Projet RAIVES (Recherche Automatique d'Informations Verbales Et Sonores) vers l'extraction et la structuration de données radiophoniques sur Internet

Nathalie Parlangeau-Vallès, Ivan Magrin-Chagnolleau, Dominique Fohr, Irina Illina, Odile Mella, Kamel Smaïli, Christine Sénac, Jérôme Farinas, Julien Pinquier, Jean-Luc Rouas, et al.

Projet RAIVES

Recherche **A**utomatique d'**I**nformations **V**erbales **E**t **S**onores :
vers l'extraction et la structuration de données radiophoniques sur Internet

Rapport d'avancement de la première année

1. Introduction

1.1. Coordinateurs

Nathalie Vallès-Parlangeau et Ivan Magrin-Chagnolleau

1.2. Partenaires du projet

STIC : **IRIT**, Université Paul Sabatier Toulouse III
118, route de Narbonne
31062 Toulouse Cedex 4

Christine Sénac, Maître de Conférences, Université Toulouse III
Christine.Dours-Senac@irit.fr

Régine André-Obrecht, Professeur, Université Toulouse III
Regine.Obrecht@irit.fr

Jérôme Farinas, ATER, Université Toulouse III
Jerome.Farinas@irit.fr

Julien Pinquier, Doctorant, Université Toulouse III
Julien.Pinquier@irit.fr

Jean-Luc Rouas, Doctorant, Université Toulouse III
Jean-Luc.Rouas@irit.fr

LORIA, INRIA-Lorraine
Campus Scientifique
BP 239
54506 Vandœuvre-les-Nancy

Nathalie Vallès-Parlangeau, Maître de Conférences, Université Nancy II
Nathalie.Valles-Parlangeau@loria.fr

Dominique Fohr, Chargé de Recherche CNRS
Dominique.Fohr@loria.fr

Irina Illina, Maître de conférences, Université de Nancy II
Irina.Illina@loria.fr

Odile Mella, Maître de conférences en délégation, Université de Nancy I
Odile.Mella@loria.fr

Kamel Smaïli, Maître de Conférences, Université Nancy II
Kamel.Smaili@loria.fr

SHS : **Laboratoire Dynamique Du Langage**, CNRS UMR 5596 - Université de Lyon 2
Institut des Sciences de l'Homme
14, avenue Berthelot
69363 Lyon Cedex 07

Ivan Magrin-Chagnolleau, Chargé de Recherche CNRS
Ivan.Magrin-Chagnolleau@univ-lyon2.fr

David Janiszek, Post-Doc.
David.Janiszek@univ-lyon2.fr

François Pellegrino, Chargé de Recherche CNRS
Francois.Pellegrino@univ-lyon2.fr

1.3. Rappel des objectifs du projet

Les documents sonores font à l'heure actuelle partie de ce que l'on appelle le « Web invisible ». Le projet a pour objectif de structurer ces documents sonores, en particulier radiophoniques, à partir de l'indexation par leur contenu, de manière à leur donner un sens du point de vue d'un utilisateur du Web, et de produire à partir de ces documents des connaissances exploitables. Ce contenu pourrait alors être accessible aux moteurs de recherche et devenir disponible aux internautes au même titre que le contenu textuel de pages HTML. Ce projet pourrait donc contribuer au développement d'une nouvelle génération de moteurs de recherche, capables d'accéder aussi à des documents sonores, et pourquoi pas visuels, par leur contenu. Les méthodes seront mises au point sur des données radiophoniques classiques avant d'être adaptées à des données radiophoniques provenant d'Internet.

1.4. Mots Clés

Indexation automatique, structuration, documents radiophoniques, information invisible sur le Web, modélisation statistique, information linguistique.

1.5. Description du projet

Internet est devenu un vecteur important de la communication. Il permet la diffusion et l'échange d'un volume croissant de données. Il ne s'agit donc plus seulement de collecter des masses importantes « d'informations électroniques », mais surtout de les répertorier, de les classer pour faciliter l'accès à l'information utile. Une information, aussi importante soit-elle, sur un site non répertorié, est méconnue. Il ne faut donc pas négliger la part du « Web invisible ». Le Web invisible peut se définir comme l'ensemble des informations non indexées, soit parce qu'elles ne sont pas répertoriées, soit parce que les pages les contenant sont dynamiques, soit encore parce que leur nature n'est pas ou difficilement indexable. En effet, la plupart des moteurs de recherche se basent sur une analyse textuelle du contenu des pages, mais ne peuvent prendre en compte le contenu des documents sonores ou visuels. Il faut donc fournir un ensemble d'éléments descripteurs du contenu pour structurer les documents afin que l'information soit accessible aux moteurs de recherche.

S'agissant de documents sonores, le but de notre projet est donc, d'une part, d'extraire ces informations et, d'autre part, de fournir une structuration des documents afin de faciliter l'accès au contenu.

L'indexation par le contenu de documents sonores s'appuie sur des techniques utilisées en traitement automatique de la parole, mais doit être distinguée de l'alignement automatique d'un texte sur un flux sonore ou encore de la reconnaissance automatique de la parole. Ce serait alors réduire le contenu d'un document sonore à sa seule composante verbale. Or, la composante non-verbale d'un document sonore est importante et correspond souvent à une structuration particulière du document. Par exemple, dans le cas de documents radiophoniques, on voit l'alternance de parole et de musique, plus particulièrement de jingles, pour annoncer les informations. Ainsi, nous pouvons considérer un ensemble de descripteurs du contenu d'un document radiophonique : segments de Parole/Musique, « sons clés », langue, changements de locuteurs associés à une éventuelle identification de ces locuteurs, mots clés et thèmes. Cet ensemble peut être bien entendu enrichi. Extraire l'ensemble des descripteurs est sans doute suffisant pour référencer un document sur Internet. Mais il est intéressant d'aller plus loin et de donner accès à des parties précises du document. Chaque descripteur doit être associé à un marqueur temporel qui donne accès directement à l'information. Cependant, l'ensemble des descripteurs appartenant à des niveaux de description différents, leur organisation n'est pas linéaire dans le temps : un même locuteur peut parler en deux langues sur un même segment de parole, ou encore sur un segment de parole dans une langue donnée, plusieurs locuteurs peuvent intervenir. Il faut donc aussi être capable de fournir une structuration de l'information sur différents niveaux de représentation.

Le schéma suivant (Figure 1) résume l'ensemble des descripteurs qui seront recherchés sur la bande sonore dans le cadre de ce projet.

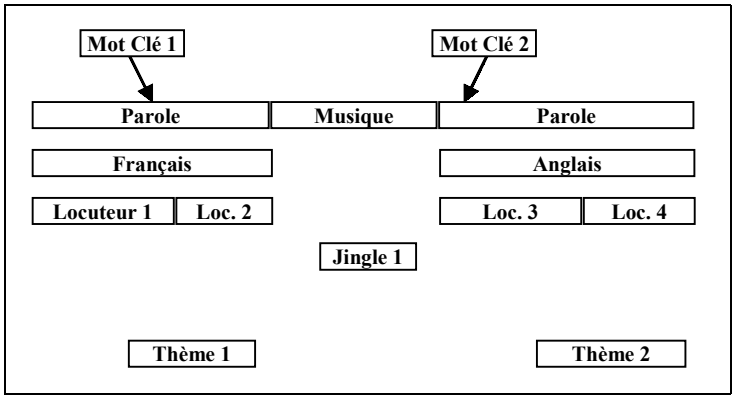


Figure 1: Quelques descripteurs sonores pouvant être extraits d'un document sonore.

1.6. Description des tâches

- **T1 : Segmentation Parole/Musique**

Objectifs :

Le but est de déterminer les segments Parole et Musique de l'enregistrement. Ces informations constituent à la fois des descripteurs et sont des points d'ancrage pour la recherche d'autres descripteurs.

Description des travaux :

Un document radiophonique est le résultat d'un mixage de différentes sources sonores (parole, musique, bruits). La combinaison de ces sources entre elles est une réelle difficulté. Partant du constat que parole et musique ne partagent pas le même espace de représentation (formantique et harmonique), l'idée repose sur une modélisation statistique différenciée afin d'effectuer une séparation de type Classe/Non Classe (Parole/Non Parole et Musique/Non Musique). Cette approche est déjà mise en œuvre et ce projet permettra de la valider sur une application à large spectre de contenu.

- **T2 : Détection de « sons clés »**

Objectifs :

Le but est d'identifier des « sons clés », et plus particulièrement ici les jingles, qui sont des éléments structurants des documents radiophoniques.

Description des travaux :

La détection de « sons clés », ou plus particulièrement ici de jingles, se fait dans les mêmes conditions difficiles que la reconnaissance de mots clés (ambiance bruitée,…). Deux approches seront explorées. La première consiste à mesurer la similitude entre une ou plusieurs références de base du « son clé » et le continuum sonore. La seconde approche est basée sur une modélisation statistique : un modèle pour le « son clé » et un « modèle du Monde ». La recherche se fait par maximum de vraisemblance.

- **T3 : Identification de la langue**

Objectifs :

Le but est de déterminer la ou les langues présentes dans le document. Cette phase permettra de sélectionner les outils et les modèles appropriés pour les tâches de détection de mots clés et de thèmes.

Description des travaux :

L'identification d'une langue se fait à partir d'une modélisation statistique de celle-ci, reposant sur les propriétés phonotactiques, rythmiques, mélodiques, phonétiques et acoustiques de cette langue. La grande diversité des langues étudiées dans le laboratoire DDL permet d'avoir une bonne connaissance des indices caractéristiques de chaque langue. Pour une langue ne faisant pas partie de l'ensemble de référence de départ, le système retournera la réponse « langue inconnue », et pourra éventuellement apprendre un nouveau modèle, en se basant sur les indices propres à cette nouvelle langue.

- **T4 : Information sur les locuteurs**

Objectifs :

Le but est double : segmenter la bande sonore en locuteurs et identifier certains locuteurs quand ceux-ci sont connus de la base de données. Une seconde étape pourra être l'appariement de locuteurs entre des documents sonores différents préalablement segmentés en locuteurs.

Description des travaux :

La segmentation en locuteurs permet de déterminer tous les changements de locuteurs dans les plages de parole, et d'apparier tous les segments appartenant à un même locuteur. Il est possible aussi d'identifier certains locuteurs dans la mesure où ils sont déjà connus de la base, et si on possède pour ces locuteurs un modèle statistique préalablement appris. Cette tâche repose sur une modélisation statistique de l'espace acoustique de chaque locuteur, mais pourra aussi intégrer des indices linguistiques qui ne sont habituellement pas pris en compte dans ce type de tâches, comme le rythme, la mélodie, la phonotactique ou encore l'espace lexical propres à chaque locuteur.

- **T5 : Détection de mots-clés**

Objectifs :

Le but est de détecter des mots clés sans s'appuyer sur une transcription complète des segments de parole.

Description des travaux :

La mise en oeuvre d'un systèmes de reconnaissnce grand vocabullaire est classiquement envisagée pour une tâhce de détection de mots clés. Cependant, pour des impératifs de temps de claculs liés à l'application visée, l'approche choisie est celle reposant sur une reconnaissance flexible à base de modèles de Markov cachés où le modèle du mot clé est construit à partir de l'ensemble des unités phonétiques qui composent le « Modèle du Monde ». L'évaluation d'un tel système se fait en fonction du nombre d'omissions et d'insertions de mots clés.

- **T6 : Extraction de thèmes**

Objectifs :

Le but est de déterminer quel est le thème d'un segment de parole à partir des mots clés précédemment détectés.

Description des travaux :

Pour plusieurs thèmes préalablement choisis, des mots caractéristiques sont automatiquement déterminés à partir de corpus d'apprentissage thématiques. Puis, pour chaque thème, les mots caractéristiques sont comparés aux mots clés détectés au cours de la tâche précédente. A l'issue de cette comparaison, un thème principal est déterminé. Plusieurs thèmes secondaires peuvent également être proposés. Le choix des mots clés à rechercher et le choix des thèmes possibles se fait bien sûr de manière dépendante.

2. La base de données RAIVES

2.1. Introduction

Un des buts du projet RAIVES est de contribuer à la recherche d'informations et à l'indexation de données audio circulant sur le web. Dans un premier temps, nous avons choisi de travailler sur des données radiophoniques non compressées. Pour cela, nous avons contacté RFI (Radio France Internationale) afin d'obtenir des données de bonne qualité et qui présente une variété d'émissions importantes dans 19 langues. Le travail sur les données compressées sera étudié ultérieurement dans le projet.

2.2. Contenu de la base de données

RFI est une radio française implantée en France et dans 18 autres pays, ce qui nous permet de collecter des programmes dans 19 langues. La base de données finale sera composée de 10 heures de programmes pour chaque langue soit un total de 190 heures. Les 19 langues sont : Allemand, Anglais, Arabe, Bulgare, Cambodgien, Chinois, Créole, Espagnol, Français, Laotien, Persan, Polonais, Portugais (Afrique), Portugais (Brésil), Roumain, Russe, Tchèque, Turc, Vietnamien. Nous avons défini la composition des programmes de façon à avoir pour chaque langue environ 3h00 d'interviews, 3h00 de programmes musicaux et 3h00 de bulletins d'information.

Afin de commencer à travailler en attendant les réceptions des divers programmes, RFI nous a fourni 30h00 de programmes divers en 4 langues (7h00 par langue). On y trouve :

- *pour le français* : des émissions comportant des interviews de chanteurs ou de groupes de musique, des extraits musicaux, des émissions de lecture de définitions extraites de dictionnaires et des bulletins d'information.
- *pour l'anglais, l'espagnol et le portugais* : des reportages comportant des éléments musicaux et des cours de français.

Nous avons pour l'instant réceptionné les enregistrements commandés pour le Français, l'Anglais et l'Espagnol.

2.3. Etiquetage

L'étiquetage manuel d'une partie de la base de données est une étape obligatoire pour la mise au point et la validation de nos algorithmes. Nous avons défini une sous-base qui satisfait au mieux toutes les demandes des divers partenaires. S'agissant d'un travail extrêmement long, cette sous base est pour l'instant relativement restreinte. Le *Tableau 1* nous donne un aperçu du contenu de cette sous base : les programmes choisis sont très différents en terme de locuteurs, contenus musicaux, style oral et surtout en terme de conditions d'enregistrement et de bruit de fond. Elle représente au total environ 3h30 d'émissions.

Pour les besoins de l'expérimentation, nous avons utilisé une partie de la base qui nous pour l'apprentissage (programmes 1 à 7) des divers modèles et l'autre partiet pour tester nos méthodes (programme 8).

	Durée totale	Nombre de locuteurs	Durée de la musique	Durée de la parole
Programme 1	23'03''	6	23'00''	9'08''
Programme 2	24'18''	7	3'35''	19'18''
Programme 3	23'45''	6	23'00''	9'00''
Programme 4	24'00''	5	3'00''	21'00''
Programme 5	23'33''	8	21'37''	11'17''
Programme 6	24'00''	6	8'05''	21'15''
Programme 7	59'59''	40	1'54''	58'10''
Programme 8	23'50''	11	2'50''	21'00''

Tableau 1: Description de la sous base d'expérimentation.

La sous-base a été manuellement étiquetée en terme de :
- parole, musique, musique voix chantée,
- langue (anglais et français pour l'instant)
- locuteurs,
- transcription orthographique.

Le logiciel utilisé pour l'annotation manuelle est Transcriber[1]. Ce logiciel permet d'annoter le signal sur trois niveaux de description qui peuvent être définis suivant l'application. Il permet de gérer des annotations fines comme des événements qui peuvent apparaître (applaudissements, rires, respirations,...), de transcrire des dialogues qui se superposent, de définir de façon précise les locuteurs,...

Afin d'harmoniser les annotations, nous avons défini un guide d'étiquetage (Annexe 1).

2.4. Perspectives

Toutes les langues de notre base de données n'ont pas été fournies par RFI. Il faut donc encore attendre des données qui doivent arriver durant la seconde année de ce projet.

L'annotation manuelle du corpus doit être étendue à au moins trois émissions supplémentaires. Nous envisageons à plus long terme un étiquetage semi-automatique basé sur les outils que nous développons.

[1] http://www.etca.fr/CTA/gip/Projets/Transcriber/

3. Segmentation Parole/Musique

3.1. Introduction

Les segments acoustiques sont de nature très diverse de par leur production et leur enregistrement : l'environnement peut être propre ou plus ou moins bruité, la qualité de l'enregistrement peut être plus ou moins soignée et liée à des éléments extérieurs (le canal téléphonique réduit la bande passante), la musique peut être traditionnelle ou synthétique, la présence de parole peut être observée en monologue ou en dialogue, les situations sont plus ou moins stressantes. De par cette multitude de facteurs, les segments acoustiques sont extrêmement variés et les transitions peuvent être rapides ou graduelles quel que soit le type de production ou d'enregistrement.

Pour tenir compte de cette extrême variabilité, si aucune connaissance forte n'est donnée *a priori*, le signal acoustique doit subir un certain nombre de pré-traitements avant de pouvoir espérer extraire une quelconque information pertinente. Ces pré-traitements peuvent être assimilés à la recherche de niveaux de description plus ou moins élémentaires : un niveau élémentaire peut être la simple décomposition de la bande sonore en ses composantes de base que sont la parole, la musique et les bruits sémantiquement significatifs. A un niveau moins élémentaire, il s'agira de rechercher des mots dans la parole ou d'identifier des séquences de notes en musique.

Nous nous sommes donc en premier lieu intéressés à la distinction Parole/Musique qui nous semble indispensable en amont de toute application d'indexation sonore. Elle permet en effet non seulement de fournir une indexation par son contenu parole ou musique, mais aussi de circonscrire une zone de recherche plus précise pour toutes les informations précédemment citées.

Parmi les méthodes de discrimination Parole/Musique classiquement trouvées dans la littérature, nombre de chercheurs se sont intéressés aux différences acoustiques qui peuvent exister entre ces deux types de sons. Les méthodes de classification, quant à elles, restent plus classiques (Modèles de Mélanges de lois Gaussiennes, k plus proches voisins,…).

De part la nature même des signaux de musique et de parole, leur indexation ne peut résulter de l'utilisation d'outils communs. Nous proposons donc ici une indexation Parole/Musique basée sur une modélisation différenciée pour chacune des composantes parole et musique. L'approche est mise en œuvre à partir de Modèles de Mélanges de lois Gaussiennes.

3.2. Méthodes

3.2.1. Le modèle théorique

S'agissant d'un contexte d'indexation, le but est de trouver les composantes parole et musique du document sonore de façon indépendante et donc non concurrente. Il n'est donc pas question de chercher à discriminer la parole de la musique, mais à les caractériser au mieux de façon indépendante afin de faire une séparation de type Classe/NonClasse (c'est-à-dire Parole/NonParole (notée P/NP) et Musique/NonMusique (notée M/NM)).

Classiquement, en discrimination, les classes que l'on cherche à séparer partagent à la fois le même espace de représentation et la même modélisation. Dans la situation présente, les différences de production qui peuvent exister entre parole et musique se retrouvent tout naturellement dans la nature des signaux eux-mêmes : la parole se caractérise par une structure formantique, tandis que la musique peut être produite de multiples manières et la définition de la musique est beaucoup plus difficile à donner. C'est pourquoi, de nombreux chercheurs se limitent (lorsqu'il s'agit d'extraire cette composante) à de la musique « instrumentale traditionnelle » dans le sens où elle est une composition de sons harmoniques (de notes au sens classique).

Le but n'est plus uniquement de trouver des paramètres qui permettent de séparer au mieux ces classes, mais aussi de trouver des ensembles de représentations qui caractérisent au mieux chacune des classes. De même, il peut être nécessaire de mettre en œuvre des modélisations distinctes pour chacune des classes. La modélisation différenciée est donc tout à fait adaptée à notre problème ; elle est basée sur le principe que les différentes classes peuvent être modélisées séparément afin de prendre en compte leurs spécificités. Ainsi, chacune des classes est définie par son espace de représentation et son ensemble de modèles.

1 classe = {Espace de représentation, Modèle Classe, Modèle Non Classe}

3.2.2. Description du système d'indexation

L'indexation de documents sonores par le contenu est un problème de reconnaissance des formes. Quel que soit le niveau de description étudié, il s'ensuit que la méthode d'extraction se décompose en deux étapes principales :

une étape de paramétrisation (appelée ici prétraitement acoustique) où l'on recherche les indicateurs discriminants ou les caractéristiques les plus pertinentes précède l'étape de décision (appelée ici reconnaissance) (Figure 2). La décomposition en Parole et /ou Musique se faisant de façon disjointe et sur le modèle Classe/Non Classe, deux systèmes totalement distincts sont nécessaires pour traiter chacune des deux classes. Ils sont suivis d'un éventuel module de fusion.

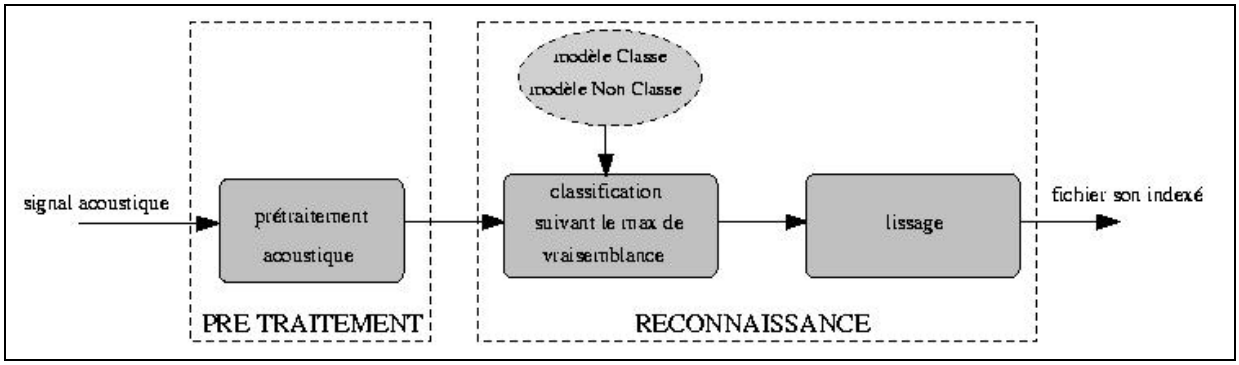


Figure 2 : Schéma général du système d'indexation

- **Prétraitement acoustique**

Compte tenu de la différence des espaces de représentation, le prétraitement acoustique doit être différent suivant que l'on recherche la Parole ou la Musique : il s'agit de définir l'espace courant des paramètres associés à chacune des classes. Pour la parole, le prétraitement consiste en une analyse cepstrale selon une échelle Mel, effectuée sur des trames de 10ms. Au total 18 paramètres sont utilisés : 8 MFCC, l'énergie et les dérivées associées. Une soustraction cepstrale permet de réduire l'effet « canal ». Pour la Musique, une simple analyse spectrale est effectuée sur ces mêmes trames. Ainsi, les paramètres extraits sont les sorties de filtres et l'énergie (soit 29 paramètres), la répartition des filtres étant linéaire par morceaux.

- **Classification**

A l'issue de la phase de prétraitement acoustique, les vecteurs de paramètres sont confrontés à chacun des modèles Classe et NonClasse (Figure 2). Nous avons choisi de modéliser la Classe et la NonClasse par un modèle de mélange de lois gaussiennes. La classification par mélange de lois gaussiennes se fait par calcul du maximum de vraisemblance pour chacun des modèles Classe et NonClasse. A la suite de cette phase de classification, une phase d'assemblage permet de concaténer sous forme de segments les trames adjacentes ayant obtenu le même index lors de la classification. Une fonction de lissage est nécessaire afin de supprimer les segments de taille négligeable.

- **Lissage**

Le but du lissage est d'éliminer les segments non significatifs issus de l'assemblage. Il s'effectue en deux phases successives :

- un pré-lissage permet d'éliminer les segments de longueur inférieure à 20ms donc non significatifs autant pour la parole que pour la musique.
- le second lissage est lié à la tâche d'indexation propre à notre application qui consiste à repérer les zones importantes (en taille) de parole (resp. de musique). Il s'agit donc ici de concaténer des segments adjacents issus du pré-lissage de façon à obtenir un seul segment indexé 'Parole' (resp. 'Musique'). Le lissage paramétrable est donc assez grossier : de l'ordre de 400ms pour la parole et de 2000ms pour la musique.

- **Apprentissage**

La modélisation est basée sur un mélange de lois Gaussiennes (GMM). Comme le montre la Figure 3 l'apprentissage s'effectue en deux temps : initialisation et optimisation. L'initialisation des modèles est obtenue par Quantification Vectorielle basée sur l'algorithme de Lloyd. L'étape d'optimisation des paramètres est réalisée par l'algorithme classique Expectation-Maximization (EM). Après expérimentation, le nombre de lois gaussiennes dans le mélange a été fixé à 128 pour tous les modèles.

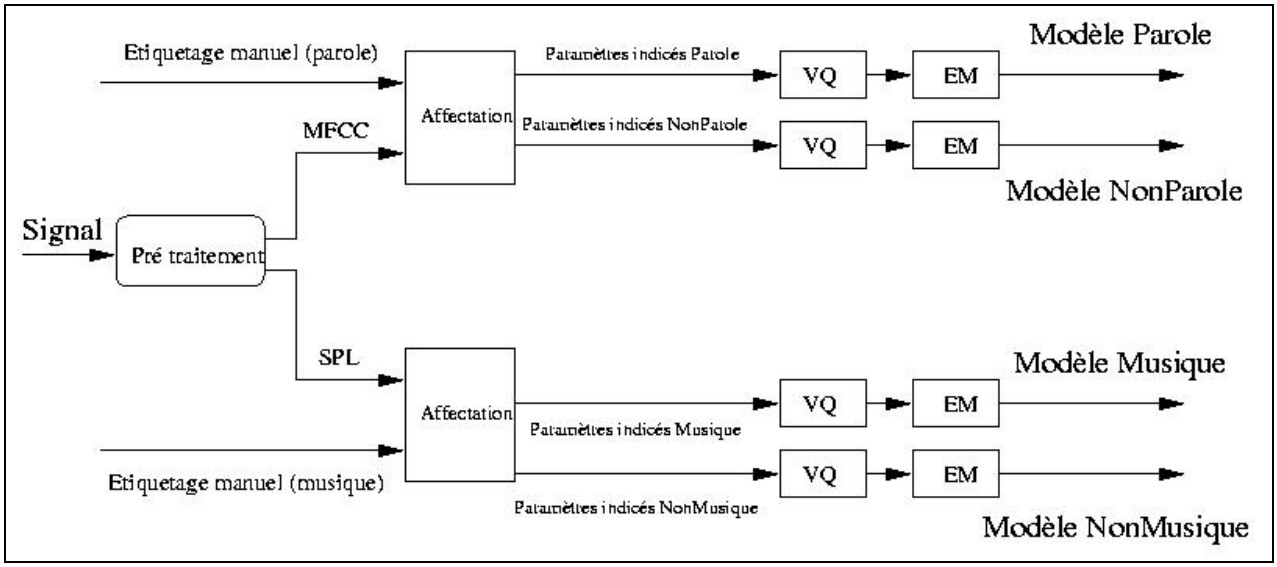


Figure 3 : L'apprentissage des GMM

3.3. Evaluation

Les programmes 1 à 6 ont été utilisés pour l'apprentissage, fournissant une durée totale de 2h15 de musique et de 1h30 de parole. Les tests ont été effectués sur le programme 8 dans un premier temps. Ce programme contient des interviews enregistrées en environnement très bruité et contient principalement de la parole (hommes et femmes) superposée avec différents types de musique. Ensuite un autre programme (programme 9) a été testé : il s'agit d'un programme en langue espagnole (1h36 dont 1h00 de parole) contenant également des interviews et de la musique. Il est à noter qu' il est beaucoup moins bruité que le programme 8.

Les résultats obtenus pour les classifications P/NP et M/NM sont reportés dans le Tableau 2 pour les 2 programmes testés. L'évaluation de la classification automatique se fait en comparant l'étiquetage manuel et l'indexation automatique. Pour cela la précision de la classification[2] est calculée selon la formule :

Précision de la classification = (longueur $_{\text{corpus de test}}$ – longueurs $_{\text{insertions}}$ – longueurs $_{\text{omissions}}$ – longueurs $_{\text{substitutions}}$) / (longueur $_{\text{corpus de test}}$)

Ensemble de test	Parole	Musique
Programme 8	84.4%	77.2%
Programme 9	94%	91%

Tableau 2 : Taux d'indexation correcte en terme de précision de la classification.

A l'examen des résultats, on observe différents types d'erreurs :

Pour la classification P/N on observe respectivement 15.6% et 6% d'erreurs sur les 2 programmes.
Sur l'ensemble des erreurs, 2% se produisent pour des segments de musique produite par des percussions qui sont classés comme des segments de parole et les 98% d'erreurs restantes concernent des segments de parole classés comme des segments de non parole.
Ces dernières erreurs sont :

a) pour 55.3%, de la parole se superpose à de la musique et du bruit,
b) pour 28%, de la parole pure,
c) pour 15.7%, du *rap* où la parole n'est pas détectée car nous avons considéré que l'étiquetage du *rap* pouvait s'assimiler à de la parole. Cet étiquetage est très certainement à revoir.
d) pour 1%, de la parole très peu audible.

Pour la classification M/NM, on observe respectivement 22.8% et 9% d'erreurs sur les 2 programmes.
Sur l'ensemble des erreurs, 1% se produisent pour des segments de musique classés comme des segments de non musique et les 99% d'erreurs restantes concernent des segments de non musique classés comme des segments de musique.
Ces dernières erreurs sont :

e) pour 68% de la parole très bruitée,

[2] Loi Toubon : Accuracy est traduit par précision de la classification

f) pour 32%, des segments de parole superposés avec de la musique très peu audible : il s'agit d'erreurs dues à un mauvais étiiquetage manuel.

En fait, la plupart de ces erreurs sont dues à une modélisation trop simpliste. En particulier, les segments de chant *a capella* ont été utilisés pour apprendre la classe NM. Ce choix permet d'expliquer de nombreuses erreurs du type e) : dans ce cas là les deux modèles M et NM ne sont pas assez 'disjoints'. Une conclusion similaire est possible pour les classes P et NP. Les segments de *rap* ont été utilisés pour l'apprentissage de la classe P : les erreurs de type a) et b) peuvent s'expliquer par ce choix. En fait ces erreurs étaient plus ou moins prévisibles mais le manque de données contenant du *rap*, de la parole chantée sur fond musical et *a capella* nous a empêchés de créer une nouvelle classe les rassemblant.

3.4. Perspectives

Parmi les perspectives envisagées sur cette tâche, nous pouvons distinguer des perspectives à très court terme et des perspectives à 1 an et plus.

A court terme, il est nécessaire :
- étiqueter des données supplémentaires en terme de parole et musique.
- faire de l'adaptation pour nos modèles de classes.

de façon à augmenter notre corpus d'apprentissage, et de pouvoir se constituer des modèles plus robustes face aux différentes conditions de fonds sonores.

A échéance d'un an, les objectifs sont de définir un plus grand nombre de classes. En effet lors de l'examen des erreurs rencontrées en classification parole (P) et musique (M), nous avons conclu que beaucoup d'erreurs étaient dues à une intersection non vide entre les classes parole et musique. En particulier, les parties de *rap* comportent à la fois de la parole et de la musique : ces parties vont donc servir à la fois pour l'apprentissage de la classe parole et de la classe musique. Intuitivement on sent bien que ces segments vont induire des problèmes de modélisation. Il en est de même pour les segments de parole chantée (sur fond musical ou bien *a capella*).
Une première perspective consiste à définir un plus grand nombre de classes. Si le corpus de données nous le permet, l'idéal serait de créer une classe supplémentaire pour le *rap*, une pour la musique superposée à du chant et pour le chant *a capella*.
Cependant, la création de ces classes va engendrer un nouveau problème lié à la fusion des index issus des différentes classifications. En effet, pour l'instant les index Parole et Musique ne sont pas concurrents, en d'autres termes, la fusion des index consiste actuellement à faire une concaténation des deux types d'index (ce qui nous donne les quatre combinaisons suivantes : NP/NM, NP/M, P/NM et P/M).
En créant de sous-classes au niveau de la musique comme nous l'envisageons, ces sous-classes vont générer des index concurrents (M, *rap*, musique chantée et *a capella*) et la fusion de ces index consiste alors à faire un choix exclusif entre ceux-ci. Cette dernière tâche ne sera pas réalisée avant la troisième année du projet.

4. Information sur les locuteurs

4.1. Introduction

On peut extraire de l'information sur les locuteurs à partir d'un document sonore de plusieurs façons différentes. On suppose évidemment que la segmentation parole/musique a déjà été faite, et que les zones de silence ont également été identifiées.
On peut alors faire du suivi de locuteur, ce qui consiste à rechercher dans les segments de parole d'un document sonore toutes les plages qui ont été prononcées par un locuteur donné. Dans ce cas, il est nécessaire d'avoir déjà un modèle du locuteur en question, ainsi qu'un modèle du monde, c'est-à-dire un modèle censé représenter l'ensemble de la population.
Si aucune information sur des locuteurs n'est disponible, la segmentation en locuteurs féminins, en locuteurs masculins et en locuteurs enfants est envisageable. Ceci permet d'avoir des segments plus homogènes, en vue de faire notamment de la détection de mots clés en utilisant des modèles plus appropriés en fonction du type de voix.
On peut aussi faire de la segmentation en locuteurs, c'est-à-dire détecter tous les changements de locuteurs et apparier entre eux tous les segments qui correspondent à un même locuteur. On est alors capable de donner toutes les plages correspondant à ce locuteur.
Finalement, en supposant que nous disposons de la segmentation en locuteurs de plusieurs documents sonores, on peut rechercher au sein de tous ces documents des locuteurs communs. Ceci s'appelle l'appariement de locuteurs.
Nous désignerons ces quatre tâches respectivement par *suivi de locuteurs*, *détection du genre*, *segmentation en locuteurs* et *appariement entre locuteurs*. Ces quatre tâches seront traitées au sein du projet RAIVES. Elles seront d'abord évaluées sur des corpus de parole radiophonique de très bonne qualité, puis sur des données radiophoniques telles qu'on peut les trouver sur le Web. Au cours de cette première année, nous avons mis au point un système de suivi de locuteurs. Nous avons également commencé à développer un système de détection de genre et un système de segmentation en locuteurs.

4.2. Méthodes

Le suivi de locuteurs consiste à retrouver dans un document sonore toutes les plages qui ont été prononcées par un locuteur donné. Cette tâche commence donc par une phase d'apprentissage au cours de laquelle on apprend un modèle statistique pour chaque locuteur dont on voudra ensuite faire le suivi. Cela signifie en particulier que nous disposons de données d'apprentissage pour ces locuteurs. Il faut également faire l'apprentissage d'un modèle du monde, qui est un modèle censé représenter l'ensemble de la population. En pratique, on utilise des données provenant de plusieurs locuteurs féminins, de plusieurs locuteurs masculins et de plusieurs locuteurs enfants. On apprend un modèle statistique avec l'ensemble de ces données. Nous allons donc maintenant détailler l'ensemble des opérations qui permettent d'obtenir un modèle statistique, que ce soit un modèle de locuteur ou le modèle du monde. Ce processus est décrit dans la Figure 4.

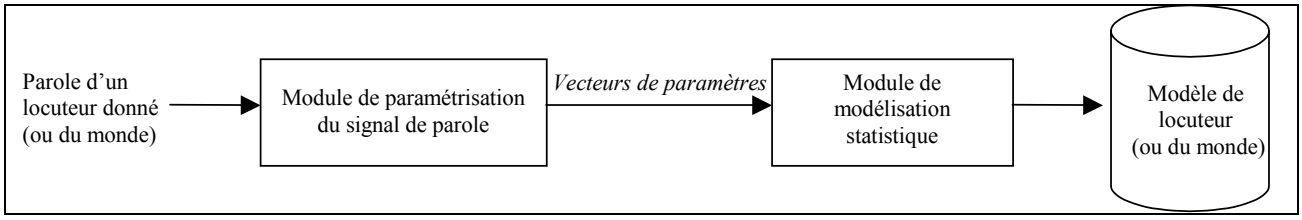


Figure 4: Représentation modulaire de la phase d'apprentissage d'un système de suivi de locuteurs.

- **Paramétrisation du signal de parole**

Comme le montre la Figure 5, la phase de paramétrisation du signal de parole consiste à transformer ce signal en une séquence de vecteurs de paramètres, utilisée ensuite pour apprendre un modèle statistique. C'est cette même paramétrisation qui sera appliquée au document sonore dans lequel on souhaite rechercher des informations sur les locuteurs. Considérons donc un signal de parole d'une certaine durée, qui peut être éventuellement la concaténation de plusieurs signaux entre eux. On applique à ce signal une analyse cepstrale.

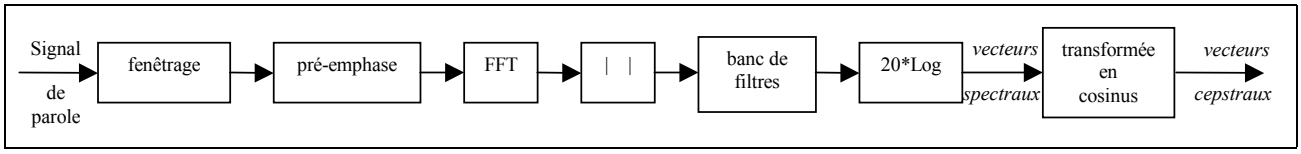


Figure 5 : Représentation modulaire d'une analyse cepstrale à base de banc de filtres.

On obtient finalement une séquence de vecteurs cepstraux représentant le signal de parole d'origine. On rajoute à ces vecteurs leurs paramètres Δ, c'est-à-dire une approximation de leurs dérivées premières. Les vecteurs de paramètres seront donc finalement constitués des coefficients cepstraux et des coefficients Δ-cepstraux.

- **Modélisation statistique**

A partir des vecteurs de paramètres, on effectue alors une modélisation statistique reposant sur un modèle par mélange de Gaussiennes. Ce modèle consiste à modéliser la distribution de probabilité des vecteurs de paramètres par une somme pondérée de Gaussiennes. Ce modèle GMM (pour Gaussian Mixture Model en anglais) est entièrement caractérisé par la donnée des poids du mélange, des moyennes et des matrices de covariance de chaque Gaussienne du mélange. L'algorithme EM (Expectation-Maximization) permet d'apprendre les paramètres du mélange de Gaussiennes à partir des données.

A l'issue de cette phase d'apprentissage, on obtient donc un modèle statistique de type GMM pour chaque locuteur dont on voudra faire ensuite le suivi. On obtient également un modèle GMM représentant le monde.

- **Suivi d'un locuteur**

La phase de suivi d'un locuteur consiste à retrouver, dans un document sonore, toutes les plages prononcées par un locuteur donné. La Figure 6 présente les différents modules mis en oeuvre.
On commence donc par paramétrer l'ensemble du document sonore, ou tout au moins les parties ayant été identifiées préalablement comme contenant de la parole. Puis on calcule pour chaque trame de parole sa vraisemblance avec le modèle du locuteur et sa vraisemblance avec le modèle du monde. On calcule le logarithme du rapport des vraisemblances. Afin de réduire l'influence des fluctuations locales de ce score, on réalise ensuite un lissage par le calcul de la moyenne des scores sur quelques trames précédentes et quelques trames suivantes. Le score d'une trame est donc en fait une moyenne des scores sur un bloc de trames autour de cette trame. Dans la pratique, on applique un fenêtrage de type Hamming sur le bloc de trames de manière à minimiser les effets de bord. On se retrouve donc finalement avec un score lissé pour chaque trame de parole du document sonore. Ce score est alors comparé à un seuil. S'il lui est supérieur, la trame est considérée comme ayant été prononcée par le locuteur considéré. Pour finir, on applique un filtre médian aux décisions trame à trame pour éviter par exemple d'avoir rejeté une trame alors que toutes les trames voisines ont été acceptées, ou *vice versa*.

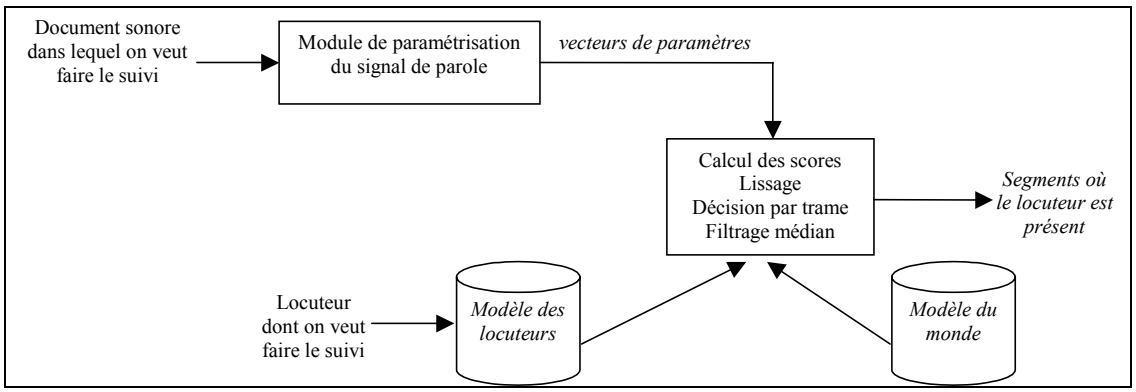


Figure 6 : Représentation modulaire de la phase de suivi d'un locuteur.

4.3. Evaluation

Seul le système de suivi de locuteurs a pour l'instant été évalué. Mais il n'a pas encore été évalué sur la base de données constituée dans le cadre du projet RAIVES. En effet, les données étiquetées pour l'instant ne permettent pas d'avoir plusieurs documents ayant un même locuteur présent dans des conditions propres, c'est-à-dire sans musique comme fond sonore. Nous avons donc décidé de

tester le système sur une autre base de données dans un premier temps. Nous avons choisi un sous-ensemble de la base de données HUB4. Cette base de données présente en effet l'avantage d'être constituée également de données radiophoniques, en langue anglaise, échantillonnées à 16 kHz, codées sur 16 bits, et entièrement étiquetées. Nous avons choisi un sous-ensemble constitué de 15 journaux télévisés de 30 minutes environ. Sept d'entre eux ont été utilisés pour l'apprentissage des modèles (un modèle du monde et trois modèles de locuteurs, une femme et deux hommes). Les huit autres ont été utilisés pour les tests.

Les valeurs utilisées pour la paramétrisation sont les suivantes : fenêtres de Hamming de 20 ms, décalées de 10 ms, pré-emphase avec un coefficient de 0.95, 24 coefficients de banc de filtres répartis sur une échelle linéaire, 16 coefficients ceptraux (sans le premier), pas de soustraction cepstrale, des coefficients Δ calculés sur 5 trames, et le rajout du Δ de la log-énergie.

Les GMMs sont constitués de 128 Gaussiennes diagonales et l'algorithme EM a été initialisé par un algorithme de quantification vectorielle (VQ). La taille des blocs pour le lissage des scores est de 31 trames. Et aucun filtre médian n'a été appliqué sur les décisions pour l'instant.

Les résultats de cette expérience sont donnés sur la Figure 7 sous la forme d'une courbe DET. Cette courbe présente les variations d'un type d'erreur en fonction d'un autre. Ici, nous avons le taux de non détection (pourcentage des trames non affectées au locuteur cible) en fonction du taux de fausse alarme (pourcentage des trames indûment affectées au locuteur cible). Ceci correspond à l'ensemble des points de fonctionnement du système pour toutes les valeurs possibles du seuil.

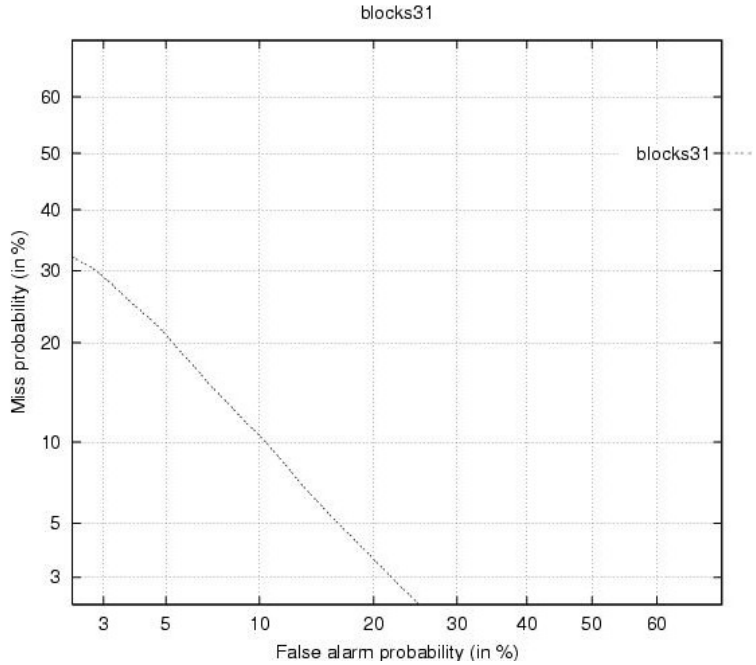


Figure 7 : Courbe DET pour le système de suivi de locuteurs.

Si on se place à l'EER, c'est-à-dire au point de la courbe pour lequel les deux taux d'erreurs sont égaux, on obtient un tout petit peu plus de 10 % d'erreur.

4.4. Perspectives

Il y a plusieurs perspectives sur la tâche de recherche d'informations sur les locuteurs pour la deuxième année du projet RAIVES. Tout d'abord, le système actuel de suivi de locuteurs sera optimisé sur le sous-ensemble de la base de données HUB4 choisi. En particulier, on fera varier plusieurs paramètres du système afin de trouver la meilleure configuration possible. On testera également le système sur les données RAIVES, dont on aura réalisé quelques étiquetages supplémentaires afin d'avoir suffisamment de données étiquetées en locuteur pour cette tâche. On continuera également le développement des systèmes de détection de genre et de segmentation en locuteurs, et on les testera aussi sur la base de données RAIVES.

Détection de genre : Dans le cas où l'on ne dispose pas de modèles particuliers de locuteurs, on peut néanmoins extraire de l'information sur les locuteurs. La première étape consiste à détecter le genre, c'est-à-dire si le locuteur d'une plage de parole est masculin, féminin, ou bien un enfant. La technique utilisée

est similaire à celle utilisée pour le suivi de locuteurs. Mais cette fois-ci, on apprend un modèle pour les locuteurs masculins, un modèle pour les locuteurs féminins et un modèle pour les locuteurs enfants. On cherchera alors pour chaque trame lequel de ces trois modèles donne la vraisemblance la plus forte, en moyenne sur un bloc autour de la trame, et on attribuera à cette trame l'étiquette correspondante. On utilise également un filtrage médian pour lisser les décisions. Cette technique est en cours de développement.

Segmentation en locuteurs : Après avoir identifié les plages de parole correspondant à des locuteurs masculins, féminins et enfants, on souhaite pouvoir segmenter chacune de ces plages en locuteurs différents, et apparier entre eux les segments correspondant à un même locuteur. La segmentation se fait à l'aide d'un système de détection de changement de locuteur reposant sur une approche BIC (Bayesian Information Criterion). L'appariement entre les segments se fait ensuite à l'aide d'un algorithme de classification hiérarchique. Cette approche est en cours de développement.

Appariement entre locuteurs : Finalement, l'appariement entre locuteurs consiste à apparier des locuteurs entre plusieurs documents sonores préalablement segmentés en locuteurs. Cette tâche nécessite donc la segmentation en locuteurs de plusieurs documents sonores, et permet de retrouver au sein de ces documents sonores des locuteurs communs. Cette tâche peut se faire également à l'aide d'un algorithme de classification hiérarchique. Cette tâche ne sera pas réalisée avant la troisième année du projet RAIVES.

5. Détection de mots-clés

5.1. Introduction

Les applications de type « Broadcast News » s'appuient sur une transcription complète du signal de parole [Gau99]. S'agissant de l'indexation par le contenu et plus particulièrement pour des applications de recherche sur le web, il n'est pas nécessaire d'avoir une transcription complète du signal, mais seulement de repérer un ensemble de mots qui correspondent à la recherche d'un utilisateur. La détection de mots-clés consiste à repérer temporellement un ensemble plus ou moins important de mots-clés dans le flux de parole.

Dans la littérature, les systèmes considérés comme donnant les meilleurs résultats reposent sur des modèles de Markov cachés [Ros90][Wil91]. Deux approches peuvent être distinguées. L'approche la plus répandue, et aussi la plus efficace, repose sur l'utilisation d'un système de reconnaissance grand vocabulaire couplé à une procédure de recherche des mots-clés dans la séquence reconnue [Wei93][Rah99].
Une autre approche utilise un « modèle du monde », c'est-à-dire modélise tout ce qui n'est pas un mot-clé [Man97]. [Kni94] donne un aperçu intéressant des différentes méthodes de « rescoring » qui peuvent être appliquées dans le cadre de cette approche. Une modélisation phonémique semble donner de meilleurs résultats qu'une modélisation globale du mot. Un des problèmes dans la détection de mots-clés est d'avoir une indication correcte des frontières. Partant du constat que les mots-clés semblent être particulièrement accentués dans un discours, l'information prosodique peut être intéressante. [Wan01] présente un système de détection de mots-clés basé sur un décodage acoustico-phonétique (MMC) couplé à un réseau de neurones prosodique. [Jun97] présente une modification de l'algorithme de Viterbi pour calculer des scores normalisés représentant la correspondance des mots-clés à différents endroits de la phrase ; la décision est prise par seuillage. Cette méthode est mise en oeuvre dans un module de post-traitement de façon très efficace [Fer01].
Si l'approche "reconnaissance grand vocabulaire" donne les meilleurs résultats, elle reste dépendante de la tâche et nécessite des bases de données d'apprentissage importantes. La méthode utilisant un « modèle du monde » permet quand à elle de pallier ces défauts. Elle donne des résultats intéressants dans le cadre d'une modélisation phonémique. Les systèmes basés sur un décodage acoustico-phonétique sont intéressants en terme d'indépendance vis-à-vis de la tâche et de la taille de la base de donnée d'apprentissage.

Dans le cadre du projet RAIVES, l'approche choisie est celle reposant sur une reconnaissance flexible à base de modèles de Markov cachés.

5.2. Méthodes

Nous avons choisi de mettre en oeuvre une méthode de détection des mots clés qui ne repose pas sur une transcription complète du signal, mais sur un décodage phonémique basé sur un "modèle du monde".
La Figure 8 décrit de façon synthétique le fonctionnement du système de détection des mots clés :

- un module d'apprentissage, permet à la fois d'apprendre nos modèles phonémiques (corpus BREF) et de les adapter au corpus Raives à partir de l'annotation manuelle disponible.
- le module de détection de mots clés permet de faire une auto-adaptation des modèles aux phrases de test avant de procéder à la détection des mots clés. Ne disposant pas de l'annotation manuelle pour les fichiers de test, l'auto-adaptation est faite à partir d'un module de décodage acoustico-phonétique. Ce module se sert des modèles adaptés en phase d'apprentissage. La détection des mots clés ne se base pas sur les résultats ne se module de DAP, mais

Chacun des modules mis en oeuvre est décrit de façon détaillé dans les paragraphes qui suivent.

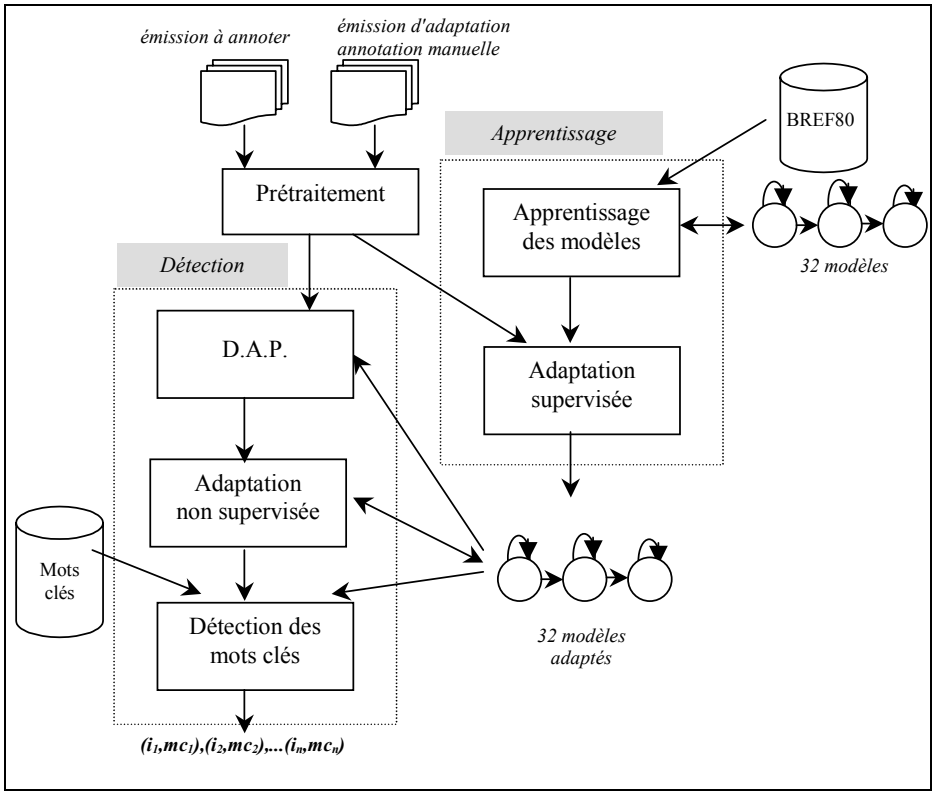


Figure 8 : Synoptique du système de détection de mots clés

5.2.1. Prétraitement

Les fichiers sont segmentés en terme de Parole/Musique. Nous ne retenons que les segments contenant de la parole. Ces segments peuvent éventuellement contenir de forts bruits de fond (musique, bruits, conversations,...). Ces segments sont à l'heure actuelle obtenus à partir de l'annotation manuelle des fichiers.

Nous avons choisi une paramétrisation de type cepstrale (Figure 5) :12 MFCC, 12 Δ MFCC, 12 $\Delta\Delta$ MFCC en ôtant C_0

Un programme de détection du canal de transmission a été mis au point afin de séparer la parole téléphonique de la parole non téléphonique. Ainsi, nous ne traitons que les segments de parole non téléphonique

5.2.2. Modélisation

La modélisation choisie est une modélisation markovienne (Hidden Markov Model). Un phonème est un modèle HMM indépendant du contexte à trois états. Chaque état est un mélange de lois gaussiennes. Cette modélisation à trois états permet de capturer les parties transitoires et stable du phonème (Figure 9).

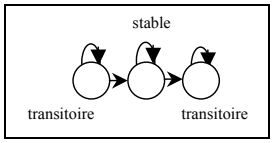


Figure 9 : modèle phonémique à trois états

Le "modèle du monde" doit représenter l'ensemble des phonèmes du corpus. Ce modèle du monde peut être soit un modèle de Markov à un état (assimilable à un GMM), communément appelé modèle "poubelle" ; soit être représenté par une boucle de modèles phonémiques parmi les 32 constituant le modèle du monde (31 phonèmes et une pause). Après différentes expérimentations, nous avons choisi d'implémenter la seconde solution (Figure 10).

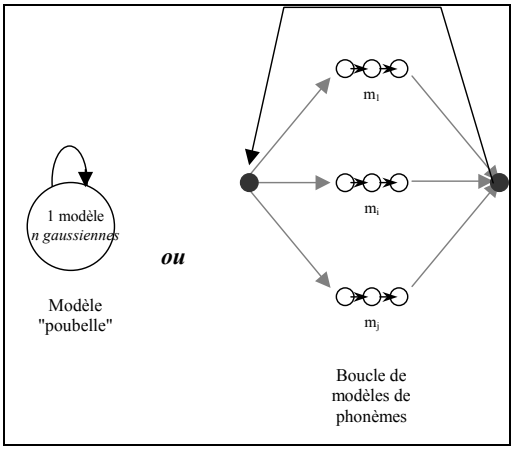


Figure 10 : Deux représentations possibles du Modèle du Monde.

Un mot clé est représenté par la concaténation des modèles de phonèmes qui le constituent. Nous prenons en compte les variantes de prononciations définies dans BDLEX et y ajoutons notre propre dictionnaire pour les mots inconnus de BDLEX. Il n'y a donc pas besoin d'apprentissage explicite des mots clés qui nécessiterait un corpus d'apprentissage important. De plus cela permet de construire de façon dynamique un nouveau mot clé (Figure 11).

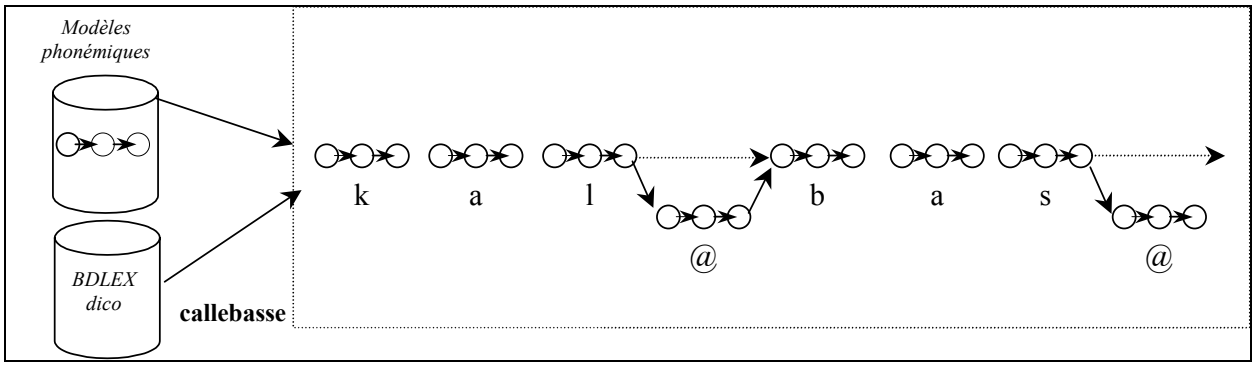


Figure 11 : Construction d'un mot clé. Prise compte des variantes de prononciations.

Avec de telles représentations, le problème qui se pose en phase de détection est que l'algorithme de décodage aura toujours tendance à privilégier le "modèle du monde" au profit des modèles de mots clés. Nous avons donc introduit une pondération qui est une fonction du nombre de phonèmes du mot clé.

5.2.3. Apprentissage et Adaptation

Le corpus utilisé dans le cadre de Raives est un corpus pour l'instant relativement restreint qui n'offre pas un volume de données suffisant pour pouvoir envisager un apprentissage satisfaisant des modèles. L'apprentissage des modèles se fait sur le corpus BREF80. Ce corpus est un corpus de 10 heures de parole lue dans des conditions d'enregistrement propres.

Les conditions acoustiques d'apprentissage sont très différentes de celles du corpus RAIVES : parole spontanée, bruits de fond, musique superposée,... Aussi, nous avons mis en oeuvre une ***adaptation supervisée*** des modèles par S-MLLR [Lau02]. Cette méthode, basée sur une structure arborescente permet d'estimer les paramètres des transformations quelque soit la quantité de données disponibles : plus il y aura de données, plus l'arbre sera profond et plus les gaussiennes seront adaptées avec précision. Avec peu de données, nous aurons une estimation plus globale des gaussiennes. L'adaptation se fait à partir des modèles appris sur BREF80, du fichier d'adaptation après prétraitement et de l'annotation manuelle du fichier (Figure 8**)**.

5.2.4. Détection des mots clés

Le corpus RAIVES contient des émissions diverses comme des bulletins d'informations (beaucoup de locuteurs natifs, bonnes conditions d'enregistrement, élocution peu hésitante,...) mais aussi des émissions de type interwiews et reportages (locuteurs non natifs, forts bruits de fond, musique, traduction simultanée...). Il a donc été nécessaire de mettre en place une ***adaptation non supervisée*** aux conditions acoustiques du fichier de test. Cette adaptation se fait par la méthode S-MLLR.

Dans la phase de détection, l'annotation manuelle n'est pas disponible. Nous réalisons une première étape de décodage acoustico-phonétique (DAP) grâce au moteur de reconnaisssance ESPERE [Foh00]. Les modèles sont adaptés à partir de la transcription obtenue.

5.3. Evaluation

Avant d'évaluer la détection des mots clés, nous avons voulu évaluer la qualité de la modélisation phonétique.

5.3.1. Evaluation du DAP sur le corpus RAIVES

Nous avons évalué les performances du décodage acoustico phonétique dans trois conditions différentes : sur BREF80, sur un bulletin d'informations (programme 7) et sur une émission 'Haute-Tension' (programme 8). Les résultats sont présentés dans le tableau (Tableau 3) ci-dessous.

	BREF80	Informations	Haute-Tension
Sans adaptation	70%	52.3%	31%
Adaptation supervisée (programme 2)	-	55%	35%
Adaptation non supervisée	-	57.7%	38.3%

Tableau 3 : Résultats de l'évaluation du DAP sur différents corpus et émissions radiophoniques.

L'étude de ces performances nous permet d'évaluer la pertinence de la mise en place des procédures d'adaptation. Les performances sur Bref80 sont de l'ordre de 70% ce qui peut être considéré comme de bonnes perfomances en DAP. La dégradation des résultats est de 17.7% entre BREF80 et le bulletin d'informations et de 39% entre BREF80 et Haute-Tension. Ceci montre bien la difficulté inhérente à la nature des signaux que nous traitons dans le corpus RAIVES : il est donc indispensable demettre en place des procédures d'adaptation. Si on analyse les résultats après adaptation (supervisée et non supervisée) nous avons un gain de 5,4% pour le bulletin d'information et de 7,3% pour l'émission 'Haute-Tension'.

5.3.2. Evaluation du système de détection de mots clés

Les modèles appris sur la base de données BREF 80 ont été adaptés de façon supervisée par S-MLLR sur le programme 2 qui a été étiqueté manuellement.
Le fichier de test est un bulletin d'information (programme 7). Ce programme comprend une grande diversité de locuteurs et très peu de musique. Nous avons défini 20 ensembles de mots clés définis autour d'un mot clé charnière : par exemple, le mot clé Amérique sera aussi associé aux mots américain(s) et américaine(s). 12 de ces mots clés appartiennent au fichier de test avec un total de 92 occurences.

Nous présentons ci-dessous (Figure 12) les résultats pour deux types de modèles différents (32 et 128 gaussiennes). Les résultats sont représentés selon le principe de la courbe ROC : pourcentage de détection correcte par nombre de fausses alarmes par mots clés et par heure de parole.

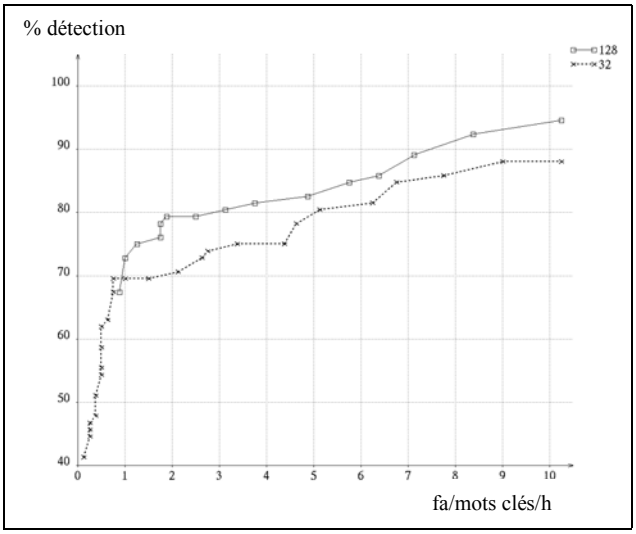


Figure 12: Courbe ROC présentant les résultats de la détection des mots clés : taux de détection par fausse alarme/mot clé et par heure

Les résultats obtenus sont satisfaisants et très encourageants. Dans la tâche de détection de mots clés nous obtenons environ 80% de bonne détection pour un

nombre de fausses alarmes de 2 par mot clé et par heure. Il est à noter que le taux de reconnaissance phonémique pour cette émission est de 57,7%. L'approche phonémique choisie est donc tout à fait appropriée pour faire de la détection de mots clés.

Ceci nous laisse penser que ce système est déjà suffisant pour envisager une tâche de détection de thèmes.

Nous présentons ci-dessous la localisation du mot-clé "Américain". Le signal temporel est présenté dans la partie inférieure et le spectrogramme dans la partie supérieure pour la phrase *"que le président américain devrait annoncer"*. La détection du mot clé *Amérique* est incluse dans le mot *américain* qui appartient au même ensemble de mots clés. La localisation du mot clé dans le flux de parole est très précise.

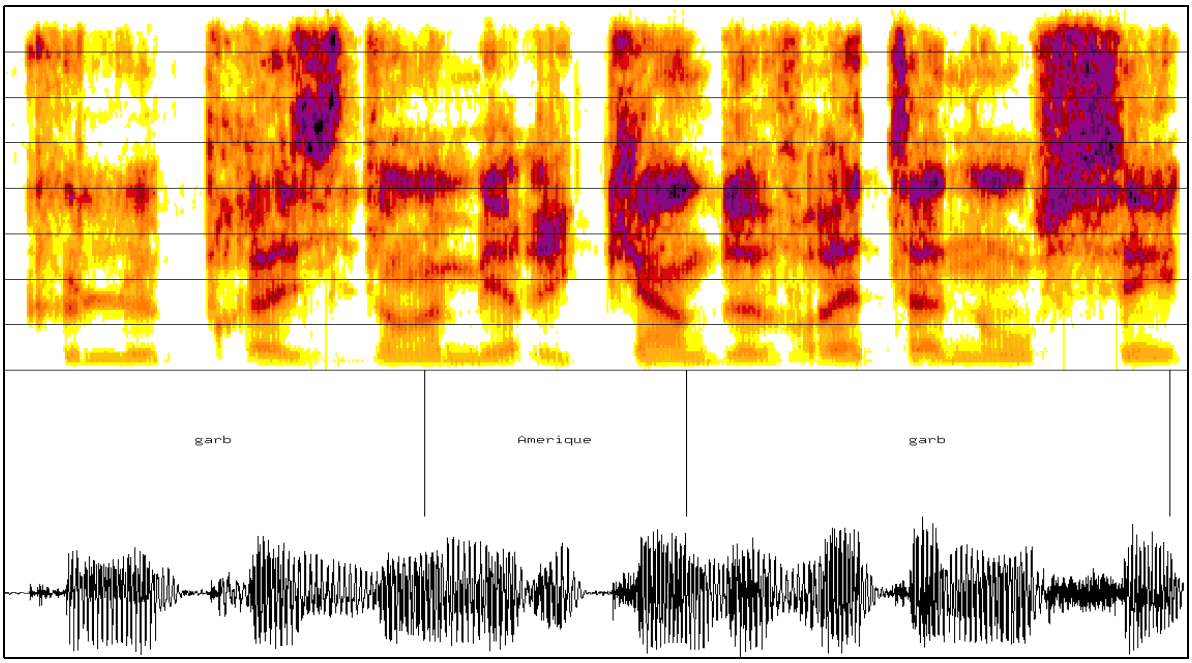


Figure 13 : Exemple de détection d'un mot clé : recherche du mot clé "Amérique" trouvé sur la séquence américain.

Il est aussi à noter que ce système est temps réel sur un PC classique, ce qui important au vu de l'application visée.

5.4. Perspectives

Parmi les perspectives envisagées sur cette tâche, nous pouvons distinguer des perspectives à très court terme et des perspectives à 1 an.

A très court terme, il est nécessaire de :
- mesurer les fréquences d'apparition des mots non outils.
- mesurer l'influence de la longueur des mots sur leur détection.

A échéance d'un an, les objectifs sont de :
- augmenter de façon importante le nombre de mots clés afin de juger de la viabilité d'un tel système par rapport à un système basé sur une transcription complète du signal.
- s'appuyer sur les résultats de segmentation automatique en locuteurs et détection de parole/musique afin d'automatiser complètement la procédure.
- mieux prendre en compte les différentes conditions de fonds sonores : musique, bruits, conversations. Nous pouvons par exemple envisager de mettre en place des méthodes de séparation de source, de débruitage ou d'adaptation plus fine des modèles.

5.5. Bibliographie

[Fer01] Ferrer L., Estienne C. (2001), "Improving performance of a keyword spotting system by using a new confidence measure", EUROSPEECH'01, pp. 2561-2564.

[For00] Fohr D., Mella O., Antoine C., "The automatic speech recognition engine ESPERE : experiments on telephone speech". In ICSLP. (Pékin, China). 2000.

[Gau99] Gauvain J.L., Lamel L., Adda G., "Audio partitioning and transcription for broadcast data indexation", CBMI'99, pp. 67-73.

[Jun97] Junkawitsch J., Ruske G., Höge H. (1997), "Efficient methods for detecting keywords in continuous speech", EUROSPEECH'97, pp. 259-262.

[Kni94] Knill K.M., Young S.J. (1994), "Speaker dependent keyword spotting for accessing stored speech", Tech. Rep. CUED/F-INFENG/TR-193, Cambridge University Engineering Department.

[Lau02] Lauri F., Illina I., Fohr D., "Comparaison de SMLLR et de SMAP pour une adaptation au locuteur en utilisant des modèles acoustiques markoviens", Actes des XXIVèmes Journées d'Etudes sur la Parole, 2002.

[Man97] Manos A., Zue V., "A segment-based wordspotter using phonetic filler models", ICASSP'97, pp. 899-902.

[Rah99] Mazin Rahim, "Recognizing connected digits in a natural spoken dialog", in Proc. of ICASSP'99, p. 153-156, Phoenix, Arizona, United States, 1999.

[Ros90] Rose R.C., Paul D.B. (1990), "A hidden Markov model based keywords recognition system", ICASSP'90, pp. 129-132.

[Wan01] Wang WJ, Lee CJ, Huang EF, Chen SH (2001), " Multi-keyword spotting of telephone speech using orthogonal transform-based sbr and rnn prosodic model", EUROSPEECH'01, pp. 2773-2776.

[Wei93] Weintraub M. (1993), "Keyword-spotting using SRI's DECIPHER large vocabulary speech recognition system", EUROSPEECH'93, pp. 1265-1268.

[Wil91] Wilcox L.D., Bush M.A. (1991), "HMM based wordspotting for voice editing and indexing", EUROSPEECH'91, pp. 25-28.

6. Publications

Le projet RAIVES a donné lieu à trois publications durant cette première année de fonctionnement.

6.1. Conférence invitée

[1]Ivan Magrin-Chagnolleau and Nathalie Parlangeau-Vallès, "Audio Indexing: What Has Been Accomplished and The Road Ahead," *Proceedings of JCIS 2002*, pp. 911-914, Durham, North Carolina, USA, March 2002 (Invited Paper).

6.2. Conférence internationale avec comité de lecture

[2] J. Pinquier, C. Sénac and R. André-Obrecht, "Speech and music classification in audio documents", *ICASSP'2002*, Orlando, Floride, Mai 2002

6.3. Soumission à conférence internationale avec comité de lecture

[3]Jérôme Farinas, Dominique Fohr, Irina Illina, Ivan Magrin-Chagnolleau, Odile Mella, Nathalie Parlangeau-Vallès, François Pellegrino, Julien Pinquier, Christine Sénac, and Kamel Smaïli, "The RAIVES Project: Toward Audio Indexing on the Web," *Submitted to ICASSP 2003*.

[4]Nathalie Parlangeau-Vallès, Jérôme Farinas, Dominique Fohr, Irina Illina, Ivan Magrin-Chagnolleau, Odile Mella, Nathalie Parlangeau-Vallès, Julien Pinquier, Jean-Luc Rouas and Christine Sénac, "Audio indexing on the web : a preliminary study of some descriptors", *Submitted to ICME'03*.

7. Budget : dépenses pour la première année

Le Tableau 4 est un récapitulatif des dépenses engagées au titre de l'année 2002. Les dépenses sont ventilées par laboratoire et catégorie. Il présente également la balance budgétaire du projet pour cette première année. L'excédent actuel est en grande partie dû au fait que nous n'avons pour l'instant engagé aucune dépense concernant la base de données. En effet, aucune convention n'a encore été signée avec RFI. Ce tableau ne fait pas état de la dotation de la deuxième année.

	Dépenses		**Dotation Globale**	**Balance**
	Missions	Matériel		
DDL	3596,41	4573,47		
IRIT	3005,62	2800		
LORIA	1382,57	3832,72		
Total	**7984,60**	**11206,19**	**30489**	**11298.21**

Tableau 4 : récapitulatif des dépenses ventilées par laboratoire et par catégorie. Dotation et balance.

8. Travail accompli au cours de la première année

Nous rappelons dans cette partie l'ensemble du travail accompli au cours de la première année du projet RAIVES.

Base de données : Nous avons recueilli un ensemble de données radiophoniques auprès de RFI (Radio France Internationale). Lorsque toutes les données auront été reçues, nous aurons environ 10 heures de documents radiophoniques dans 18 langues différentes, soit un total d'environ 180 heures de données. Pour l'instant, nous avons reçu l'intégralité de la partie du corpus en français, en espagnol et en allemand. Les données continuent à arriver au fur et à mesure.

Etiquetage des données : Nous avons étiqueté pour l'instant 8 programmes radiophoniques différents en langue française, ce qui correspond à environ 5 heures de données. Ces données ont été étiquetées en fonction de la présence ou non de musique, en locuteurs, et en transcription orthographique, ce qui a constitué un gros travail d'étiquetage, mais nécessaire pour l'évaluation des méthodes développées.

Segmentation parole/musique : Un premier système de segmentation parole / musique, reposant sur une modélisation différenciée à base de modèles par mélanges de Gaussiennes, a été réalisé et évalué sur la base de données RAIVES.

Information sur les locuteurs : Un premier système de suivi de locuteurs a été réalisé en utilisant une paramétrisation de type cepstrale du signal de parole et une modélisation statistique par mélange de Gaussiennes. Ce système a pour l'instant été évalué sur un sous-ensemble de la base de données HUB4, en attendant de pouvoir être évalué sur les données RAIVES.

Détection de mots clés : Un premier système de détection de mots clés a été réalisé. Les résultats actuels ont été évalués sur une petite partie du corpus RAIVES. Ils sont très encourageants et permettent d'envisager une tâche de détection de thèmes sans problèmes.

9. Liste des tâches à accomplir pour l'année 2003

9.1. Segmentation Parole/Musique

1. étiqueter des données supplémentaires en terme de parole et musique ;
2. faire de l'adaptation pour les modèles de classes ;
3. définir un plus grand nombre de classes ;
4. fusion des décisions.

9.2. Information sur les locuteurs

1. 1optimiser le système de suivi ;
2. faire tourner le système de suivi sur les données RAIVES ;
3. tester le détecteur de genre sur HUB4 et RAIVES ;
5. mise au point du système de segmentation en locuteurs et test sur HUB4 et RAIVES.

9.3. Détection de mots clés

1. améliorer l'adaptation aux données ;
2. augmenter le nombre de mots clés pour tester la fiabilité et l'utilisabilité d'un tel système ;
3. s'appuyer sur les résultats de segmentation automatique en locuteurs et détection de parole/musique afin d'automatiser complètement la procédure ;
4. mieux prendre en compte les différentes conditions de fonds sonores : musique, bruits, conversations. Nous pouvons par exemple envisager de mettre en place des méthodes de séparation de sources, de débruitage ou d'adaptation plus fine des modèles en utilisant les résultats de la segmentation parole/musique et de la segmentation du locuteur ;
5. amorcer une réflexion commune avec la tâche de détection de thème.

9.4. Identification de la langue

1. appliquer un système complet sur les données RAIVES : décodage acoustico-phonétique, multigrammes et pseudo-syllabes ;
2. gérer les changements de langue dans le flux de parole.

9.5. Détection de thèmes

Une étude sera menée à partir des résultats du module de détection des mots clés.

9.6. Détection de sons clés

Une étude prospective sera menée sur cette tâche.

9.7. Ingénierie

Des stagiaires seront employés pour faire de l'ingénierie dans le cadre du projet :
1. intégration des trois modules et test sur des fichiers tests.
2. réalisation d'un prototype de démonstration

ANNEXE 1

Petit guide de l'annotation sous Transcriber

1. Information sur le locuteur (*procédure en cours de révision*)
2. Transcription orthographique
3. Annotation Parole/Musique

1. Annotation des locuteurs

Création d'un locuteur
Ajouter l'information communication téléphonique sur la fenêtre locuteur.
Ajouter l'information relative au sexe.

2. Transcription orthograhique

La transcription orthographique se fait sur *la ligne verte*. Sur cette ligne apparaîtront aussi un ensemble d'événements notables.

Important : s'il y a un doute sur un mot, une expression, il faut l'isoler en le mettant sur une nouvelle ligne, cela évite de perdre une partie de parole trop importante.

a) Règles de transcription orthographique

Les paroles doivent être transcrites en suivant ce qui est exactement prononcé, même si c'est une faute de français. Par exemple, le locuteur dit "y'a pas", on transcrit "y a pas".

Acronymes. Les acronymes épellés comme SNCF devront être notés S_N_C_F. Les autres sigles, comme FNAC par exemple s'écriront en majuscule.

Allongements de syllabe finale. Terminer le mot par $. *Ex: Théophile$*

Chiffres et les nombres doivent être écrits en toutes lettres. *Ex : un, quatre-vingt deux francs cinquante, trente quatre, cinq millions,...*

Interjections sont à transcrire. *Ex : euh, hum, ah, oh, ouh...*

Mot coupé, pas terminé. Transcrire ce qui a été prononcé en terminant par une *. *Ex : si capital est coupé à la dernière syllabe, on transcrit capi*.*

Néologisme. Transcrire phonétiquement et terminer par *. *Ex : keuf rissaillezer.*

Noms propres doivent être noté avec la première lettre en majuscule. *Ex: Théophile*

Orthographe incertaine donner une transcription phonétique probable et terminer le mot par *. *Ex : Aouakoundé*.*

Reprises. Transcrire ce qui est prononcé. *Ex : cent trente non cent quarante.*

Ponctuation. Ne pas mettre de virguel, point virgules, point,...

Majuscules. Les majuscules sont à conserver uniquement pour les noms propres et un point de rupture ou une nouvelles phrse ne motivent pas une majuscule.

b) Les événements

Il existe plusieurs catégories d'événements (Figure 14 : Ensemble des types d'événements disponibles). Ces catégories sont : bruit, prononciation, lexique et langue. La catégorie "comment" ne servira pas dans le cadre de l'application. ***Bien penser à noter les silences, pauses, respirations qui sont des événements.***

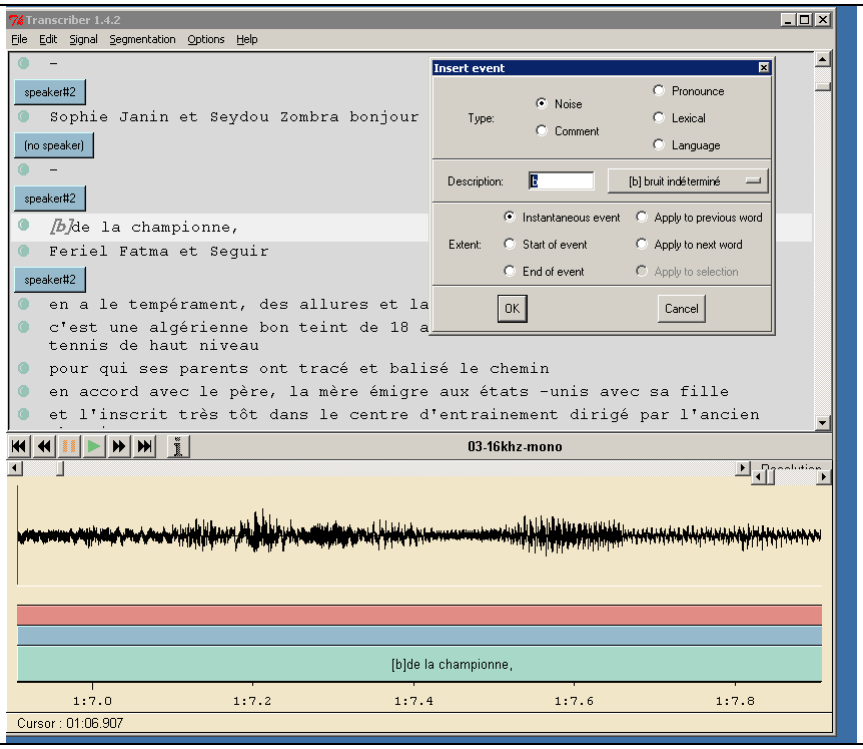
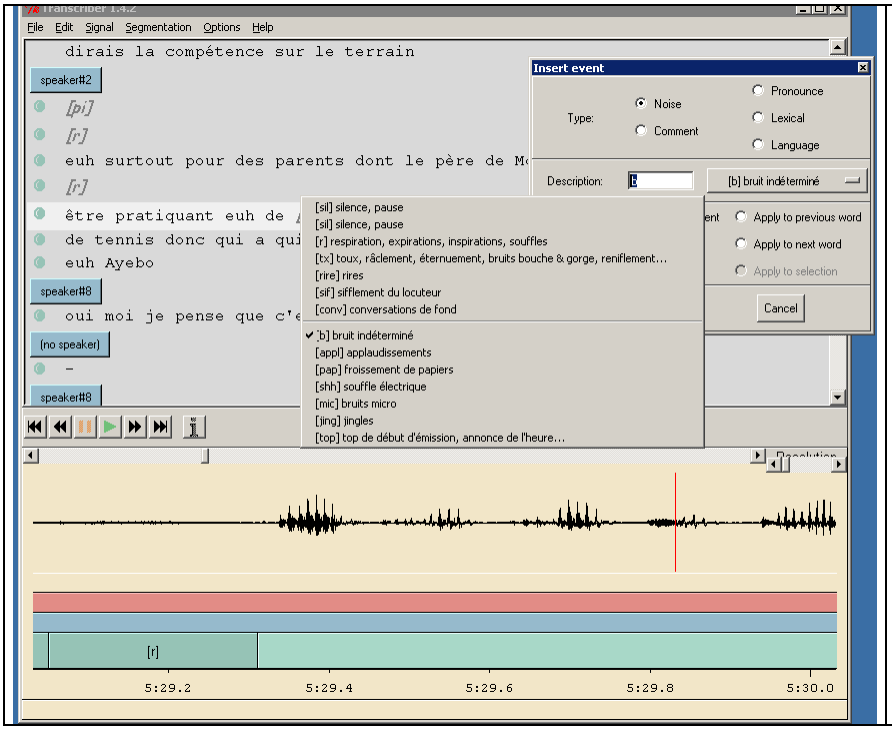
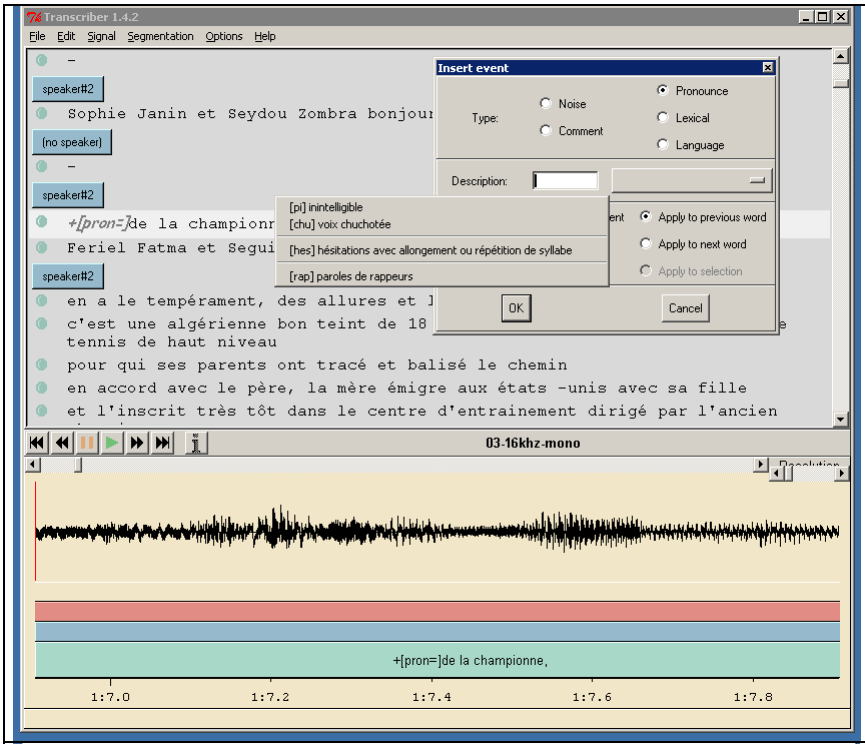


Figure 14 : Ensemble des types d'événements disponibles



La plupart des étiquettes parlent d'elles-même...

Si des événements ***sont superposés*** à de la parole, transcrire la parole sur la même ligne.

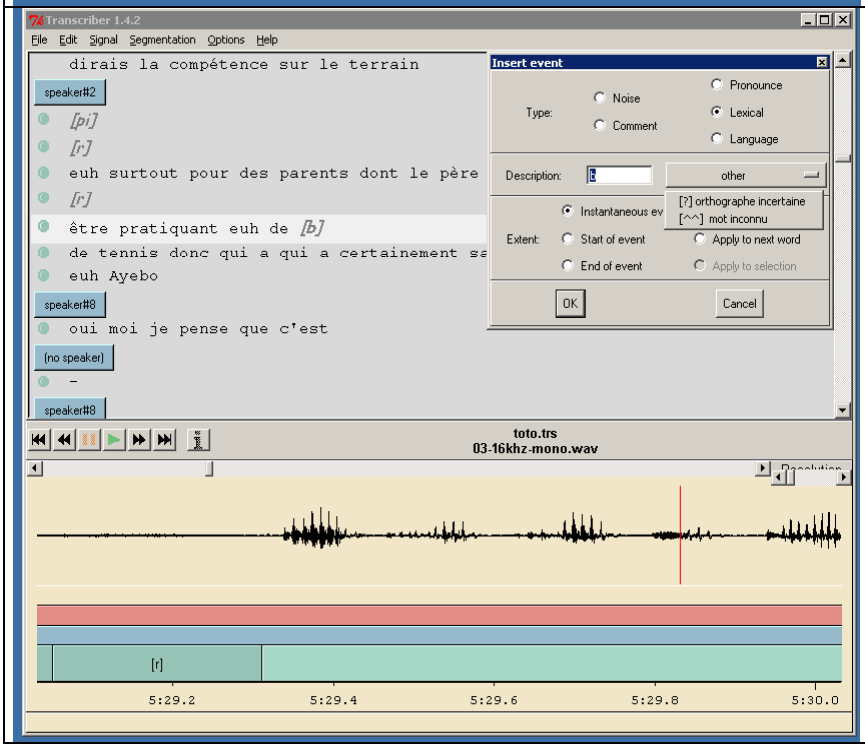


[pi] parole inintelligible, on ne sait vraiment que transcrire...

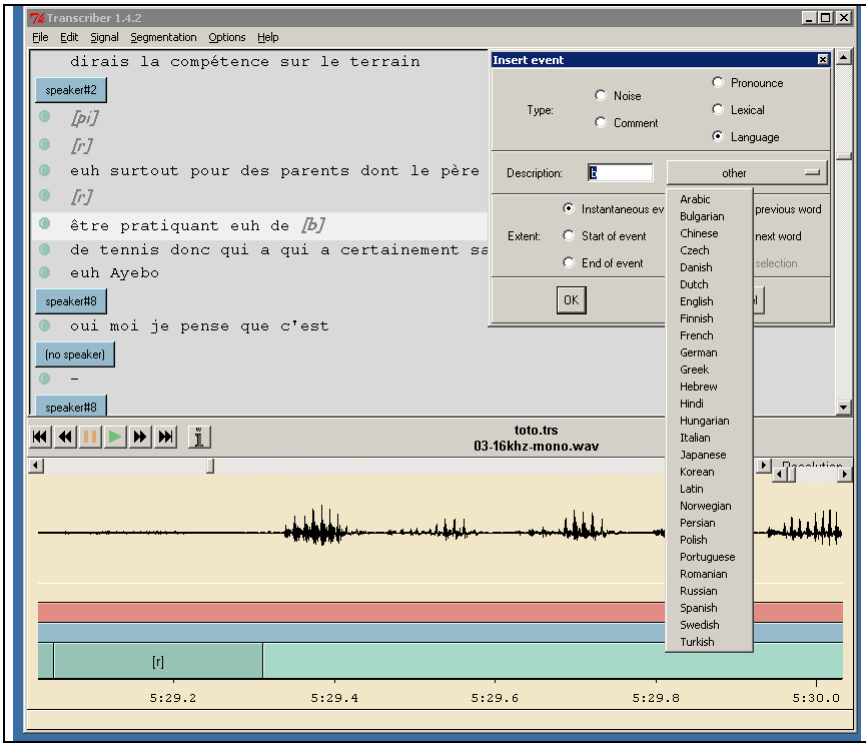
[chu] voix chuchotée. Intelligible, mais niveau sonore vraiment bas.

[hes] hésitations. Soit une syllabe répétée (mamaman) soit une syllabe allongée en fin de mot.

[rap] paroles de rappeurs.



Isoler le mot inconnu ou mal orthographié

	Isoler le mot ou la phrase en langue étrangère.

3. Annotation Musique/NonMusique

a) Ensemble des catégories

Catégorie - Etiquette	Signification		Exemple sonore
M	musique	Tous les styles de musique instrumentale	
VC	voix chantée	Chant a capella, sans musique	
MVC	musique & voix chantée	Chant avec musique en fond	
NM	non musique	Tout ce qui n'est pas l'une des catégories précédentes. Ex : parole, cris, bruits isolés...	

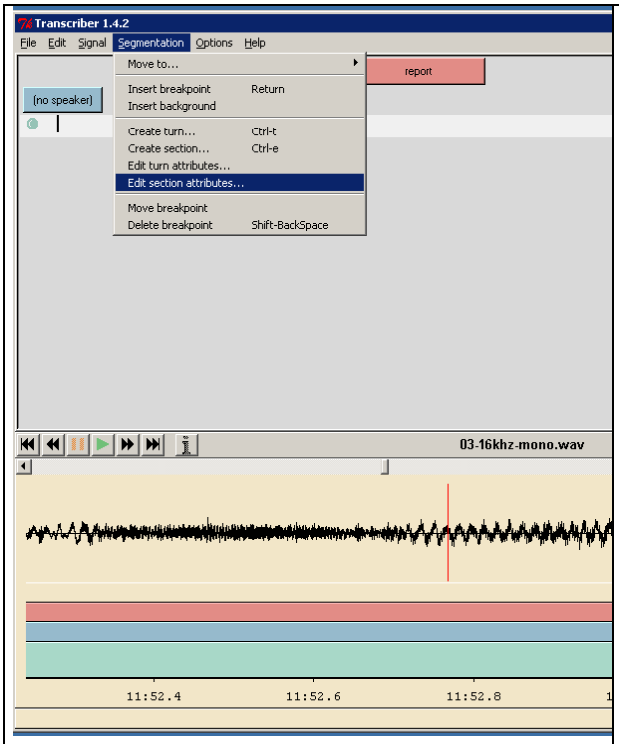
Tableau 5 : tableau des catégories concernant l'annotation Musique/ Non Musique.

b) Annotation

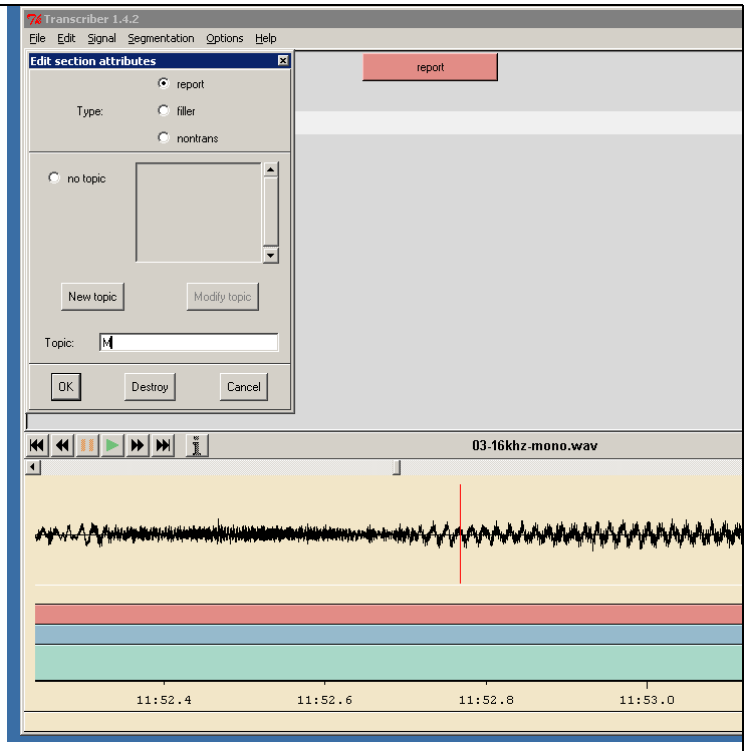
L'annotation de la musique se fait sur *la ligne rouge*, c'est-à-dire la ligne des *SECTIONS.*

Préambule important :

Il faut recréer l'ensemble des catégories lorsque l'on ouvre une nouvelle transcription.
Pour cela, il faut :



1- Choisir le menu segmentation;edit section attributes.



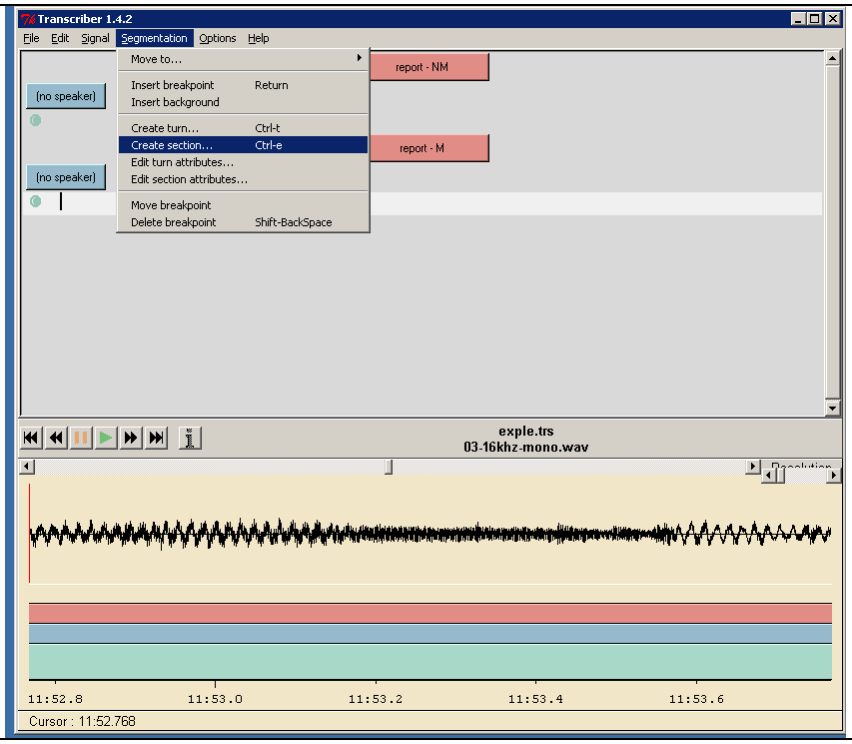
2- Créer chacune des catégories :
- avec l'étiquette correspondante
- et dans l'ordre présenté dans le *Tableau 5.*

Ceci est **important** afin de faciliter la cohérence entre les différentes transcriptions.

Procédure d'annotation :

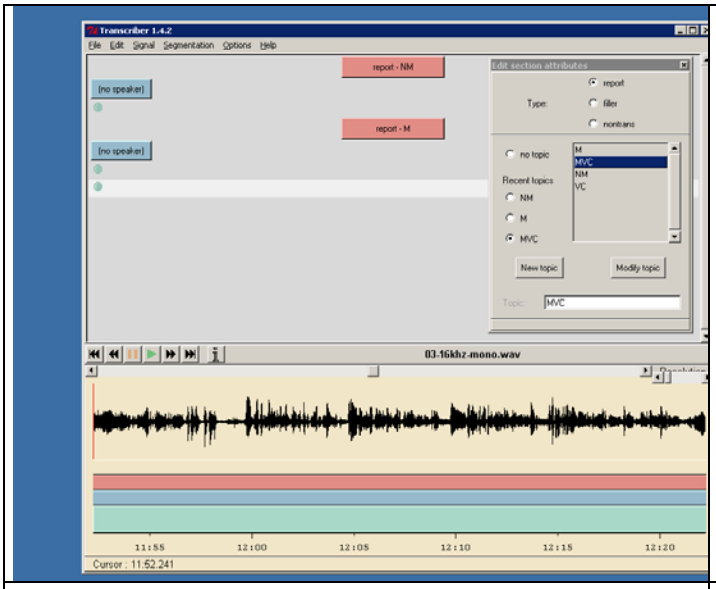
1- Insérer un point de rupture (return ou segmentation;insert breakpoint), sans cela il est impossible de créer une section.

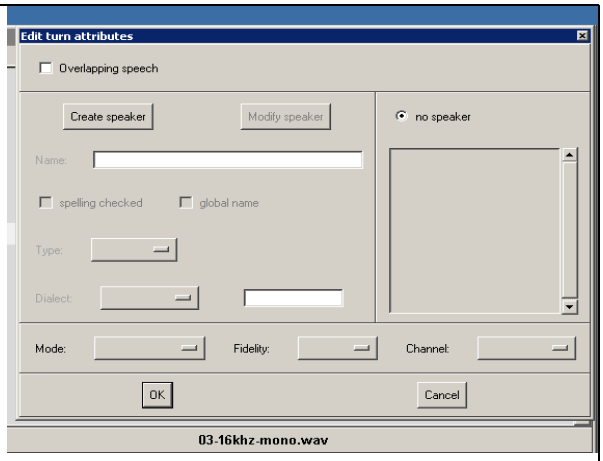
2- Créer la section



➢ Créer la section. menu segmentation;create section.

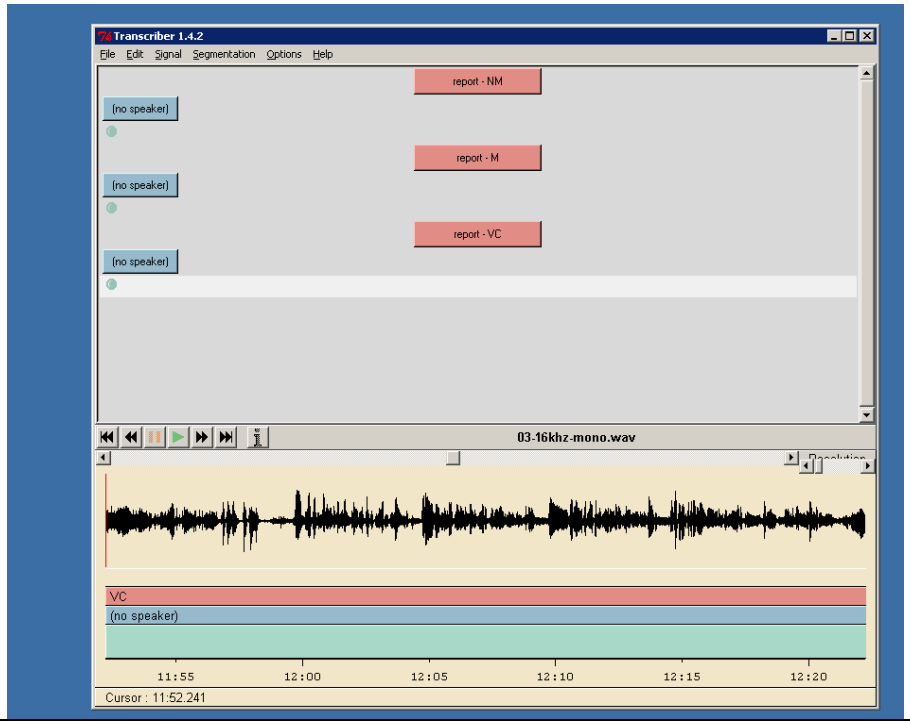
3- Affecter une catégorie à la section.





- Ne pas changer le type, toujours laisser le type report.
- Choisir la section : M, NM, VC, MVC
- Valider par *return*

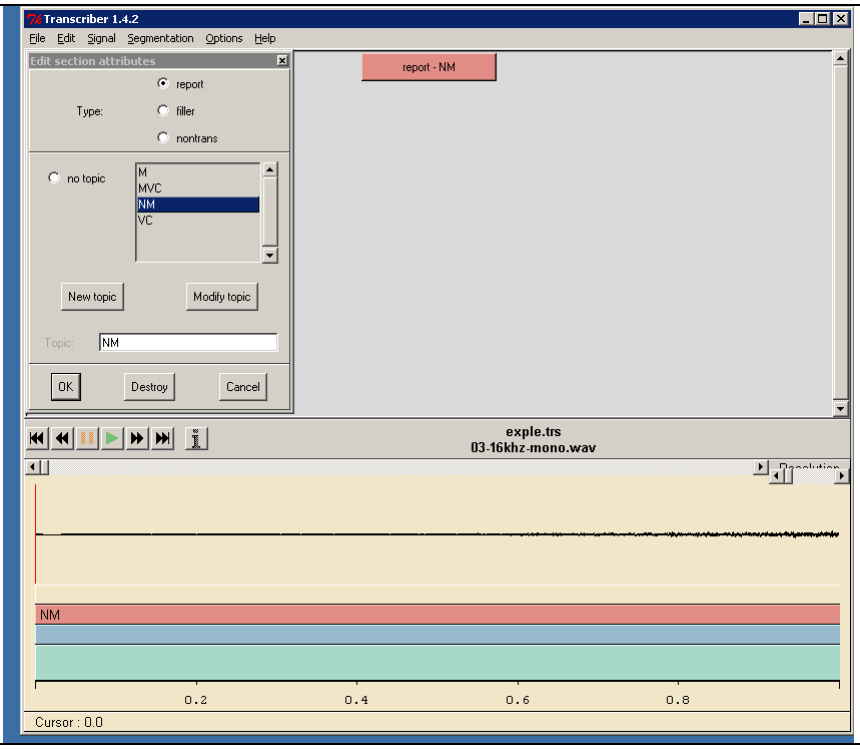
- Cliquer sur ok sans rien modifier



- Voici le résultat

En cas d'erreur

Que ce soit une erreur de catégorie ou alors une erreur dans la création de la section, on peut suivre la démarche suivante :



- Cliquer sur la section
  - Soit on peut changer de catégorie.
  - Soit on peut détruire la section. Il peut être utile ensuite de détruire aussi le point de rupture (segmentation;delete breakpoint).

Transcription en format transcriber

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<!DOCTYPE Trans SYSTEM "trans-13.dtd">
<Trans   scribe="(unknown)"   audio_filename="03-
16khz-mono" version="1" version_date="020614">
<Topics>
<Topic id="to1" desc="NM"/>
<Topic id="to2" desc="M"/>
<Topic id="to3" desc="VC"/>
<Topic id="to4" desc="MVC"/>
</Topics>
<Episode>
<Section        type="report"        startTime="0"
endTime="702.502" topic="to1">
<Turn startTime="0" endTime="702.502">
<Sync time="0"/>


</Turn>
</Section>
<Section        type="report"        topic="to2"
startTime="702.502" endTime="712.241">
<Turn startTime="702.502" endTime="712.241">
<Sync time="702.502"/>

</Turn>
</Section>
<Section        type="report"        topic="to3"
startTime="712.241" endTime="1432.0933125">
<Turn startTime="712.241" endTime="1432.0933125">
<Sync time="712.241"/>

</Turn>
</Section>
</Episode>
</Trans>
```

Ensemble des catégories

Transcriber prend des conventions de notations.
to1 = NM

Description de la 1ère section

Le format est différent pour la première section. La catégorie se trouve après les marqueurs temporels.

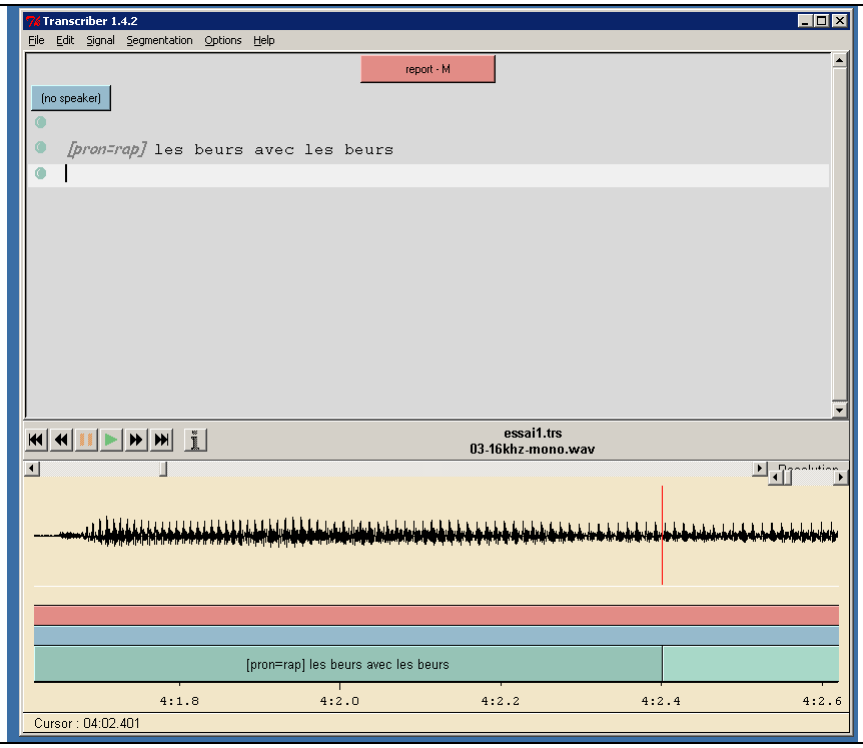
Description d'une section

Cas particulier du Rap

Pour le rap, nous n'avons pas transcrit les paroles... Je laisse néanmoins cette partie.

Etant donné que l'on ne peut réellement considérer le rap comme de la voix chantée, il faut un traitement particulier. Les paroles sont généralement accompagnées de musique, mais pas toujours. Il faut donc pouvoir donner l'information rap en dehors de l'étiquetage de la ligne musicale.
Le choix se porte sur un événement [rap] qui sera suivi de la transcription orthographique.

Cela donnera le résultat suivant :



- Insérer un événement de type isolated noise : edit;insert event; isolated noise
- Choisir un événement de type prononciation
- Choisir [rap]
- Mettre la transcription orthographique
- Les lignes vides signifient que c'est de la musique seule.