Séparation en locuteurs de conversations via IP

Daniel Moraru, Laurent Besacier

CLIPS-IMAG, Univ. J. Fourier ; équipe GEOD
BP 53 - 38041 GRENOBLE Cedex 9
Tél.: +33 (0) 4 76 63 56 95 – Fax: +33 (0) 4 76 63 55 52
Mél: {Laurent.Besacier, Daniel.Moraru}@imag.fr

ABSTRACT

In this paper we are interested in speaker segregation, meaning to recognize who speaks, and at which time, on an audio document containing the speech from several people. At first the theoretical ideas concerning our subject are presented. The signals which will have to be speaker segregated contain two-speaker conversations over IP. No statistical speaker or speech model is available *a priori*. The algorithm used is based on the Bayesian Information Criterion (BIC). This paper mainly brings contribution to evaluation procedures in this new field which is speaker segregation, especially when the speakers speak in the same time. For performance comparison, the VoIP database on which experiments are done is made available by the authors.

1. INTRODUCTION

En indexation automatique de documents multimédias, il existe différentes tâches liées au traitement de signaux audio et de la parole [Bon00], [Fon00] : sous titrage automatique d'un film, classement des éléments de la bande son d'un film, séparation parole/musique, détection des changements de locuteur. Cet article adresse la séparation en locuteurs des signaux audio. En d'autres termes, le but est de savoir qui parle, et à quel moment. Nous traitons les documents sonores en supposant qu'aucune information préalable sur les locuteurs n'est disponible (aucun modèle de locuteur ou de parole, aucune phase d'entraînement). La séparation en locuteurs peut être divisée en deux parties principales (figure 1). D'abord, l'étape de segmentation recherche des changements de locuteurs sur le signal. L'étape suivante, appelée regroupement, vise à regrouper les segments de parole liés à un même locuteur.

Dans la première partie de cet article nous présentons d'abord les aspects théoriques concernant notre sujet. Dans cette section nous discutons également de l'évaluation dans ce domaine nouveau qu'est la séparation en locuteurs. Dans la deuxième partie nous proposons une procédure qui permet d'avoir une segmentation de référence lorsque plusieurs locuteurs parlent en même temps. Finalement nous présentons les résultats de nos expérimentations. A la fin de l'article, nous montrons également la nécessité d'évaluer séparément les deux étapes de séparation en locuteurs afin d'avoir une idée précise du comportement du système complet.

Les données utilisées sont des enregistrements de dialogues via IP (NetMeeting ; toujours exactement 2 personnes qui parlent) obtenus pendant le projet européen NESPOLE [Bur01] de traduction automatique de la parole. Comme les résultats présentés sont les premiers obtenus sur cette base, les auteurs de cet article proposent de la mettre à disposition d'autres laboratoires (sur simple demande), afin que ces résultats puissent être comparés à d'autres obtenus avec des techniques différentes

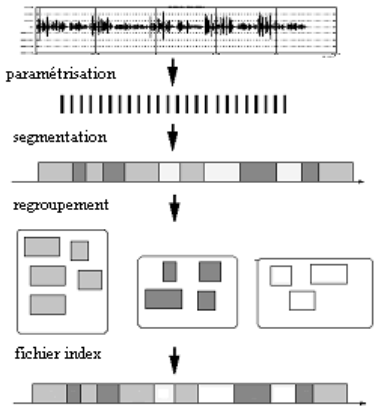

Figure 1: Le processus complet de séparation en locuteurs

2. SEPARATION EN LOCUTEURS

Cette partie présente les aspects théoriques des deux étapes de la séparation en locuteurs de signaux audio : d'abord la segmentation et ensuite le regroupement des segments issus de la première phase.

2.1 Segmentation à l'aide du critère d'information Bayesien (BIC)

En considérant que nous avons deux segments consécutifs de vecteurs de paramètres spectraux acoustiques (ex: coefficients MFCC, LPCC, etc): x_n, $n = 1 \ldots N_1$ et y_n, $n = 1 \ldots N_2$.

Nous devons tester en utilisant le critère BIC (Bayesian Information Criterion) les deux hypothèses suivantes:

- H_0 : les séquences ont été prononcées par le même locuteur ; i.e. elles ont été générées par un seul modèle gaussien (μ, Σ),
- H_1 : les séquences ont été prononcées par des locuteurs différents ; i.e. elles ont été générées par deux modèles gaussiens différents: (μ_1, Σ_1), (μ_2, Σ_2)

Si nous considérons que λ dénote le rapport de vraisemblance [Gis91] entre les deux hypothèses, alors nous obtenons:

$$\lambda = \frac{L(z; \mu, \Sigma)}{L(x; \mu_1, \Sigma_1) \cdot L(y; \mu_2, \Sigma_2)}$$

où z est l'union des deux séquences x et y

Nous estimons le logarithme de ce rapport $R = -\log \lambda$ à l'aide de deux fenêtres glissantes [Bon00] déplacées le long du signal audio paramétrisé (figure 2) et nous obtenons une séquence de valeurs de $R(t)$.

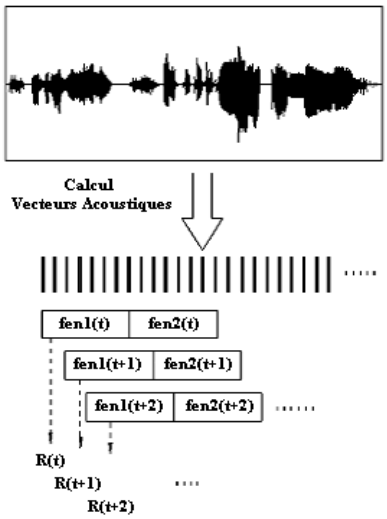

Figure 2: Fenêtres glissantes pour la détection de changements de locuteur

Ensuite la détection de points de changement de locuteur consiste à chercher des maximums sur la courbe $R(t)$ que nous appellerons par la suite *la courbe BIC*.

2.2 Regroupement Hiérarchique

Après avoir détecté les changements de locuteurs et avoir obtenu des segments contenant le discours prononcé par un et seulement un locuteur, la seconde étape de la séparation en locuteurs est de regrouper les différents segments de signal appartenant à un même locuteur.

Dans notre cas nous connaissons *a priori* le nombre de locuteurs dans chaque enregistrement. Un regroupement par agglomération est alors utilisé.

Pour le calcul de distance entre les classes C_m et C_n le critère BIC est à nouveau utilisé :

$$d_{BIC}(C_m, C_n) = \frac{L(C_m; \mu_m, \Sigma_m) \cdot L(C_n; \mu_n, \Sigma_n)}{L(C_{m\cup n}; \mu_{m\cup n}, \Sigma_{m\cup n})}$$

où (μ_m, Σ_m) sont les paramètres de modèle gaussien de la classe m, (μ_n, Σ_n) sont les paramètres de modèle gaussien de la classe n et $(\mu_{m\cup n}, \Sigma_{m\cup n})$ sont les paramètres de modèle gaussien de la réunion des deux classes.

Cette fois le rapport de vraisemblance est inversé pour être homogène à une distance. Une valeur de rapport petite signifie que les deux classes sont proches.

Le processus s'arrête quand on a obtenu le nombre final de classes (locuteurs) désiré (2 dans notre cas).

2.3 Méthode d'Evaluation

L'évaluation pour la tâche de séparation en locuteurs n'est pas du tout un problème facile. Les difficultés pour l'évaluation de la séparation en locuteurs sont dues d'une part à la difficulté d'avoir une bonne référence et d'autre part à la façon de présenter les résultats.

La référence pour ce type de tâche est difficile à obtenir parce que la précision avec laquelle l'oreille humaine détecte les changements de locuteurs est assez faible. Pour la tâche de séparation en locuteurs il faut d'abord évaluer la segmentation et après évaluer le regroupement.

Une bonne séparation en locuteurs doit fournir des changements de locuteurs corrects et des segments ne contenant que la parole d'un seul locuteur.

On peut donc définir pour l'évaluation de la segmentation le Taux de Fausses Alarmes (TFA) et le Taux de Détections Manquées (TDM) comme suit:

$$TFA = 100 \times \frac{Nb.de\,FA}{Nb.de\,changements\,réels + Nb.de\,FA}$$

$$TDM = 100 \times \frac{Nb.de\,DM}{Nb.de\,changements\,réels}$$

On a choisi cette façon de calculer le TFA en rapportant le nombre de fausses alarmes au nombre de changements réels plus le nombre de fausses alarmes pour rendre les deux taux d'erreurs comparables.

Par ailleurs, il est important d'avoir des segments "purs" à la fin de la segmentation. Un segment qui contient la parole de plusieurs locuteurs peut détériorer le résultat du regroupement. Au contraire une fausse alarme peut être corrigée pendant le regroupement. Pour évaluer le regroupement il faut en fait mesurer la qualité des partitions obtenues.

Soit N_S le nombre total de locuteurs réellement présents, N_C le nombre total de classes obtenues par le système automatique et N_U le nombre total de trames. Le nombre total de classes peut être différent ou non du nombre de locuteurs (cela dépend si l'on connaît ou non *a priori* le

nombre de locuteurs). Soit n_{ij} le nombre de trames mises dans la classe i et prononcées par le locuteur j. On a donc $n_{.j}$ le nombre total des trames qui ont été prononcées par le locuteur j et $n_{i.}$ la taille de la classe i trouvée par le système:

$$n_{.j} = \sum_{i=1}^{N_C} n_{ij} \; ; \; n_{i.} = \sum_{j=1}^{N_S} n_{ij}$$

Maintenant on peut définir la pureté [Sol98] d'une classe i comme:

$$p_i = \sum_j \frac{n_{ij}^2}{n_{i.}^2}$$

La pureté d'une classe peut en fait être interprétée comme la probabilité que deux trames prises au hasard dans cette classe appartiennent effectivement à un même locuteur ; ou comme une approximation du pourcentage moyen des trames d'une classe qui appartiennent à un même locuteur.

3. EXPERIMENTATION ET RESULTATS

3.1 Base de signaux utilisée

Les données utilisées sont des conversations via IP (enregistrés avec le logiciel de vidéoconférence NetMeeting) en français obtenues pendant le projet européen NESPOLE [Bur01] concernant la traduction automatique de la parole (parole réelle, codée, les locuteurs peuvent parler en même temps, 15 à 20 min par enregistrement, 2 locuteurs par enregistrement, 19 enregistrements au total). Les enregistrements sont disponibles sur simple demande pour les personnes intéressées. La base NESPOLE contient également des enregistrements de dialogues dans d'autres langues.

Chaque dialogue a été enregistré en stéréo et contient la parole d'un locuteur sur chaque voie. Ceci nous a permis de faire un étiquetage de référence en utilisant les voix séparées, puis de tester notre système sur le signal mono résultant du mixage entre les deux voies séparées.

3.2 Obtention de l'étiquetage de référence

Nos enregistrements de départ sont des enregistrements stéréo, chaque locuteur parle sur une des voies. Après l'application d'un système de détection de silence sur chacune des voies, nous obtenons les segments de parole de chaque locuteur. Le résultat de la détection de silence peut être vu comme une courbe de 0 (pour les trames de silence) et de 1 (pour les trames de parole).

Quand les deux locuteurs parlent en même temps on peut faire deux étiquetages de référence : un où l'on considère tous les changements de locuteurs y compris ceux où l'on passe d'un seul locuteur à deux locuteurs et un où l'on considère seulement les changements de locuteur situés à la frontière entre des segments appartenant à un seul et seulement un locuteur et un segment appartenant à un autre et seulement un autre locuteur. On peut obtenir les changements de locuteurs dans les deux situations précédentes très facilement à partir des courbes de détection de silence sur les deux voies. A partir des deux courbes, on va construire deux références qu'on va appeler courbe OR et courbe XOR. Les courbes OR et XOR sont en fait un simple OR respectivement XOR (OU exclusif) logique sur les courbes de détection de silence.

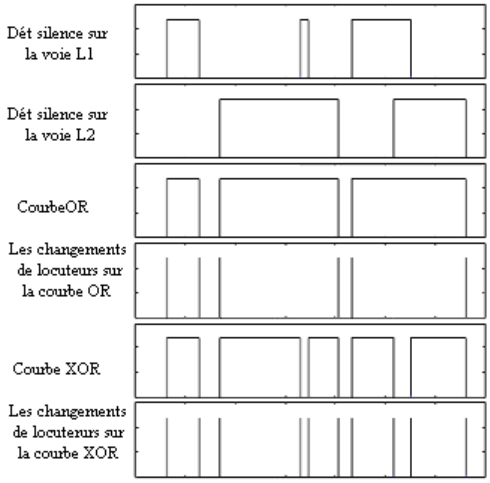

Figure 4: Obtention de l'étiquetage de référence sur nos données

On voit sur la courbe OR, que seuls les passages entre les deux locuteurs sont visibles alors que sur la courbe XOR on peut voir aussi les passages entre un segment où il y a un seul locuteur qui parle et un segment où il y a deux locuteurs qui parlent. Nous présenterons les résultats de l'étape de segmentation seule, évalués avec ces deux types de références.

3.3 Paramétrisation et modélisation du signal de parole

Le signal de parole est paramétrisé en utilisant 16 coefficients MFCC [Rab93] calculés toutes le 10ms sur des fenêtres de 20 ms. Pour la phase de regroupement, nous utilisons en plus l'énergie de chaque fenêtre de 20 ms de signal. La distance est calculée seulement sur les trames qui ont suffisamment d'énergie. On évite ainsi d'utiliser des trames de silence pour le calcul de la distance.

Pour le calcul de la courbe BIC nous utilisons pour les deux fenêtres (cf figure 2) des modèles mono-gaussiens avec des matrices diagonales. Etant donné que les interventions de chaque locuteur sont parfois très courtes la taille de fenêtres a été choisie d'environ 1 seconde. Les modèles pour le regroupement sont aussi mono-gaussiens mais cette fois avec des matrices pleines. Avec ces modèles mono-gaussiens, la totalité des 19 signaux (environ 5h de signal) est traitée (système complet de séparation) en 6h avec un Pentium III à 735 MHz, soit environ 1.2 fois le temps réel.

3.4 Résultats

Pour ces expériences, nous avons utilisé la plate-forme ELISA développée par le consortium ELISA [Eli01] et utilisée pour les évaluations NIST de systèmes de vérification du locuteur.

Pour évaluer les performances de segmentation, nous considérons deux valeurs de tolérance autour des points de changement de locuteur de respectivement 0,5 et 1,5 secondes. Les taux d'erreurs (TFA et TDM) ont été calculés pour les deux références et pour les deux tolérances.

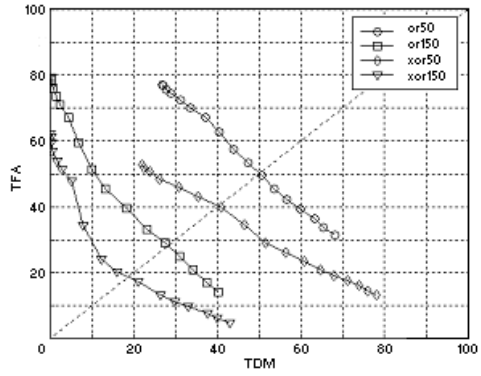

Figure 5: Résultats de la segmentation

La référence XOR contient en plus de la référence OR, les changements entre des segments où il y a deux locuteurs qui parlent en même temps. En conséquence si on regarde les courbes ROC on voit que pour un même TDM et une même tolérance le TFA est plus grand pour la référence OR. De même, pour un TFA fixé on voit que le TDM est plus petit pour la référence XOR. Cela pourrait signifier que notre système détecte aussi les changements de locuteur entre des segments où il y a deux locuteurs qui parlent. On voit qu'il y a une grande différence entre les taux d'erreurs pour les deux tolérances et cela veut dire qu'un certain nombre de changements de locuteur trouvés sont situés à une distance entre 0,5 et 1,5 secondes (les deux tolérances) des changements de locuteur de référence.

En ce qui concerne le regroupement, la pureté moyenne obtenue est de 79.04%. La pureté a été calculée, comme dans le cas des évaluations NIST [Nis01] seulement pour les segments où les deux locuteurs ne parlent pas en même temps.

Dans notre évaluation, nous avons donc évalué aussi la segmentation seule, ce que ne permet pas le scoring NIST. D'ailleurs, pour savoir si les performances de segmentation ont une influence sur la qualité finale de la partition en locuteurs obtenue, nous avons calculé séparément, pour chacun des 19 enregistrements le taux d'égale erreur de segmentation (EER) pour lequel TFA=TDM, ainsi que la pureté. Le coefficient de corrélation obtenu entre les 19 valeurs d'EER et les 19 valeurs de pureté est de 0.07 pour une tolérance de 0.5s autour du point de changement de locuteur, et de -0.17 pour une tolérance de 1.5s. Il semblerait donc que l'étape de segmentation, avec ses performance actuelles, n'a pas une influence significative sur la performance finale du système de séparation en locuteurs, évalué par la pureté. Ceci tendrait à confirmer que, lorsqu'on évalue un système de séparation en locuteurs avec un critère homogène à une pureté (comme cela est fait dans les évaluations NIST), la phase de segmentation n'est peut être pas l'étape la plus critique et qu'elle peut éventuellement être remplacée par une initialisation du système avec des segments courts de taille fixe, avant le regroupement.

4. CONCLUSIONS

Dans ce papier nous avons présenté et évalué un système de séparation en locuteurs de conversations via IP. Les résultats ont montré que le critère BIC, simple à mettre en œuvre et pratiquement temps-réel si on utilise une modélisation monogaussienne, donne des résultats corrects à l'issue du regroupement (pureté moyenne d'environ 80%). Cependant, il semblerait que la performance de segmentation (détection de points de changement de locuteurs), n'a pas une influence significative sur la pureté finale de la partition en locuteurs. Nous pensons cependant que trois indicateurs (TFA, TDM, pureté) sont nécessaires pour une évaluation complète d'un système de séparation en locuteurs.

BIBLIOGRAPHIE

[Gis91] Gish H., Siu M.-H., Rohlicek R. (1991), " Segregation of Speakers for Speech Recognition and Speaker Identification ", ICASSP 91, pp. 873-876.

[Rab93] Rabiner L., Juang B.-H. (1993), Fundamentals of Speech Recognition, Prentice Hall, 1993.

[Sol98] Salomonoff A., Mielke A., Schmidt M., Gish H. (1998), " Clustering Speakers By Their Voices ", Proc. of ICASSP 98, pp. 757-760

[Bon00] Bonastre J. F., Delacourt P., Fredouille C., Merlin T., Wellekens C., (2000)," A Speaker Tracking System Based on Speaker Turn Detection for NIST Evaluation ", ICASSP2000, pp 1177-1180

[Fon00] Fontaine L., Sénac C., Vallès-Parlangeau N., André-Obrecht R. (2000) " Indexation de la bande sonore : les composantes Parole/Musique ", JEP'2000, pp. 65-68. Aussois, France.

[Bur01] Burger S., Besacier L., Coletti P., Metze F., Morel C. (2001), " The NESPOLE! VoIP Dialogue Database ", Eurospeech 2001, Aalborg, Danemark.

[Eli01] The ELISA Consortium (2001), " Overview of the ELISA Consortium Research Activities ", Proc. of A Speaker Odissey, pp. 67-72, 2001.

[Nis01] http://www.nist.gov/speech/tests/spk/