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

Suivi Automatique du Locuteur Utilisant une Segmentation Equidistante.

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

*USTHB, Institut d'Electronique, BP 32 Bab Ezzouar, Alger, Algérie

ouamour@ifrance.com

sayoud@ifrance.com

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

Résumé : Une des tâches les plus importantes en suivi de locuteur est la segmentation de la parole en zones délimitant l'identité des différents locuteurs parlant et les zones du silence. Dans cette étude, nous avons proposé une nouvelle méthode basée sur une segmentation équidistante à pas glissant pour ensuite indexer les différentes zones de parole, dans une fin de suivi de locuteur.
Les tests ont été faits sur des discours de TIMIT, comprenant deux à dix locuteurs différents dans chaque discours. A titre indicatif, le taux d'erreur moyen obtenu, en milieu sain est de 5%.

Mots clés : Suivi du locuteur, Indexation de document audio, segmentation équidistante, mesures de similarité.

1 Introduction

Lors de conversations multi-locuteur, telles que les interviews télévisées, on cherche à suivre automatiquement le locuteur parlant durant la discussion (par exemple par caméra). Ce problème nous a incité à développer un système automatique pour le suivi de locuteur (Bonastre & al, 2000) basé sur une segmentation équidistante à pas glissant. Les tests qui ont été effectués sur des discours comprenant des locuteurs de TIMIT (Fisher & al, 1986), avec et sans considération des zones de silence (dans le discours) ont donné des résultats intéressants.

2 Méthode d'identification utilisée

Notre méthode se compose en deux étapes : l'apprentissage et le tes de suivi.
Durant l'apprentissage, on procède à l'extraction des MFSC ou Mel-énergies (Dauddington, 1983), puis pour chaque prononciation on fabrique le vecteur moyenne x et la matrice de covariance X ; ainsi il existe une moyenne x pour chaque locuteur et une covariance X pour chaque locuteur. Le couple (x,X) représente la référence statistique d'ordre 2 pour le locuteur ***X*** (Bimbot & al, 1995) utilisé dans le dictionnaire des références.

En phase de test, une modélisation similaire de la phrase de test générera le couple (y,Y) représentant le modèle statistique de test pour le locuteur inconnu ***Y***.

L'identification est basé, alors, sur la distance minimale (plus proche voisin) au sens de la métrique statistique du 2e ordre: $\mu_{Gc0.5}$. La $\mu_{Gc0.5}$ appelée mesure de vraisemblance gaussienne symétrique à covariance [Bim95] est définie par :

$$\mu_{Gc0.5}(X,Y) = \frac{1}{2}\alpha - 1 \qquad (1)$$

avec

$$\alpha = \frac{1}{p}\left(tr\left(YX^{-1}\right) + tr\left(XY^{-1}\right)\right) \qquad (2)$$

p étant la dimension du vecteur des caractéristiques acoustiques et "tr" dénote la trace d'une matrice.

X représente la référence et ***Y*** représente le locuteur à reconnaître.

3 Méthode de suivi

3.1 Prétraitement du signal de parole

Le prétraitement est constitué des étapes d'analyse du signal de parole (Boite, 1987) précédant

l'apprentissage et les tests de suivi du locuteur. Ainsi, chaque discours est analysé selon les étapes suivantes.

- Découpage du signal de parole en zones de silence et d'activité en utilisant un détecteur d'activité basé sur l'énergie moyenne avec un pas d'analyse de 0.5 seconde (Nishida & al, 1998).
- Le signal de parole est ensuite décomposé en segments d'analyse de durée 2 secondes pour chaque segment, avec un recouvrement de 50%.
- Chaque segment est à son tour décomposé en fenêtres d'analyse de 512 échantillons.
- Une fenêtre rectangulaire est appliquée à chaque fenêtre avec un recouvrement de 50% (256 échantillons).
- La FFT pour chaque fenêtre fournit 256 valeurs représentant un spectre de Fourier (Coulon, 1987) dans la bande 0-8 kHz
- Finalement, les coefficients ainsi obtenus (notés $Y_k(m)$), pour chaque fenêtre "m" sont stockés dans des vecteurs de dimension p (p : étant le nombre de filtres du banc), appelés les MFSC ou *Mel Frequency Spectral Coefficients*.

3.2 Algorithmes de suivi

Dans cette étude, l'étape de suivi est basée sur un nouveau type de segmentation que nous avons appelé : Segmentation équidistante à pas glissant.

Le suivi automatique du locuteur est défini comme la reconnaissance de celui qui a parlé et quand est-ce qu'il a parlé, sur une zone de parole prononcée par un ou plusieurs locuteurs, dans un dialogue multi-locuteurs. Toutefois, une des taches les plus importantes en suivi de locuteur est la segmentation de la parole en zones délimitant l'identité des différents locuteurs parlant et en zones coïncidant avec le silence, avec une grande résolution temporelle et un grand taux d'identification. Les principaux algorithmes employés pour le suivi sont les suivants.

3.2.1 Algorithme d'apprentissage

L'apprentissage représente la phase permettant de construire le dictionnaire de référence des caractéristiques propres à chaque locuteur. La durée moyenne d'apprentissage (durée des phrases prononcées pour l'apprentissage) est de 9 secondes et les étapes de son algorithme sont les suivantes.

- Présentation de la liste des différents locuteurs participant dans le discours.
- Calcul de la matrice de covariance X pour chaque locuteur par la formule (2.1.b).
- Sauvegarde de la matrice X avec l'indice du locuteur, afin de construire le dictionnaire de référence.

3.2.2 Algorithme de suivi : *(La Segmentation équidistante à pas glissant)*

Le test correspond à la phase de suivi dans le but de segmenter et d'indexer tout le signal correspondant à la discussion proposée pour le suivi. Cet algorithme est décrit par les étapes suivantes.

- Détection des zones de silence et des zones d'activité (Nishida & al, 1998).
- Décomposition du discours en segments d'analyse équidistants de durée 2 secondes avec un recouvrement de 50%.
- Décomposition de chaque segment d'analyse en fenêtres d'analyse de 512 échantillons chacune (de durée 32 ms) avec un recouvrement de 50%.
- Extraction des coefficients MFSC (voir paragraphe précédent).
- Calcul de la matrice de covariance (Bellanger, 1987) du segment étudié.
- Comparaison de cette matrice avec l'ensemble des matrices de référence, en utilisant la mesure μ_{Gc} (voir section 2) et recherche de la distance minimale.
- Attribution de l'identité du locuteur reconnu au segment étudié.
- Correction Entrelacée des erreurs de confusion

Nous pouvons observer le principe général de notre algorithme de suivi sur les figures 1et 2.

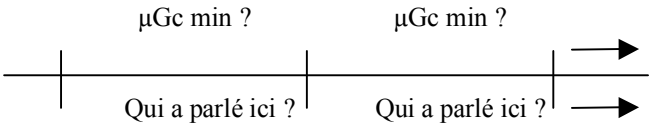

Figure 1: Segmentation équidistante et Recherche des distances minimales.

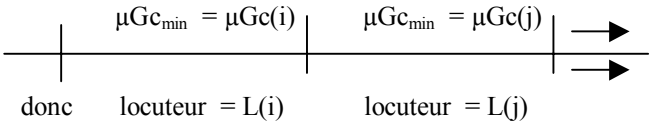

Figure 2: Indexation (étiquetage) et déplacement à pas constant.

Notons que la règle du plus proche voisin (distance minimale) est applicable ici, puisque nous travaillons dans un ensemble fermé, c'est à dire que la liste des locuteurs de test est la même que celle de référence.

4 Expériences

Nous avons utilisé des discours multi-locuteur extraits de la base de données parlée TIMIT. Ces discours sont organisés de la manière suivante :

- **15 discours bi-locuteurs :** dans chaque discours, chacun des locuteurs prononce 4 phrases différentes et la durée moyenne d'un discours est de 29 secondes environ.

Les 15 discours sont partagés en trois groupes différents répartis comme ci-dessous.

- **1er groupe :** composé de 5 discours, sachant que chaque discours correspond à une discussion entre deux locuteurs féminins.
- **2ème groupe :** composé de 5 discours, sachant que chaque discours correspond à une discussion entre deux locuteurs masculins.
- **3ème groupe :** composé de 5 discours, sachant que chaque discours correspond à une discussion entre un locuteur féminin et un locuteur masculin.

▪ **Et 9 discours multi-locuteurs :** qui sont partagés comme suit.

- **3 discours correspondant à 3 locuteurs différents** dans chacun d'eux et dont la durée moyenne de chaque discours est de 47 secondes environ. Ici, chaque locuteur prononce 5 phrases différentes.
- **3 discours correspondant à 5 locuteurs différents** dans chacun d'eux et dont la durée moyenne de chaque discours est de 74 secondes environ. Ici, chaque locuteur prononce 4 phrases différentes.
- **3 discours correspondant à 10 locuteurs différents** dans chacun d'eux et dont la durée moyenne de chaque discours est de 128 secondes environ. Ici, chaque locuteur prononce 4 phrases différentes.

Notons enfin, que les enregistrements sont pris avec un microphone de haute qualité dans un format de 16 bits et une fréquence d'échantillonnage de 16 kHz.
La durée moyenne des discours varie de 30 secondes à 130 secondes.
Le suivi du locuteur est fait en utilisant la mesure μ_{Gc} (Ouamour & al, 1999) (Sayoud & al,2000).

Les résultats de suivi du locuteur sont exposés dans les tableaux 1 et 2 suivants.

Tableau 1 : Taux d'erreurs moyens (avec considération de silence).

Nombre de locuteurs par discours	Taux d'erreur moyen %
2 locuteurs féminins	8.6
2 locuteurs masculins	7.17
2 locuteurs de sexes différents	5.67
2 locuteurs (en moyenne)	7.15
3 locuteurs	8.07
5 locuteurs	7.91
10 locuteurs	10.31

Tableau 2 : Taux d'erreurs moyens (sans considération de silence).

Nombre de locuteurs par discours	Taux d'erreur moyen %
2 locuteurs féminins	5.93
2 locuteurs masculins	5.59
2 locuteurs de sexes différents	4.39
2 locuteurs (en moyenne)	5.3
3 locuteurs	7.28
5 locuteurs	5.91
10 locuteurs	7.98

5 Observations

Cas de considération du silence :

• Cas des discours à 2 locuteurs :

Dans ce cas, nous nous sommes intéressés à suivre les locuteurs dans des discours prononcés par deux locuteurs différents en milieu sain avec détection du silence et avec une résolution de 0.5 seconde.
Les taux d'erreur sont partagés en 3 groupes : dans le 1er groupe nous faisons le suivi de 2 locuteurs féminins, dans le 2e groupe nous faisons le suivi de 2

locuteurs masculins et dans le 3e groupe nous faisons le suivi d'un locuteur féminin avec un locuteur masculin.
Les taux ainsi obtenus montrent que le meilleur taux moyen est obtenu dans le cas de 2 locuteurs de sexes différents (tableau 1), ceci peut être expliqué par le fait que la différence entre les caractéristiques de ces deux voix est grande, la valeur de ce taux est de 5.67% (tableau 1). Donc, nous pourrons conclure que le suivi de locuteur est meilleur quand les deux locuteurs sont de sexe différent. Donc, il est astucieux d'utiliser dans une interview bi-locuteurs, un journaliste féminin si le locuteur interrogé est masculin.
Le taux d'erreur moyen de tous les discours est de 7.15%, ce qui correspond à un taux de suivi de 92.85% et qui représente un résultat appréciable. Ainsi, nous pourrons dire que la paramétrisation MFSC utilisée pour la caractérisation des locuteurs convient bien en suivi du locuteur.

• Cas des discours multi-locuteur
Le 1er groupe de test correspond à des discours contenant 3 locuteurs à la fois dans chaque discours et le 2e groupe correspond à des discours contenant 5 locuteurs différents par discours. Nous remarquons que pour ces 2 groupes de test, le taux d'erreur moyen (tableau 1) reste assez faible, par contre pour le 3e groupe, où chaque discours contient 10 locuteurs à la fois, le taux d'erreur moyen augmente nettement et vaut 10.31% ; ceci est explicable par le fait que lorsque le nombre des locuteurs augmente, le nombre de transitions (changements des locuteurs) augmente et par conséquent le nombre de confusions augmente aussi.
Nous remarquons par ailleurs que si l'intervalle d'analyse ne coïncide pas parfaitement avec la zone de silence, ce dernier ne sera pas détecté puisqu'il sera partagé entre 2 zones d'activité, ce qui fait augmenter le taux d'erreur.
Toutefois, quand le nombre des locuteurs augmente, la probabilité de confondre 2 locuteurs augmente, ceci parce que la mesure μ_{Gc} est utilisée avec un protocole long-court, long apprentissage (9 secondes) et court test (2 secondes) et comme c'est montré dans les travaux de Bimbot, le taux d'identification n'est pas très bon pour ce protocole. Par conséquent, le taux d'erreur se voit à la hausse. Mais nous pourrons toutefois améliorer le taux d'erreur si la durée de l'intervalle d'analyse devient plus grande.

Quelles sont les causes des erreurs ?

Le taux d'erreur exprime la confusion entre 2 locuteurs faite par l'algorithme. Ce taux augmente dans les cas suivants :

- Quand le nombre des zones de transition entre les locuteurs est grand.
- Quand l'intervalle d'analyse ne coïncide pas avec la zone de silence, dans ce cas le silence va être partagé entre 2 zones d'activité.
- Quand le nombre de locuteurs est grand, donc le nombre de confusions entre les locuteurs augmente évidemment.

Cas de non considération du silence :
L'étude suivante correspond aux résultats des tests effectués, sans considération du silence, sur les discours bi-locuteurs et multi-locuteurs. Le tableau 2 présente les taux d'erreurs moyens dans le cas de 2 locuteurs, 3 locuteurs, 5 locuteurs et 10 locuteurs. Nous remarquons que le meilleur taux est celui des discours mixtes et qui vaut 4.39%, ce qui confirme le résultat précédent obtenu avec détection du silence. Nous remarquons aussi, que le taux moyen des discours à 2 locuteurs vaut 5.3% correspondant à 94.7% de bon suivi et qui est un résultat très encourageant pour le suivi. Par ailleurs, nous constatons que le taux d'erreur augmente quand le nombre de locuteurs augmente et ceci pour les mêmes raisons expliquées dans le cas précédent.
La dernière remarque à faire est que le taux d'erreur obtenu, dans le cas où nous ne détectons pas le silence, est meilleur que celui obtenu si les zones de silence sont prises en considération, car les erreurs de confusion du type locuteur/silence ne sont pas comptabilisées dans le deuxième cas.

6 Conclusion et Discussion

Dans cette étude, nous avons proposé une nouvelle méthode de suivi de locuteur basée sur une segmentation équidistante à pas glissant (avec une détection du silence de résolution 0.5 seconde sur la durée).

Dans un but comparatif, nous avons rassemblé tous les résultats obtenus dans un format graphique illustré par la figure 3.
Ainsi, nous remarquons aussi que le taux d'erreur du suivi augmente quand le nombre d'interlocuteurs augmente; ce qui est évident, car le risque de confusion augmente aussi. A titre d'exemple l'erreur est d'environ 5.3% pour les discours à deux locuteurs et augmente jusqu'à atteindre une valeur de 7.98% pour les discours à 10 locuteurs.
La méthode utilisée est plus fiable dans le cas où les discours contiennent des locuteurs de sexes différents. Ce qui nous incite à diversifier le sexe des locuteurs dans les conversations ou les interviews pour permettre un bon suivi de ces derniers.

Par ailleurs, nous constatons aussi que les résultats obtenus sans considération du silence, sont meilleurs que ceux obtenus en considérant le silence ; et ceci est dû au fait qu'on ne comptabilise pas les erreurs du type silence/parole.

Globalement on pense que les résultats obtenus sont encourageants, du fait qu'on arrive à suivre, avec une faible erreur de détection, les locuteurs parlant dans un débat ou une discussion multi-locuteur.

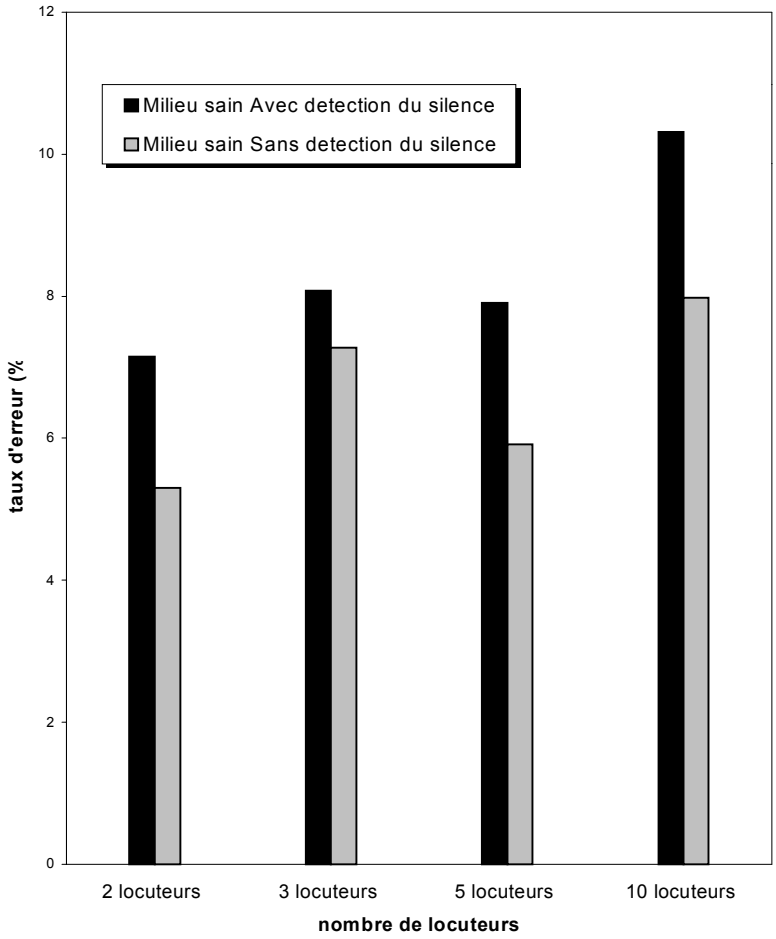

Figure 3: Taux d'erreur en fonction du nombre des locuteurs.

Références bibliographiques

(Bellanger, 1987) Bellanger, M., 1987. *"Traitement Numérique du Signal"*,. Collection technique et scientifique des télécommunications, édition MASSON, Paris.

(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.

(Boite, 1987) Boite, R., Kunt, M., 1987. *"Traitement de la Parole, complément traité d'électricité",* Presses Polytechniques Romandes, Lausanne, Paris, 1987.

(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.

(Coulon, 1987) De Coulon, F., 1987. Théorie et Traitement du Signal, Presses Polytechniques Romandes, Lausanne, 1987.

(Dautrich & al, 1983) B.A. Dautrich, L.R. Rabiner and T.B. Martin 1983, *"The effects of selected signal processing techniques on the performance of a filter bank based isolated word recognizer"*, Bell System Technical Journal, 1983.

(Doddington, 1998) G. R. Doddington 1998, *" Speaker Recognition Evaluation Methodology. An Overview and Perspectives"*, RLA2C Avignon, 20-23 April 1998, pp 60-66.

(Fisher & al, 1986) W. Fisher, V. Zue , J. Bernstein and D. Pallet 1986, *"An acoustic-phontic database"*, JASA, suppl. A, Vol. 81(S92) 1986.

(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.

(Ouamour & al, 1999) Ouamour, S., Kernouat, N., 1999. *"Système Robuste pour la Reconnaissance Automatique du Locuteur –Application à la Voix Téléphonique-"*. Thèse d'Ingéniorat, USTHB, Alger, Oct 1999.

(Sayoud & al, 2000) Sayoud, H., Ouamour, S. et al, 2000. *"Reconnaissance automatique du locuteur en milieu bruité"*. JEP'2000, pp 345-348, Aussois 19-23 juin 2000.

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