Outils de navigation dans les fichiers audio

Chris J. Wellekens

Department of Multimedia Communications
Institut Eurécom, France
christian.wellekens@eurecom.fr

Résumé

Les services audio de la nouvelle génération requièrent des outils d'édition des fichiers audio qui permettent de traiter un tel fichier aussi simplement qu'un fichier texte. L'indexation des fichiers est une solution permettant l'accès rapide à l'information qui peut être l'identité d'un locuteur, la localisation de son intervention dans le temps, le sujet de la conversation. D'excellents outils de traitement de texte existent depuis de nombreuses années et une solution pour la navigation aisée dans un fichier audio pourrait se réduire à la reconnaissance du contenu entier du fichier à éditer (parole vers texte) mais requiert en général des reconnaisseurs de grands vocabulaires et indépendants des locuteurs qui fournissent actuellement des résultats acceptables seulement pour des locuteurs coopératifs limitant le thème de leur intervention au domaine applicatif correspondant à un modèle de langage prédéfini. Et même dans ce cas, la détection des interludes musicaux, l'identification d'un intervenant et la segmentation de l'enregistrement en zones correspondant à un seul locuteur restent des problèmes d'intérêt.

La maitrise complète des techniques d'indexation ouvrira un grand marché d'applications attrayantes pour le consommateur telles que la production de programmes audio (et aussi video) à la demande mais aussi pour le monde professionnel de la radiophonie qui pourrait composer du contenu par réutilisation.

Il tombe aussi sous le sens que de tels outils faciliteront l'accès des bases de données et des archives multimedia.

1. Introduction

Depuis de nombreuses années, les éditeurs de texte offrent une panoplie d' outils essentiels telles que "*chercher*", "*couper*" et "*coller*" dont aucun utilisateur ne peut se passer aujourd'hui. L'accroissement des capacités de stockage et le progrès en réseaux de communication joints aux algorithmes de compression ont fait flamber les besoins en et les utilisations de l'information numérique audio et video qui sont les modes les plus naturels d'interaction entre les humains et leur environnement.

Nous sommes à présent confrontés au problème de l'accès rapide à l'information contenue dans les fichiers multimedia.

Les fichiers audio sont sans doute plus difficiles à manipuler sur le plan du traitement du signal. En effet, le système auditif humain est à ce point sophistiqué que nous sommes capables de comprendre la parole et de reconnaitre la voix d'une personne mais dans des conditions de bruit élevé. Un modèle du système auditif est donc extrêmement difficile à construire Le processus de l'audition est encore dans une large mesure mal maitrisé. Il suffit de se rendre compte des efforts entrepris pour le développement de prothèses cochléaires et les difficultés rencontrées pour produire un semblant de conversation pour les sourds profonds.

Les signaux video bi-dimensionnels sont paradoxalament plus faciles à traiter par exemple en termes de segmentation mais la difficulté réside dans une plus difficile interprétation sémantique due à l'extrême variabilité d'une scène: orientations différentes, éclairages différents, échelles différentes ... y compris les différents aspects d'une même scène (personne portant des lunettes ou non, différents vêtements, barbue...) ou à la signification des séquences d'actions.

L'audio est plus facile de ce point de vue car la variabilité n'est généralement pas intentionnelle (à l'exception des imposteurs en reconnaissance du locuteur) et peut dès lors être prise en compte par des méthodes d'adaptation.

Le développement d'outils puissants permettra un nombre incalculable de nouvelles applications par exemple pour les journalistes ou médecins (recherche d'images, d'enregistrements video ou audio dans de grandes bases de données) mais aussi le consommateur non-professionnel (documentation audio-visuelle, TV et audio à la demande, analyse automatique du contenu de la boite vocale personnelle). Les fabricants d'appareils électriques domestiques préparent actuellement des produits nouveaux capables de stocker de très importantes quantités d'émissions TV. A l'aide d'outils d'indexation efficaces, ces données video et audio pourront être indexées (et aussi segmentées) et de la video à la demande pourra être construite à partir des données traitées. Pour être aisée pour l'utilisateur, tout le processus de reconstruction devra être automatique et se basera donc sur la qualité de l'indexation des contenus multimedia qui reste encore le maillon faible de la chaine.

L'archivage de grandes quantités d'information audio/video est presque illusoire sans outils de recherche efficaces. Inutile de rappeler le rôle crucial des moteurs de recherche sur le Web. En audio, une requête peut être simplement orale et peut être reconnue ou immédiatement comparée à l'information enregistrée.

La formulation d'une requête video est certainement plus sophistiquée car montrer une image d'une scène n'est généralement pas suffisamment caractéristique pour trouver l'information correspondante permettant de rejouer la scène, recherchée (une photo ne permet pas de retrouver toutes les apparitions d'une personne dans un film; l'identification judiciaire exige face et profil au moins!).

Le problème de l'archivage est particulièrement crucial pour les institutions nationales chargées de la conservation du patrimoine audio-visuel (cinémathèques, Institut National de l'Audiovisuel (INA) en France dont la mission est de stocker l'ensemble de la production audio-visuelle française (films, émissions , enregistrements audio,...)).

Dans de nombreux cas, l'utilisation de l'information audio

est une aide importante pour l'indexation video. L'indexation audio-visuelle peut bénéficier de l'apport mutuel des deux media par exemple au niveau de l'interprétation sémantique mais aussi au niveau plus bas de la reconnaissance de parole audio-visuelle où la lecture labiale augmente les taux de reconnaissance de la parole dans des environnements bruités.

Le problème de la fusion d'informations véhiculées par des media différents reste un problème ouvert: en effet, soit des traits d'information appartenant à différents medias sont combinés pour réduire le reconnaisseur à un simple outil de décision (classificateur) sur des données hybrides soit des probabilités d'hypothèses multiples peuvent être combinées (scores) en vue d'une décision finale.

2. La segmentation

Un problème très important est la détection de la nature d'un signal ou plus spécifiquement, comment segmenter un signal en blocs à contenu homogène selon un critère bien déterminé (même locuteur, musique, parole ou bruit...). Le développement récent de la composition vocale de l'appel pour les GSM mais aussi la limitation de bande passante qui interdit la transmission du bruit ou de l'absence de parole remplacés par du bruit généré localement à la réception ont mis l'accent sur les détecteurs d'activité vocale (VAD).

Le besoin de VAD fiable était déjà bien connu pour l'amélioration de la parole utilisant la soustraction spectrale du bruit pour laquelle le spectre de bruit est évalué dans les segments bruités avant d'être soustrait des segments de parole suivants: le rôle du VAD est critique et difficile puisque par définition de l'application, il doit être efficace à faible rapport signal/bruit.

Différentes techniques ont été proposées qui utilisent la puissance du signal et la détection du taux de passage par zéro. L'utilisation de la puissance peut être mise en oeuvre comme une prise de décision avec hysteresis [**?**]: sous un seuil donné, la décision est "non-parole" et au-dessus d'un autre seuil la décision est "parole". L'ambiguité entre les deux seuils est levée par une décision à vote majoritaire sur la nature des segments voisins.

Pour les GSM, la faible quantité de mémoire disponible empêche le stockage de longs segments requis pour les évaluations des niveaux de puissance.

Une autre solution est d'entrainer un très simple modèle de Markov caché (HMM) sur la parole et la non-parole et d'ensuite étiqueter les segments en parole/non-parole par reconnaissance. L'entrainement sur la non-parole est critique puisque les modèles de parole sont très caractéristiques (surtout les segments voisés) tandis que la variabilité de la non-parole est grande et peut dégrader le comportement du détecteur; l'adaptation du modèle de non-parole est dès lors recommandée pour la plupart des applications.

Une approche fort intéressante a été proposée récemment par Renevey [**?**] dans laquelle on utilise le fait que la cohérence des segments de parole est beaucoup plus forte que celle du bruit. En conséquence, l'entropie est utilisée comme critère pour discriminer parole et non-parole (entropie plus élevée).

Une autre application est la séparation entre parole et musique. On peut entrainer des réseaux de neurones et les utiliser ensuite pour discriminer musique et parole mais aussi différents types de musique (jazz, pop, classical music,...) [**?**].

L'indexation automatique des nouvelles radiodiffusées exige ce genre d'outils discriminants comme préprocesseur avant analyse afin d'éliminer les jingles associés aux présentations parlées. En cas de musique de fond, la séparation de sources serait nécessaire mais l'information disponible sur les enregistrements est généralement insuffisante pour résoudre ce problème.

Nous analyserons dans une section suivante le rôle de la segmentation d'un fichier de parole multilocuteurs en locuteurs et celle de l'identification tant pour l'indexation que pour l'amélioration des taux de reconnaissance [**?**].

D'une façon semblable, la reconnaissance de la langue peut être utilisée pour activer les modèles de phonèmes ou de mots dans une tache de reconnaissance multilingue et même pour activer un outil de traduction en ligne qui est presque le but ultime du traitement de la parole.

3. L'utilisation de reconnaisseurs grand vocabulaire et indépendants du locuteur. (LVSIR)

La façon la plus triviale de résoudre le problème de l'indexation audio est d'utiliser un reconnaisseur de parole puissant pour reconnaitre le texte dit et d'ensuite analyser celui-ci à l'aide d'un éditeur de texte. Cependant, cette méthode est source de nombreux problèmes puisque le reconnaisseur n'est jamais parfait surtout s'il est indépendant du locuteur et à large vocabulaire.

Une application intŕessante des LVSIR est la transcription des messages vocaux en texte écrit comme démontré aux laboratoires de AT&T. Le temps réel est loin d'être atteint mais pour ce type d'application cela n'est pas indispensable puisque par définition des délais sont toujours prévus entre l'enregistrement et la lecture. A partir du texte reconnu, on peut rechercher l'information en utilisant des éditeurs de texte ordinaires.

Un grand nombre d'expériences ont été menées sur la base de données Broadcast News Hub4 dans des laboratoires européens, japonais et américains [**?**, **?**, **?**, **?**, **?**]. Plusieurs projets européens comme THISL ont été consacrés à l'indexation par reconnaissance.

Les LVSIR ne sont pas conçus pour fournir de l'information sur l'identité des locuteurs; bien au contraire, le reconnaisseur évite l'usage de toute information sensible à l'identité. Nous discuterons dans une section suivante comment l'identification du locuteur peut servir à l'indexation.

4. Recherche de mots-clés

La recherche des mots-clés a été l'un des premiers outils spécialisés pour l'accès à l'information audio.

4.1. Modèles de rejet

Lors des premiers essais, des HMM furent construits et entrainés pour les mots-clés ainsi qu'un modèle de rejet justifiant tous les autres mots. Le taux de détection se dégrade trés rapidement avec le nombre de mots-clés qui doivent par ailleurs être tous spécifiés d'avance.

La définition d'un modèle de rejet dynamique est une amélioration intéressante. Les probabilités d'émission de ce modèle ne sont plus calculées à partir de distributions paramétriques associées aux états mais sont pour chaque vecteur d'entrée la moyenne des probabilités d'émission des états les plus probables des modèles des mots-clés à l'exclusion des deux ou trois meilleurs. Le modèle est appelé dynamique car il n'engendre pas ses propres probabilités d'émission mais les emprunte aux modèles des mots-clés. C'est la représentation d'un évènement qui sera moins bien justifié par les modèles de

mots-clés que les mots-clés eux-mêmes.

4.2. Treillis phonémiques

Afin d'être plus souple par rapport au nombre de mots-clés et contrairement à ce qui est fait lors de l'emploi de LVSIR qui reconnaissent des mots, on utilise aussi le décodage acoustico-phonétique. Tous les efforts sont portés sur la qualité de ce décodage et sur la génération d'un treillis des N-meilleures solutions [**?**]. Le treillis est engendré hors ligne pour chaque fichier. Ce treillis est stocké comme fichier adjoint du fichier audio. Lors d'une requête, un mot spécifique sera décrit par sa transcription phonétique et recherché dans le treillis en respectant des règles de continuité et de recouvrement entre hypothèses de localisation de phonèmes pour compenser les imprécisions du décodage.

L'avantage de cette méthode est qu'aucune liste de mots-clés ne doit être connue a priori et que n'importe quel mot peut être recherché (bien sur, comme dans la recherche textuelle, les mots très courts sont détectés partout!).

Cependant, puisqu'une grande variabilité est observée dans les prononciations des mots, la transcription phonétique canonique reprise dans les bons dictionnaires, n'est pas la seule entrée à utiliser. La génération de variantes de prononciation est indispensable pour accroitre le taux de rappel de mots-clés.

Différentes techniques ont été expérimentées pour la génération de treillis phonétiques incluant l'étiquetage de trames, [**?**], les HMM [**?**] et la technique REMAP [**?**, **?**, **?**] où il est fait usage de probabilités d'émission plus discriminantes engendrées par un étiqueteur phonémique HMM/ANN hybride. Des expériences ont été faites sur TIMIT mais aussi sur les résultats sportifs retransmis par CNN.

Des critères d'évaluation universels sont difficiles à définir [**?**]: en effet la nature de la base de données joue un rôle important ainsi que la fréquence du mot-clé recherché dans la base.

Les courbes ROC (Receiver Operating Curves) sont traditionnellement utilisées lorsqu'il s'agit d'évaluer le comportement d'un système en termes de fausses alarmes et de non-détections (typique pour l'identification des locuteurs). En effet un système est décrit dans le plan de ces deux types de taux d'erreur et la courbe ROC est le lieu engendré par la variation d'un seuil de décision.

Cependant, dans une application de recherche de mots-clés, les scores devraient être indépendants de la fréquence des mots-clés dans le texte et le taux de fausses alarmes est remplacé par le probabilité de fausses alarmes par mot-clé et par heure.

Pour l'application du tri par le contenu (c'est à dire si l'on essaye de trier les messages en fonction des mots-clés contenus), un autre critère peut être utilisé comme défini ci-dessous. Suuposons que le mot-clé recherché apparaisse N fois dans la base de données. Pour chaque phrase, la vraisemblance qu'elle contienne le mot-clé est calculée. Le détecteur de mots-clés trie les phrases en ordre décroissant des vraisemblances.

Dans un système idéal, les N premières phrases contiennent le mot-clé et la précision est 1. Bien sur, des erreurs sont possibles et supposons que la m-ème apparition du mot-clé arrive à la n-ème phrase ($n > m$): la précision est alors n/m pour cette apparition. La précision moyenne sur toutes les apparitions donne une bonne mesure du taux de rappel du système.

Il est indépendant de la taille de la base de données si la fréquence des mots-clés reste constante tout au long de la base de données. Ceci n'est plus vrai dans le cas de recherche de mots-clés: en effet plus la base est grande, plus faible devient la précision.

Une autre mesure d'efficacité est le gain de temps réalisé en triant la base. Il est défini comme le rapport entre le nombre de phrases que l'utilisateur ne doit pas écouter par rapport au nombre de celles qu'il aurait du écouter s'il ne disposait pas de détecteur de mots-clés.

5. Information sur le locuteur

L'identification du locuteur a fait l'objet d'un incroyable nombre de techniques différentes et la première section ne vise en aucun cas à leur analyse exhaustive et comparative: un livre entier ne suffirait pas. Seules quelques directions essentielles seront mises en évidence et liées à différents domaines applicatifs. Une seconde section introduira le problème de la segmentation en locuteurs d'un enregistrement de conversation à deux ou à plusieurs locuteurs.

5.1. Identification du locuteur

Les buts de l'indexation sont très différents de ceux du commerce électronique où l'identité du locuteur est requise pour valider une transaction commerciale et est donc une vérification du locuteur. En effet, les modèles du locuteur en commerce électronique doivent être construits d'avance à partir d'une quantité raisonnable de données collectées hors ligne. L'utilisateur proclame son identité qui est comparée à son modèle.

Des phrases promptées par le système permettent d'éviter l'utilisation de mots de passe frauduleusement enregistrés. Dans ce cas, les modèles de phonèmes sont entrainés hors ligne et le volume de données d'entrainement est très élevé.

Lors d'une tache d'indexation, on n'utilisera que le fichier à indexer mais d'autre part, le problème des imposteurs n'existe plus.

Plusieurs solutions pour l'identification ont été proposées:

- la quantification vectorielle (VQ) où chaque locuteur est représenté par son " codebook"; l'identification est alors basée sur l'erreur de quantification cumulée [**?**],
- une généralisation de la VQ est le modèle ergodique où des probabilités de transition controlent le saut d'un centroide à l'autre,
- l'utilisation de HMM spécifiques entrainés sur des données de l'utilisateur [**?**],
- des réseaux de neurones entrainés pour chaque locuteur et qui requièrent un énorme volume de données pour l'enregistrement du locuteur (enrolment) [**?**],
- la comparaison de matrices de covariance des données enregistrées et des données de test avec différentes mesures de similarité [**?**]
- la comparaison avec les centroides dans l'espace vectoriel des locuteurs engendré hors ligne (eigenvoices) [**?**, **?**, **?**].
- l'utilisation du logarithme du rapport de vraisemblance généralisé avec un modèle du monde où chaque modèle de locuteur et du monde sont des mélanges de Gaussiennes (GMM). Les données d'entrainement sont enregistrées pendant une phase d'enregistrement (enrolment) [**?**].

Toutes ces applications requièrent un enregistrement (enrolment). La qualité de l'identification du locuteur dépend du volume des données et est estimée par les ROC's.

5.2. La segmentation en locuteurs

Le but de la segmentation est de diviser des sessions audio enregistrées en segments de sorte que chacun d'eux ne contienne de la parole que d'un seul locuteur. Dans une opération suivante, les différents segments sont regroupés pour former une base de données pour chaque locuteur. A partir de ces données homogènes (pour autant aque la segmentation ait été précise), des modèles de locuteur (GMM) peuvent être entrainés. Faisant usage ensuite de ces modèles, la session peut être segmentée exactement comme des phrases peuvent être segmentées en mots en reconnaissance de parole continue en utilisant un algorithme de Viterbi.

Les premiers travaux sont dus à Gish [**?**] qui développa un analyseur automatique des dialogues entre pilotes d'avion et controleurs aériens. Bien sur, les différents rapports signal/bruit rendaient la qualité des transmissions très asymétrique si bien que des segments de longueur égale pouvaient facilement être étiquetés comme parole du pilote ou du controleur avec la résolution de la longueur d'un segment.

Il devint rapidement clair que pour des applications plus générales, d'autres techniques sont nécessaires et une technique trés répandue est l'utilisation de de fenêtres divisées glissantes le long du fichier audio. Plus spécifiquement, des modèles gaussiens du signal enregistré sont estimés dans des fenêtres contigües ainsi que sur l'union de deux fenêtres. Le produit des vraisemblances des deux fenêtres est comparé à la vraisemblance de leur union en utilisant un critère BIC (Bayesian Indexing Criterion) qui pénalise la représentation ayant globalement plus de degrés de liberté [**?**, **?**, **?**, **?**, **?**, **?**, **?**, **?**]. Un critère de vraisemblance généralisé est tabulé et ses maxima sont détectés sous des contraintes évitant des artefacts dus aux erreurs numériques et à la faible représentation locale du signal. Les maxima correspondent aux localisations de possibles changements de locuteurs. Dans une seconde passe, les points de segmentation sont confirmés [**?**].

Dans l'étape suivante, les segments trouvés sont regroupés si on peut les considérer comme appartenant au même locuteur. Puisque le nombre de locuteurs intervenant dans l'enregistrement est a priori inconnu, une recherche est faite dans un arbre de groupement et un seuil dépendant du critère BIC à nouveau décide quand unir des segments (la réunion des feuilles de l'arbre conduit à un seul locuteur) [**?**].

Lorsque le regroupement est terminé, la base segmentée peut être considérée comme une nouvelle base contenant un volume suffisant de données par locuteur permettant d'entrainer des modèles (par exemple des GMM). A l'aide des modèles entrainés et d'un algorithme de Viterbi une nouvelle segmentation en locuteurs peut être obtenue et itérativement, des modèles plus précis peuvent être construits tout en améliorant la segmentation. Cette méthode est semblable à l'entrainement Viterbi in situ qui fournit une segmentation comme produit dérivé.

6. Conclusions

On a montré que l'analyse automatique du contenu des fichiers audio peut engendrer de nombreuses applications. Différentes technologies convergent pour améliorer la recherche d'information selon le contenu depuis des outils de base comme les détecteurs d'activité vocale jusqu'aux reconnaisseurs de grands vocabulaires et indépendants du locuteur capables de transcrire un fichier entier. Cet article a essayé de faire la liste des techniques qui peuvent contribuer à la création d'un moteur de recherche audio et de décrire les technologies sous-jacentes. Notre volonté n'a pas été d'être exhaustif et cette contribution est certainement biaisée en décrivant l'activité du groupe Parole de l'Institut Eurécom. La liste de références est loin d'être le reflet des efforts de tous les laboratoires multimedia impliqués dans ce domaine crucial pour le développement de la société de l'information.

7. References

[1] Ph. Gelin, C.J. Wellekens, Keyword Spotting for Video Soundtrack Indexing, *IEEE Conf. Acoustics, Speech and Signal Processing, ICASSP-1996* , Atlanta (USA).

[2] Ph. Gelin, C.J. Wellekens, Keyword Spotting Enhancement for Video Soundtrack Indexing,*ICSLP 96*, Philadelphia, October 96.

[3] Ph. Gelin, C.J. Wellekens, REMAP for video soundtrack indexing, *ICASSP 97*, Munchen, 1997.

[4] Ph. Gelin, C.J. Wellekens, Keyword Spotting for Multimedia Document Indexing, *SPIE 97*, Dallas, Nov 97.

[5] P. Delacourt, D. Kryze, C.J. Wellekens, Speaker-Based Segmentation for Audio Data Indexing, *ESCA-ETRW Workshop: Accessing Information in Spoken Audio*, Cambridge (UK), April 1999.

[6] P. Delacourt, C.J. Wellekens, Audio Data Indexing: Use of Second Order Statistics for Speaker Based Segmentation, *Proc. ICMCS 99*, Florence, June 1999.

[7] P. Nguyen, R. Kuhn, J-C. Junqua, N. Niedzielski, C.J. Wellekens, Voix propres: Une représentation compacte de locuteurs dans l'espace des modèles, *CORESA 99*, Sophia Antipolis, France.

[8] P. Delacourt, C.J. Wellekens, Segmentation en locuteurs d'un document audio, *CORESA 99*, Sophia Antipolis, France.

[9] P. Nguyen, C.J. Wellekens, J.-C. Junqua, Maximum Likelihood Eigenspace and MLLR for Speech Recognition in Noisy Environmemt, *Eurospeech 1999*, Budapest, Hungary.

[10] P. Delacourt, C.J. Wellekens, "A first step into speaker-based indexing", *1st European Workshop on Content-based Multimedia Indexing (CBMI'99)*, Toulouse, France, October 25-27, 1999.

[11] P. Delacourt, J.F. Bonastre, C. Fredouille, S. Meignier, T. Merlin, C.J. Wellekens, "Différentes Stratégies pour le Suivi du Locuteur" , *RFIA2000 : Reconnaissance des Formes et Intelligence Artificielle*, Paris, 01-03 Février 2000.

[12] P. Delacourt, J.F. Bonastre, C. Fredouille, T. Merlin, C.J. Wellekens, "A Speaker Tracking System Based on Speaker Turn Detection for Nist Evaluation", *ICASSP-2000*, Istanbul, Turkey, 05-09 juin 2000.

[13] P. Delacourt, C.J Wellekens, "DISTBIC: a speaker-based segmentation for audio data indexing", Speech Communication, vol. 32, 2000.

[14] P. Delacourt, C.J. Wellekens, "Regroupement par le locuteur de messages vocaux", *Coresa 2000*, Poitiers, 19-20 octobre 2000.

[15] P. Nguyen, R. Kuhn, J.-C. Junqua, N. Niedzielski, C.J. Wellekens, "A compact representation of speakers in the model space", *Annales de Télécommunications*, nov. 2000.

[16] H. Gish, " Robust discrimination in automatic speaker identification", *ICASSP 1990*, pp. 289-292, 1990.

[17] P. Renevey, Doctoral thesis, EPFL, 2001.

[18] H. Bourlard,Y. Konig, N. Morgan, "REMAP: Recursive Estimation and Maximization of A Posteriori Probabilities. Application to Transition-Based Connectionist Speech Recognition", Internal report of ICSI, TR-94-064, August 1995.

[19] F. Bimbot, Y. Magrin-Chagnolleau, L. Mathan, "Second order statistical measures for text independent speaker identification", Speech Communication, vol. 17, nos 1-2, pp.177-192, Aug. 1995.

[20] S. Chen, J.F. Gales, P. Gopalakrishnan, R. Gopinath, H. Printz, D. Kanevsky, P. Olsen, L. Polymenakos, "IBM's LVCSR system for transcription of broadcast news used in the 1997 HUB4 English evaluation", *DARPA Speech Recognition Workshop*, 1998.

[21] F. Kubala, H. Jin, L. Nguyen, R. Schwartz, J. Makhoul, "The 1996 BBN Byblos Hub-4 transcription system", *DARPA Speech Recognition Workshop*, 1997.

[22] P. Woodland, J.F. Gales, D. Pye, S. Young, "The development of the 1996 HTK Broadcast News transcription system", *DARPA Speech Recognition Workshop*, 1997.

[23] M. Harris, X. Aubert, R. Haeb-Umbach, P. Bayerlein, "A study of broadcast news audio stream segmentation and segment clustering", *Eurospeech 1999*.

[24] P. Delacourt, "Indexation de données audio: segmentation et regroupement par locuteurs", Thèse doctorale, Institut Eurécom, 2000.

[25] P. Gelin, "Détection de mots clés dans un flux de parole: application à l'indexation de documents multimedia", Thèse doctorale, Institut Eurécom, 1997.

[26] C. Montacié, M.-J. Caraty, "Sound Channel Video Indexing", *Eurospeech 1997*, pp. 2359-2362, 1997.

[27] T. Matsui, S. Furui, "Comparison of text independent speaker recognition method using VQ-distortion and discrete/continuous HMMs", *ICASSP*, 1992

[28] D.A. Reynolds, "Speaker identification and verification using Gaussian mixtures models", Speech Communication, vol. 17, pp. 91-108, 1995.

[29] J-L. Gauvain, L. Lamel, G. Adda, M. Jardino, "The LIMSI 1998 HUB4-E transcrption system", *DARPA Broadcast News Workshop*, 1999.

[30] Y. Bennani, P. Gallinari, "Connectionist methods for speaker verification (Tutorial)", *ESCA Workshop on Speaker Recognition, Identification and Verification*, Martigny, 1994.