SEGMENTATION ET INDEXATION PAR LOCUTEURS D'UN DOCUMENT AUDIO

Perrine Delacourt

Institut EURECOM, 2229 route des Crêtes, 06904 Sophia Antipolis, France
perrine.delacourt@eurecom.fr
http://www.eurecom.fr/~delacour

RESUME

Mon travail de thèse consiste à segmenter et indexer par locuteurs des documents audio. En d'autres termes, il s'agit de reconnaître la séquence de locuteurs présents dans la conversation. Ce travail est réalisé avec les hypothèses suivantes : aucune connaissance a priori sur les locuteurs n'est disponible, le nombre de locuteurs est inconnu et les personnes ne parlent pas simultanèment. Notre système d'indexation se décompose en trois parties principales : la segmentation en locuteurs, le regroupement des segments appartenant à un même locuteur, la construction des modèles de locuteurs en ligne et l'utilisation de ces modèles pour le raffinement de la segmentation et la reconnaissance de la séquence de locuteurs.

1. INTRODUCTION

Avec l'augmentation du volume d'archives sonores (radio et télévision) ces dernières années, il devient désormais indispensable d'indexer automatiquement ces documents pour être exploitables. La clé d'indexation qui nous intéresse ici est l'identité du locuteur : nous voudrions savoir qui parle et quand. Le système d'indexation par locuteurs peut servir également comme étape préliminaire pour des tâches de transcription [1], [2] ou pour le suivi de locuteurs [3]. Concernant la transcription, les taux de reconnaissance de parole sont améliorés quand les modèles de parole sont adaptés aux locuteurs. L'indexation préalable par locuteurs permet alors d'utiliser les données de parole de chaque locuteur présent dans la conversation pour adapter ces modèles. Quant au suivi de locuteur, l'utilisation du système d'indexation a pour conséquence de prendre la décision d'identification du locuteur cible sur des segments contenant plus d'informations que quelques trames comme cela est fait traditionnellement.

2. LE SYSTÈME D'INDEXATION

Le système d'indexation par locuteurs est composé de trois étapes principales, comme décrit figure 1. La première étape consiste à segmenter le signal de parole paramétrisé en locuteurs, i.e. obtenir les segments les plus longs possibles et homogènes en termes de locuteurs. La deuxième étape consiste à regrouper les segments appartenant à un même locuteur. Enfin, le but de la troisième étape est de construire un modèle de locuteur à partir de chaque groupe de segments résultant de la deuxième étape et d'utiliser les modèles de locuteurs ainsi obtenus pour raffiner la segmentation et pour reconnaître la séquence de locuteurs.

Ce travail est financé par le Centre National d'Etudes des Télécommunications (CNET) par le contrat n^o 98 1B

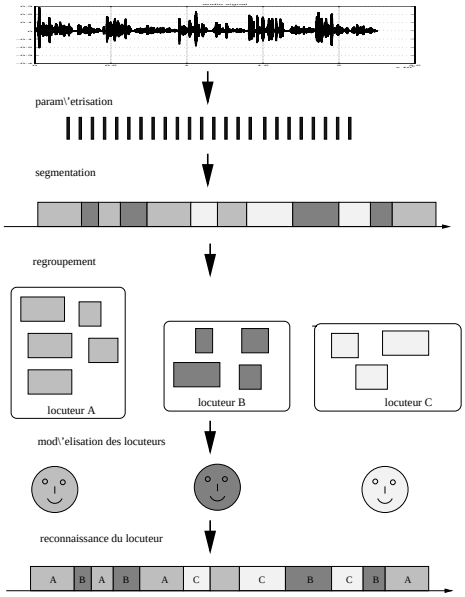

FIG. 1 – *Système d'indexation par locuteurs*

3. LA SEGMENTATION EN LOCUTEURS

La segmentation en locuteurs revient à détecter les changements de locuteurs. La technique de segmentation que je propose est réalisée en deux passes : la première détecte les changements de locuteurs les plus probables et la seconde les valide ou non. Cet algorithme est décrit en détails dans [4].

Segmentation basée sur le calcul d'une distance La première passe de l'algorithme repose sur le calcul d'une distance entre deux portions de signal. Une forte valeur de cette distance indique que ces deux portions de signal paramétrisé ont été générées par deux locuteurs différents. A l'inverse, une faible valeur signifie que les deux portions ont été prononcées par un même locuteur. La mesure de distance que nous utilisons est le rapport de vraisemblance généralisé proposé par H.Gish en identification du locuteur [5, 6]. Cette distance est calculée pour une paire de fenêtres

de signal de parole adjacentes. Cette paire de fenêtres est décalée à l'itération suivante et une nouvelle distance est calculée. Ce processus est répété jusqu'à avoir parcouru la totalité du signal à traiter. Ainsi, nous obtenons une courbe de distances, dans laquelle nous détectons les maxima, qui correspondent aux points de changements de locuteurs potentiels.

Raffinement à l'aide du Critère d'Information Bayesien (CIB) La détection des changements de locuteurs utilisée a été conçue pour minimiser le nombre de détections manquées (un changement n'est pas détecté alors qu'il existe) mais cela se fait au détriment du nombre de fausses alarmes (un changement est détecté alors qu'il n'existe pas). Pour réduire ce nombre de fausses alarmes, un raffinement de la segmentation est nécessaire et c'est l'objet de cette seconde passe. Cette passe utilise le critère d'information Bayesien (CIB). Ce critère est un critère de vraisemblance pénalisé par la complexité des modèles utilisés pour les segments. Nous l'appliquons sur des paires de segments résultant de la première passe. Si le critère est vérifié alors le changement de locuteur est validé et vice-versa. Ce critère a été proposé par S.Chen dans [7] pour la segmentation en locuteurs.

4. LE REGROUPEMENT EN LOCUTEURS

L'étape suivante consiste à regrouper les segments appartenant à un même locuteur, i.e le "cluster" correspondant à un locuteur donné ne doit contenir que les segments appartenant à ce locuteur et tous les segments relatifs à ce locuteur doivent se trouver dans ce même "cluster". Nous décrivons deux méthodes possibles.

Le "clustering" hiérarchique Le principe du"clustering" hiérarchique est expliqué dans [8]. Nous détaillons ici le "clustering bottom-up" : à chaque itération, les deux "clusters" (i.e. groupes de segments) les plus proches au sens de la distance considérée sont regroupés. Ce processus est répété jusqu'à ce qu'un critère d'arrêt soit atteint. Il y a donc deux paramètres à choisir : la distance (le critère de regroupement) et le critère d'arrêt du processus. Parmi les critères de regroupement déjà utilisés, nous pouvons citer le critère de vraisemblance généralisé ou encore le rapport de vraisemblance croisé [9]. Puisque le nombre de locuteurs est inconnu, le critère d'arrêt n'est pas évident. Deux solutions existent : soit le processus continue jusqu'à n'obtenir qu'un seul "cluster" et la partition idéale est ensuite choisie dans l'arbre de regroupement obtenu, soit la distance est pénalisée, i.e. elle ne doit pas dépasser un certain seuil pour que les deux "clusters" les plus proches soient regroupés. Ce type de "clustering" ne prend pas en compte les relations temporelles entre les différents segments.

Le "clustering" séquentiel prend en compte les relations temporelles entre les différents segments. Les segments résultant de la segmentation sont considérés dans l'ordre temporel : un premier "cluster" est créé avec le premier segment. Puis le segment suivant est examiné : s'il est suffisamment proche du "cluster" au sens de la distance choisie alors il est ajouté à celui-ci, sinon un nouveau "cluster" est créé, et ainsi de suite, jusqu'à avoir examiné tous les segments. Le critère d'arrêt est ici évident. Par contre, le critère de regroupement est semblable à une distance pénalisée.

5. LA MODÉLISATION ET LA RECONNAISSANCE DES LOCUTEURS

La dernière étape consiste à construire des modèles de locuteurs à partir de chaque "cluster" obtenu. Les modèles de locuteurs peuvent être assez sophistiqués puisque chaque "cluster" est censé contenir toutes les données d'un locuteur. Ces modèles sont ensuite utilisés pour reconnaître la séquence de locuteurs intervenant dans la conversation, à l'aide de techniques classiques d'identification du locuteur [10]. Pour plus de détails, consulter [11].

6. REFERENCES

[1] J.-L. Gauvain, L. Lamel, and G. Adda, "Partitioning and transcription of broadcast news data," in *ICSLP*, 1998.

[2] P. Woodland et al., "The development of the 1996 HTK broadcast news transcription system," in *DARPA Speech Recognition Workshop*, 1997.

[3] A. E. Rosenberg et al., "Speaker detection in broadcast speech databases," in *ICSLP*, 1998.

[4] P. Delacourt, D. Kryze, and C. Wellekens, "Detection of speaker changes in an audio document," in *EUROSPEECH*, 1999.

[5] H. Gish, M.-H. Siu, and R. Rohlicek, "Segregation of speakers for speech recognition and speaker identification," in *ICASSP*, pp. 873–876, 1991.

[6] H. Gish and N. Schmidt, "Text-independent speaker identification," *IEEE Signal Processing Magazine*, oct. 1994.

[7] S. Chen and P. Gopalakrishnan, "Speaker, environment and channel change detection and clustering via the Bayesian information criterion," in *DARPA speech recognition workshop*, 1998.

[8] R. Duda and P. Hart, *Pattern classification and scene analysis*. John Wiley and Sons, Inc., 1973.

[9] D. Reynolds et al., "Blind clustering of speech utterances based on speaker and language characteristics," in *ICSLP*, 1998.

[10] S. Furui, "An overview of speaker recognition technology," in *ESCA Workshop on Automatic Speaker Recognition, Identification and Verification*, 1995.

[11] P. Delacourt, , and C. Wellekens, "A first step into a speaker-based indexing system," in *Workshop on Content-Based Multimedia Indexing*, 1999.