB**	Bruit de fond	26,0	55,7	53,7	67,2
	Bruit de bureau	19,9	24,3	57,6	66,1
	Bruit humain	9,1	7,9	23,0	19,9
▪ **Parole bruitée à 6 dB**	Bruit de fond	32,8	58,4	64,7	79,1
	Bruit de bureau	28,1	37,7	63,4	70,6
	Bruit humain	11,8	12,9	15,5	24,3

Finalement en ce qui concerne la durée segmentale, la comparaison entre les différentes durées utilisées montre que la meilleure segmentation est obtenue pour une durée de segments de 3s (voir figure 3).

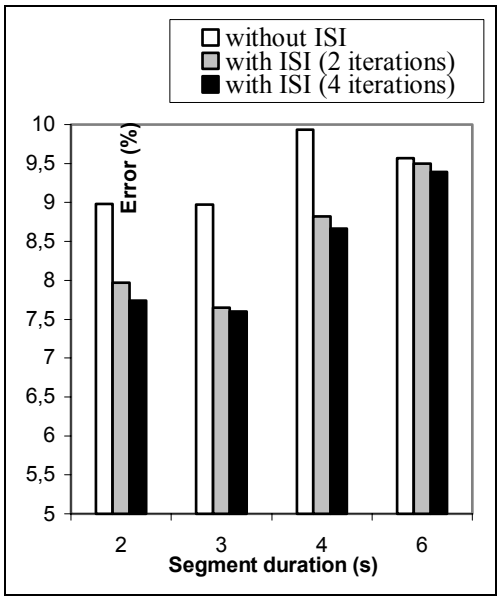

Figure 3 : Erreur d'indexation sur Hub4.

5. Travaux apparentés

Dans le même cadre du suivi du locuteur, Meignier [1] a utilisé un modèle de Markov progressif (durant le procédé d'indexation, le modèle change à chaque nouvelle détection de locuteur) et l'a testé sur un sous-ensemble de Switchboard. Magrin-Chagnolleau [2] a utilisé un modele GMM pour le suivi du locuteur cible. Cettolo [3] a utilisé une approche similaire (i.e un GMM pour modéliser chaque locuteur et chaque classe audio générique) sur le corpus italien Broadcast News. Une nouvelle technique/ règle de suivi présentée par Johnson [4] est basée en même temps sur un clustering agglomératif et divisif et est appliquée sur « the 1996 Hub 4 development data ».

Les performances rapportées dans les travaux cités ci-dessus varient largement selon l'approche, la tache et le corpus, et sont par conséquent difficiles à comparer avec les performances de notre approche conduite selon notre protocole expérimental adopté.
Cependant nous croyons que cette approche représente un bon compromis, puisque simple d'implantation, pas chère en calculs et a présenté de hautes performances durant les expériences.

6. Conclusions

Nous avons conçu une nouvelle technique pour le suivi automatique du locuteur, appelée ISI (*Interlaced Speech Indexing*), utilisant une segmentation équidistante entrelacée. Les expériences ont montré que l'association des mesures statistiques avec la technique ISI est très efficace: Bien que les mesures statistiques nécessitent des segments de durée 2s au minimum, ce qui signifie que la meilleure résolution segmentale est de 2s, l'association avec la technique ISI a permis de réduire cette résolution jusqu'à 0.5s seulement. Par ailleurs, cette nouvelle approche améliore considérablement la précision de suivi en corrigeant les erreurs de confusion. De plus quand on augmente le nombre d'itérations de l'algorithme ISI, l'erreur de suivi décroît continûment.
En résumé, deux grandes conclusions, dépendant de la baes de données, peuvent être déduites.

- Sur TIMIT, la meilleure performance est obtenue avec un score de 95% (pourcentage de segments correctement étiquettes), pour le cas non bruité.
Quand du bruit est ajouté, les expériences montrent que le score diminue avec le RSB. Cependant l'insertion de musique dans le signal de parole (par concaténation) n'altère pas le score de suivi. Par contre, ces mêmes expériences montrent que l'erreur de suivi augmente quand le nombre de locuteurs augmente, ce qui est évident.

- Les expériences faites sur Hub4 indiquent que la meilleure mesure statistique est la $\mu_{Gc\beta}$ et que la meilleure durée de segments est 3s. Par ailleurs, la technique ISI s'avère très intéressante tant pour l'amélioration de la résolution segmentale que pour le raffinement de la précision de suivi.

Qu'en est-il de l'implémentation ? Comparée à des travaux similaires, notre méthode offre des performances acceptables, bien qu'il soit difficile d'en faire une comparaison très objective. En tout cas, nous sommes convaincus que cette nouvelle association représente une technique intéressante pour le suivi automatique du locuteur, puisqu'elle est facile à implémenter, non coûteuse en calculs et offrant de bonnes performances.

REMERCIEMENTS

Nous remercions les équipes de recherche de l'IRIT et du LIA pour leurs aides quant à l'achèvement de ce travail.

REFERENCES

[1] S. Meignier et al. Modèle de Markov évolutif pour les tâches de suivi de locuteurs. JEP'2000 conférence, Aussois France, June 2000, pages 69-72.
[2] I. Magrin-Chagnolleau, et al. Detection of Target Speakers in Audio Databases. Proceedings of ICASSP'99, March 1999, Phoenix, Arizona, USA, vol.2, pages 821-824.
[3] M. Cettolo. Speaker Tracing in a Broadcast News. The speaker Recognition Workshop, June 2001, Crete, Greece, pages 163-168.
[4] S.E. Johnson. Who Spoke When? – Automatic Segmentation and Clustering for Determining Speaker Turns. Proc. Eurospeech, Budapest, Sept. 99, vol. 5, pages 2211-2214.
[5] F. Bimbot, et al. Second-Order Statistical measures for text-independent Broadcaster Identification. Speech Communication, vol. 17, Num 1-2, Aug. 95, pages 177-192.
[6] J.F. Bonastre, et al. A speaker tracking system based on speaker turn detection for NIST evaluation. .ICASSP. Istanbul, June 2000.