# Recherche des expressions clés caractéristiques de l'interaction entre locuteurs dans les documents audiovisuels

par Archange Giscard Destiné

**Responsables de stage : Isabelle Ferrané**
**Jérôme Farinas**

**Mots Clés :** Extraction d'informations, extraction d'entités nommées, détection du rôle des locuteurs, segmentation et regroupement en locuteur, détection de termes clé, reconnaissante automatique de la parole

**Résumé :** La détection du rôle du locuteur permet de classer les intervenants d'une émission radiophonique en 3 grandes catégories *Présentateur, Journaliste et Autre.* Il est possible d'affiner les rôles (*Journaliste chroniqueur, envoyé spécial, correspondant, …*) et de récupérer des informations sur les locuteurs (nom, leur statut) pour caractériser de façon plus précise leur fonction dans les échanges. Cette étude consiste à déterminer les expressions clés qui permettent de cibler les endroits qui contiennent potentiellement ces informations et d'étudier la robustesse de ces termes.

**Keywords:** Information and named entity extraction, speaker role detection, speaker diarization, keyword detection, automatic speech detection

**Abstract:** Detection of speaker's role enables to classify stakeholders of a radio broadcast in three broad categories: presenter, journalists and others. It is possible to refine these roles (Journalist reporter, special correspondent, correspondent...) and retrieve information about users (name, status) to characterize more precisely their role in discussions. The aim of this study is to identify terms in recognized sentences that allow the tracking of information and to study the robustness of these terms.

**Année 2010 - 2011**

## Remerciements

Au terme de cette expérience, j'adresse mes profonds remerciements à tous ceux qui ont contribué à ma réussite. J'exprime toute ma reconnaissance envers mes parents et amis pour leur soutien moral toujours présent.

Je compte aussi remercier tous les enseignants et professeurs qui avaient participé à ma formation avant ces études et à tous les professeurs qui ont fait de leur mieux pendant cette année académique pour transmettre leur savoir sur les domaines. Je remercie plus particulièrement les professeurs Régine André-Obrecht et Christine Sénac qui ont su montrer toute la beauté du domaine du traitement automatique de la parole.

J'adresse mes remerciements aux différents cadres de l'IRIT, et plus particulièrement de l'équipe SAMoVA. Je remercie les encadrants Isabelle Ferrané et Jérôme Farinas pour leur confiance et leur disponibilité dès les premiers moments où j'ai manifesté un intérêt pour ce sujet de recherche, jusqu'aux moments de la rédaction de ce rapport. Je tiens aussi à remercier Benjamin Bigot pour ses conseils et son aide inestimable qui m'a permis de mieux appréhender des concepts liés à son travail de thèse. Je remercie tous les membres cette institution (chercheurs, thésards, stagiaires …) qui permettent de créer un bon environnement de travail avec parfois des moments de franches camaraderies.

Je remercie aussi tous ceux et celles qui ont guidé pour une bonne réussite de mon séjour sur Toulouse. Mes remerciements vont aussi à tous ce qui ne sont pas mentionné ici et qui ont directement ou indirectement contribué à la réussite de ces études.

# Table des matières


| | |
|---|---|
| **Remerciements** | **2** |
| **Table des matières** | **3** |
| **1 – Introduction** | **6** |
| 1.1 – Contexte | 6 |
| 1.2 – Problématique | 7 |
| 1.3 – Objectifs du travail | 8 |
| 1.4 - Organisation du document | 9 |
| **2 – Cadre théorique** | **10** |
| 2.1 - Le décodage acoustico-phonétique | 10 |
| 2.2 - La reconnaissance automatique de la parole | 11 |
| 2.3 - L'extraction des entités nommées | 12 |
| 2.4 - Structuration en zones d'interactions et détection de rôle | 14 |
| 2.4.1 - Segmentation et regroupement en locuteurs | 14 |
| 2.4.2 - Détection des zones d'interaction | 15 |
| 2.4.3 - Détection de rôle | 15 |
| **3 – Cadre applicatif** | **16** |
| 3.1 - Le projet EPAC | 16 |
| 3.2 - Les ressources utilisées | 16 |
| **4 - Termes clé comme marqueurs des entités nommées** | **19** |
| 4.1 - Les termes clé | 19 |
| 4.2 - Les acte de dialogue | 20 |
| 4.3 – Approches de la recherche des termes clés | 21 |
| 4.4 – Construction de la liste des termes clés | 21 |
| **5 – Fiabilité et Robustesse des termes clé** | **24** |
| 5.1 – Approche | 24 |
| 5.2 - Interprétation des résultats | 35 |
| **6 - Conclusions et perspectives** | **36** |
| **Annexes** | **38** |
| Annexe A : Une émission radiophonique interactive | 38 |
| Annexe B : Les variations dans la prononciation des mots | 41 |
| Annexe C : L'alphabet phonétique international dans le système | 43 |
| Annexe D : Liste des termes clés obtenus | 44 |
| **Bibliographie** | **46** |

## Terminologie

Avant de détailler le sujet et de présenter ma contribution, il est nécessaire, pour une meilleure compréhension de la suite, de définir les concepts de base utilisés. Le sens proposé ici est celui qui sera retenu et considéré comme connu et sans ambigüité dans les pages qui suivent sauf indication contraire.

**Locuteur**

Il désigne toute personne qui prend la parole. Il peut aussi être désigné par le terme *intervenant.* Placé dans le contexte de la recherche en cours, le terme *locuteur* fait référence à la personne qui produit le signal de parole.

**Rôle**

Du fait du contexte de la thèse de Benjamin Bigot, le terme *rôle* désignera le profil du locuteur. Les sources audio qui seront principalement utilisées sont des émissions présentant plutôt des interactions entre locuteurs : des émissions de débat par exemple. Dans la présente étude, on considèrera qu'il existe, suivant la fonction du locuteur, trois grandes catégories de rôles : présentateur, journaliste, autre.

Suivant la durée et la fréquence de leurs interventions, on distingue les locuteurs qui sont ponctuels (qui parlent une seule fois) de ceux qui ne le sont pas. Un présentateur n'étant pas ponctuel par principe, cette distinction est plus pertinente pour les deux autres catégories : *Journaliste* et *Autres*. Ceci conduit à considérer les cinq *rôles* suivants :

- Présentateur
- Journaliste *ponctuel*
- Journaliste *non ponctuel*
- Autre *ponctuel*
- Autre *non ponctuel*

Le *Présentateur* est le locuteur qui intervient le plus souvent dans un programme, il introduit les autres locuteurs au cours de l'émission.

Le *Journaliste* peut être par exemple un chroniqueur, un envoyé spécial.

Les autres locuteurs sont classés dans la catégorie *Autre*. On y retrouve, entre autres : un interviewé, un spécialiste, un invité, …

**Tour de parole**

Ce terme sera utilisé dans ce document pour désigner l'intervention d'un locuteur dans une conversation [Bunt2007]. La distribution des tours de parole est régie par des normes sociales et contextuelles qui font entre autres qu'il ne peut pas y avoir de trop longs silences dans un tour de parole.

# 1 – Introduction

## 1.1 – Contexte

Ce travail constitue le rapport du stage que j'effectue à l'Institut de Recherche en Informatique de Toulouse (IRIT) dans le cadre de la formation en Master 2 Recherche en Informatique et Télécommunication suivie à l'Université de Toulouse III. Il porte sur la recherche des expressions clé qui caractérisent l'interaction entre locuteurs dans les documents audiovisuels.

Dans ce cadre, l'équipe Samova de l'IRIT travaille entre autres sur la structuration de contenu audiovisuel. Le travail de thèse, effectué par Benjamin Bigot, a porté sur la détection du rôle du locuteur et sur l'apport de la connaissance du rôle pour la structuration. L'objectif est de segmenter le contenu audio visuel en macro-segment pouvant correspondre à des interviews (interaction entre deux locuteurs) ou à des débats (interaction entre plusieurs locuteurs). En effet, la connaissance du rôle peut aider à caractériser ces segments, le présentateur (animateur de débat par exemple) ayant une place centrale notamment en distribuant la parole aux autres interlocuteurs. Ce travail s'est basé sur 3 grandes catégories de rôle : *Présentateur*, *Journaliste* et *Autre* et cherche à détecter et à caractériser les zones d'interaction entre locuteurs. Ces trois catégories de rôle seront précisées plus tard dans le document.

Dans le travail mené par Benjamin Bigot sur la détection du rôle il apparaît que le rôle du *Présentateur* est très bien détecté (90%). Tandis que la confusion entre *Journaliste* et *Autres* reste importante. Dans les données à notre disposition (essentiellement des émissions de radio comme nous le verrons ultérieurement), la catégorie *Autres* regroupe des rôles assez diversifiés allant du simple auditeur appelant pour poser une question à une personne invitée en tant qu'expert sur un sujet particulier. *Journaliste* et *Autres* étant des catégories assez génériques, il devient nécessaire des les affiner. Une autre difficulté qui apparait dans cette étude (celle menée par Benjamin Bigot) est que lors d'interviews assez longues, l'invité peut être confondu avec le présentateur.

L'objectif du travail qui m'a été proposé est de mener une étude sur les différentes possibilités d'affiner la détection de rôle en extrayant les informations sur ce qui est dit dans le contenu audio : le nom de l'intervenant et son statut éventuel (auditeur, invité, expert, …).

## 1.2 – Problématique

Pour accéder à ce qui est dit, il faut accéder à la transcription. Il existe déjà des méthodes qui permettent d'extraire les informations contenues dans un texte. En effet, beaucoup de techniques d'extraction d'informations sur des enregistrements audio consistent d'abord à le transcrire de façon automatique pour ensuite applique un des méthodes d'extraction sur le texte obtenu pour récupérer par exemple des noms propres, des noms de villes, des noms de professions.

| Exemple : |
| --- |
| Prononcé : Françoise dans le Var est en ligne bonsoir Françoise. |
| Transcrit : **Françoise dans le Var** est en ligne bonsoir **Françoise** |

Cependant cette façon de procéder présente quelques difficultés. Il n'est toutefois pas toujours possible de récupérer les informations par ces méthodes qui dépendent de la transcription. En effet, l'application de ces méthodes présente les limites que voici.

**- La transcription n'est pas toujours disponible**

**- La transcription (automatique) contient parfois des erreurs**

La transcription peut être soit manuelle, soit automatique. Il est possible d'extraire les informations pertinentes à partir du texte. Mais appliquer ces techniques d'extraction d'information sur un volume important de donnée n'est pas toujours envisageable. Les techniques d'extraction d'information à partir de sources audio-visuelles comportent donc souvent une étape de transcription du contenu. Les transcriptions automatiques sont alors les plus utilisées. L'extraction d'entités nommées présente alors des problèmes liés à cette dépendance à la qualité de la transcription. Mais ces données automatiques contiennent beaucoup d'erreurs relativement (par rapport aux manuelles).

En effet, même avec les systèmes les plus performants, il existe des erreurs de transcription au niveau des termes introduisant les entités nommées et aussi au niveau de ces entités qui rendent presqu'inenvisageable une technique d'extraction des noms des locuteurs à partir du texte qui aura été obtenu après le traitement du signal de parole par un système de reconnaissance vocale. Illustrons une partie des erreurs obtenues un outil de reconnaissance vocale affichant des performances de l'état de l'art.

| |
|---|
| Prononcé : France-Inter le journal de *dix neuf heures* **et c'est** *Denis Astagneau*<br>Transcrit : france-inter j' en mal de dix neuf heures **essais** denis astagneau |
| Prononcé *:* … une véritable impression de gâchis. ***Ecoutez*** *Valérie Rodrigo.* Elle est élue du ...<br>Transcrit : … une véritable impression de gâchis **et coûter** valérie rodrigo elle est élue du … |

Dans les deux exemples mentionnés dans les exemples 1 et 2 ci dessus, on remarque que les erreurs se portent au niveau des mots ou expressions introduisant l'information pertinente. Ceux-ci seront désignés par *termes clés* ou *expressions clés*. A la place des mots corrects, le système de transcription propose des mots ayant le même son. Ces termes clés sont donc récupérables automatiquement. Par contre, les noms des personnes (dans ces cas précis) sont bien transcrits. Il faut préciser aussi que le moteur utilisé pour la transcription des ces contenus audio conversationnels avaient les noms des différents locuteurs dans le dictionnaire utilisé.

Il existe parfois aussi des erreurs de transcription au niveau même de l'information recherchée. Dans l'exemple ci-dessus, l'erreur se porte sur le statut d'une personne dans l'émission.

| |
|---|
| Prononcé *: ...sur place* ***notre envoyé*** *spécial Philippe Reltien…*<br>Transcrit : … sur place **tout premier** spécial philippe reltien … |
| Prononcé *:* ... en tout cas une pratique disons régulière. ***Professeur Guy-Grand*** *!*<br>Transcrit : … en tout cas une pratique disant régulière **l'un des grands**. |

On constate donc que les transcriptions peuvent aider à extraire les informations sur les locuteurs quand elles ne présentent pas des erreurs sur les termes clés ou sur les données recherchées. Dans le cadre du stage, je m'intéresse à l'extraction de ces informations à partir d'un enregistrement audio dans le cas ou la transcription automatique ne peut pas fournir les informations nécessaires (à cause des types d'erreurs mentionnés).

## 1.3 – Objectif du travail

Une observation sur les interactions entre locuteurs montre que les informations recherchées se trouvent presque systématiquement autour de certaines expressions (ex. *Bonjour* M. Dupont).

Une première phase du travail consiste à générer une liste d'expressions permettant de connaître la localisation des informations pertinentes permettant de raffiner les rôles des locuteurs, d'extraire des informations supplémentaires sur les locuteurs (nom, statut, spécialité)

Une étude de la robustesse de ces termes clé permettra de vérifier si effectivement ces termes peuvent être utilisés comme marqueurs d'entités nommées pour récupérer les informations dans les cas où la transcription automatique comporte des erreurs au niveau des termes clés ou des informations pertinentes.

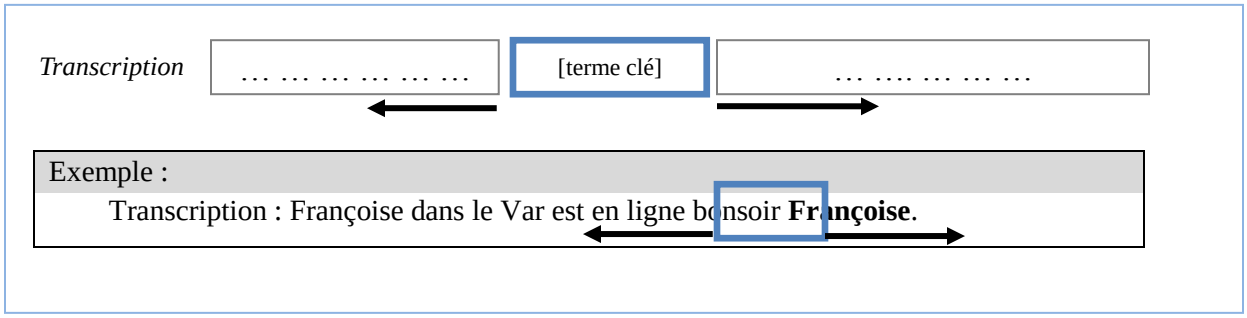


*Figure 1 : Localisation des informations pertinentes*

## 1.4 – Organisation du document

Pour mener cette étude sur les expressions clés dans l'interaction entre locuteurs dans les documents audiovisuels, différents niveaux de traitement et résultats sont utilisés et méritent d'être brièvement présentés.

En effet, les chapitres 2 et 3 présentent respectivement le cadre théorique (le décodage acoustico-phonétique, la reconnaissance automatique de la parole et l'extraction d'entités nommées sont développées de façon sommaire) et les corpus utilisés dans le déroulement de cette étude.

Les chapitres 4 et 5 concernent les deux grandes parties des travaux. La première constitue une étude sur les termes clés. Le chapitre 5 rapporte l'étude sur la robustesse de ces termes clés comme marqueur des entités nommées.

A la fin du document, quelques lignes sur les apports de cette étude dans le domaine de l'extraction de l'information et plus précisément sur l'extraction des entités nommées et les perspectives envisageables constituent la conclusion de ce travail.

# 2 – Cadre théorique

Au cours de mes travaux, j'ai eu à utiliser des traitements et résultats de traitements qu'il a été nécessaire de maitriser : le décodage acoustico-phonétique (DAP), la reconnaissance automatique de la parole (RAP), la segmentation et regroupement en locuteur (SRL), l'extraction d'informations, la détection du rôle. Cette partie du document est consacrée à une brève présentation de ces systèmes.

## 2.1 - Le décodage acoustico-phonétique

Le terme décodage acoustico-phonétique de la parole désigne le processus automatique qui permet de générer à partir du son produit une suite d'unités linguistiques qui soient la plus proche possible du signal produit. Cette unité linguistique, de façon classique, est le phonème. Mais elle peut aussi être un dérivé du phonème (diphones, phonèmes en contexte...). [Tuba1989]

*La paramétrisation*

Le signal de parole présente des parties stationnaires sur une période de temps (fenêtres) relativement petite, 20-30ms. La paramétrisation acoustique (ou simplement paramétrisation) consiste à extraire de ces fenêtres, des vecteurs de valeurs (coefficients ou paramètres) qui contiennent les informations nécessaires pour décrire le signal. Il s'agit donc de l'une des premières étapes de tout traitement automatique de la parole.

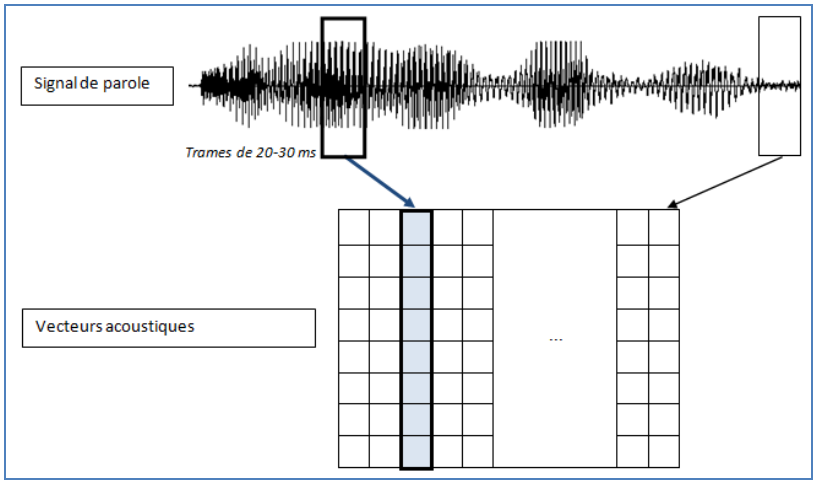


*Figure 2 : Paramétrisation*

*Le moteur de décodage*

Cette étape constitue le cœur de la partie applicative du système. Il produit une transcription en unités acoustiques à partir de l'analyse des paramètres extraits du signal de parole. Ce traitement utilise pour cela des données issues d'une phase d'apprentissage : les modèles acoustiques, un vocabulaire et éventuellement un modèle de langage.

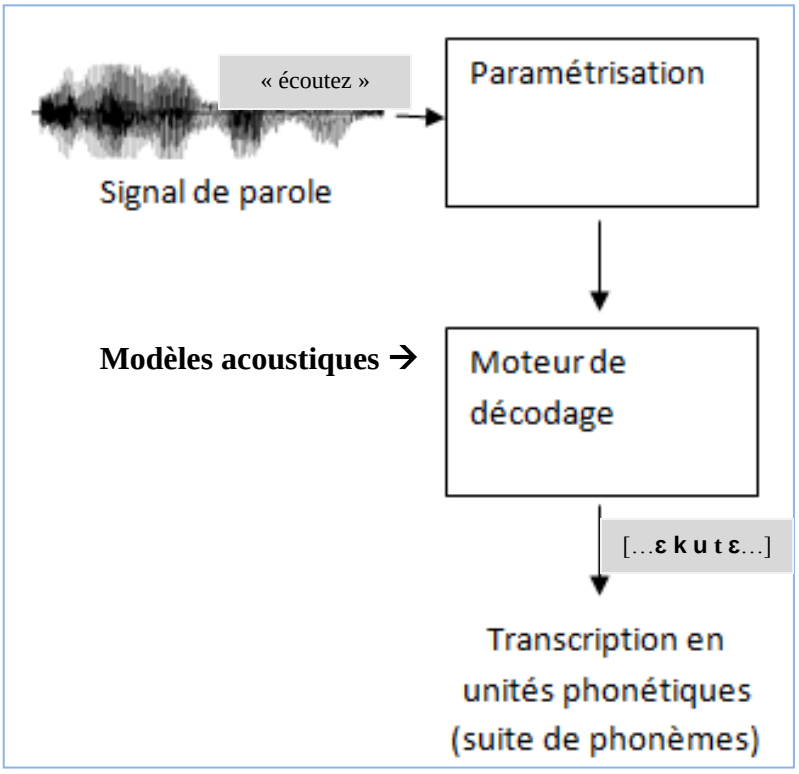


*Figure 3 : Principe du Décodage Acoustico-Phonétique*

## 2.2 - La reconnaissance automatique de la parole

La reconnaissance automatique de la parole est le domaine du traitement de la parole qui permet de produire à partir du signal de parole, le texte correspondant. La constitution d'un système de traitement automatique de la parole suit un principe général qui est présenté dans cette partie du document. Les différences entre les systèmes viennent des exigences fixées lors de la conception. Par là, j'entends des exigences en termes de vocabulaire qu'un tel système doit pouvoir reconnaitre (allant de 10 à 50000 mots), le cadre d'utilisation (monolocuteur, multilocuteur, indépendant du locuteur), du milieu dans lequel il est appelé à être utilisé (grand public, difficiles, ou conditions de laboratoire).

*Reconnaissance de parole grand vocabulaire*

Il s'agit des systèmes conçus pour déterminer, à partir du signal de parole, la phrase qui a été prononcé. Ceci exige un vocabulaire de grande taille (pouvant aller jusqu'à 65,000 mots) prenant en compte les noms propres qui sont plus ou moins connus. Un modèle de langage et une grammaire permettent de décider sur une suite cohérente de mots quand plusieurs suites proches du signal produit. [Hato2006] [Mari2002]

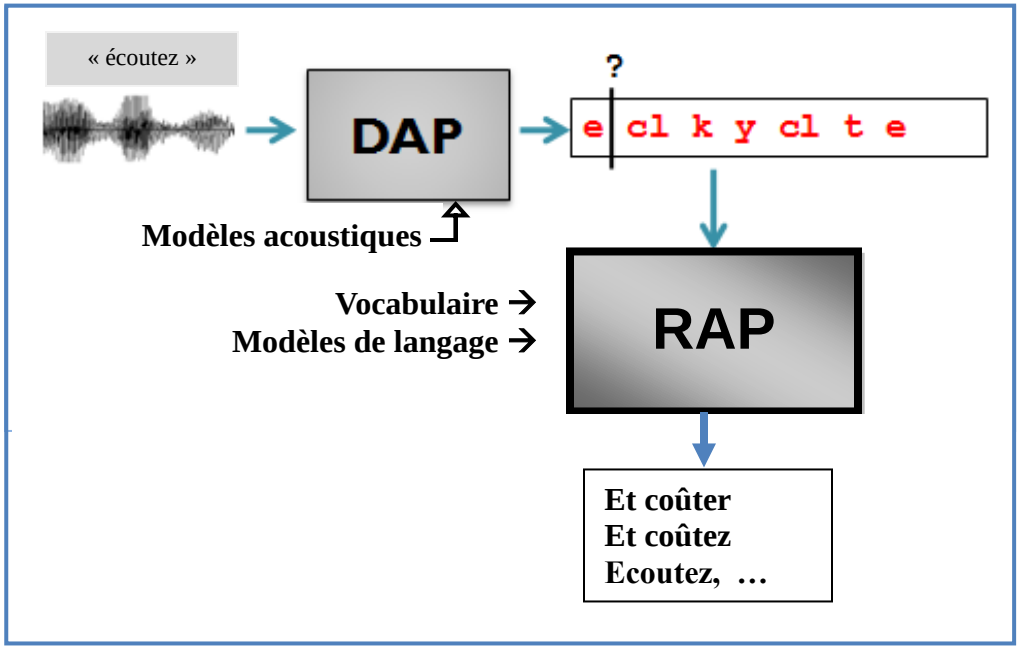


*Figure 4 : Architecture générale du système de transcription automatique*

## 2.3 - L'extraction des entités nommées

***Entités nommées***

Le terme *Entités nommées* désigne des expressions faisant référence à des personnes, des objets, des lieux … Dans le cadre de ce travail, la taxonomie utilisée sera celle proposée par dans la campagne ESTER. En effet, l'ensemble d'entités nommées choisies est constitué de 8 grandes catégories : personnes, lieux, organisations, groupes sociopolitiques, quantités, temps, produits, des installations. [Grav2004]

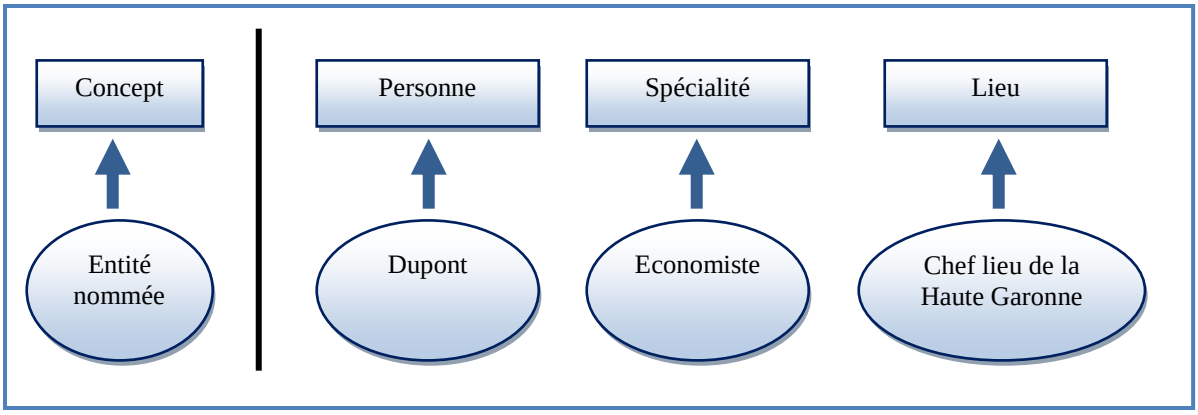


*Figure 5 : exemple d'entités nommées*

Un document sera jugé informatif ou pas suivant qu'il contienne ou non des entités nommées. La valeur sémantique de l'information qu'elles véhiculent constitue tout l'intérêt qui leur est porté [Bres2005]. Extraire les entités nommées (noms propres ou chiffre) permet en effet d'extraire de l'information pertinente de la source audio.

***Extraction des entités nommées***

Si les entités nommées transportent une grande partie de l'information du document, il faut non seulement pouvoir extraire ces noms propres (ou chiffre) mais aussi pouvoir déterminer à quoi ils font référence. Un moyen de pouvoir le déterminer est de prendre en compte le mot clé qui l'introduit. A titre d'exemple, ces deux phrases présentent un nom identique : « Je suis **Monsieur Dupont**, … » et « Bonjour **Monsieur Dupont**. Bienvenue à … ». Mais l'extraction de cette entité nommée sans prendre en compte les termes clé l'introduisant réduit considérablement la valeur de l'information qu'elle véhicule.

Il existe plusieurs techniques d'extraction d'entités nommées. Ces techniques sont en général basées sur des recherches à partir de la transcription du contenu. L'extraction des entités nommées étant du domaine de l'extraction d'information suscite beaucoup d'intérêt dans la communauté scientifique.

Il est difficile de donner un niveau de performance pour les systèmes qui existent. En effet, un système affichant un très bon taux de réussite dans un domaine donné peut ne pas être utilisable dans d'autres domaines d'application.

Les tous premiers travaux d'extraction des entités nommées (EN) au début du XX siècle visaient l'extraction d'information dans des articles de journal. D'autres domaines n'allaient pas tarder à utiliser ces traitements : traitement des rapports militaires, les conversations téléphoniques, le secteur médical.

*Les performances*

Appliqués sur la transcription manuelle, les meilleurs systèmes ont un taux d'erreur de 10%. Tandis que le taux d'erreurs affiché pour quand la transcription automatique est utilisée varie entre 40 et 50%. [Gall2009]

## 2.4 – Structuration en zone d'interaction et détection de rôle

### 2.4.1 – La segmentation en locuteurs

La segmentation en locuteur consiste à détecter, dans un signal de parole, les différents changements de locuteur dans le signal de parole. Ces dix dernières années des recherches autour du traitement automatique de la parole ont manifesté un grand intérêt pour la segmentation en locuteurs d'un flux audio. Ceci s'explique par le fait que la segmentation en locuteur a un apport considérable dans plusieurs autres applications telles que classification des contenus audio, la transcription de la parole conversationnelle, le suivi du locuteur.

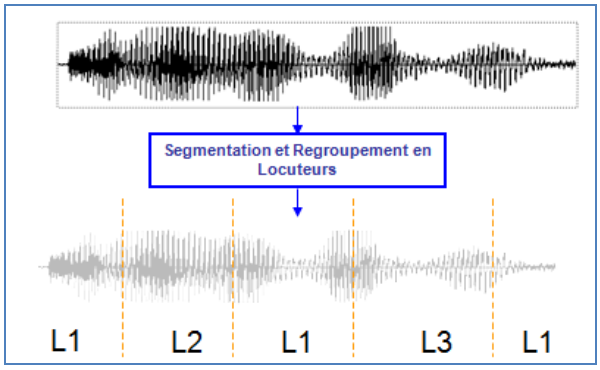


*Figure 6 : Localisation des informations pertinentes*

Dans l'équipe la technique utilisée pour effectuer cette tâche est celle développée par El Khouri [Khou2010]

### 2.4.2 – Détection des zones d'interaction

Etant donnée qu'il est devenu possible avec la SRL de détecter les plages d'intervention des locuteurs. Le travail de Bigot consiste à détecter les zones dans l'enregistrement vidéo ou il y a eu des échanges de parole entre locuteurs. Elles sont désignées par le concept zones d'interactions.

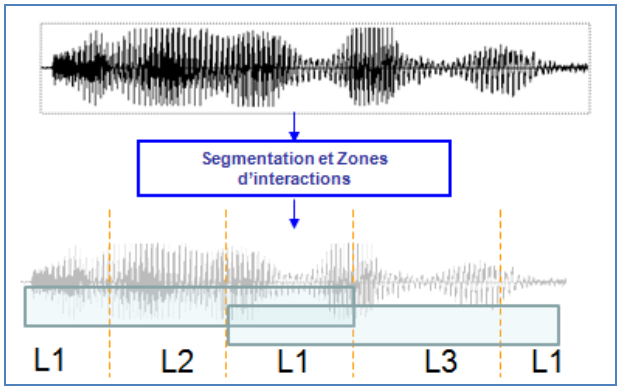


*Figure 7 : Détection des zones d'interaction*

### 2.4.3 – Détection de rôle [Sala2010]

La distribution des locuteurs sur le flux de parole permet non seulement de détecter les zones d'interaction mais aussi de détecter, sans avoir recours à la transcription, les rôles des locuteurs. Comme il est mentionné plus tôt dans le document, le locuteur sera classé Présentateur, Journaliste ou Autre. [Bigot2011]

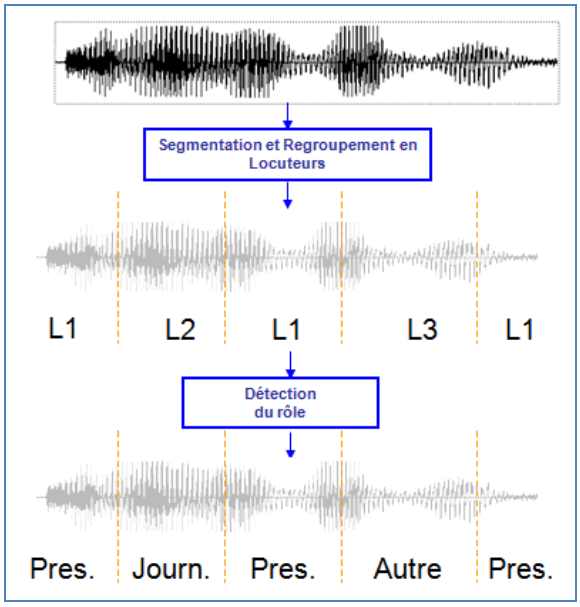


*Figure 8 : Détection du rôle du locuteur*

# 3 – Cadre applicatif

Le chapitre précédent nous a permis d'expliquer les notions qui seront nécessaires à la compréhension de mon travail. Dans ce chapitre est détaillé le cadre applicatif des recherches menées dans l'équipes SAMoVA et en particulier le projet EPAC et les ressources de utilisées dans mon stage de recherche.

## 3.1 – Le projet EPAC

Les méthodes d'extraction d'information, jusqu'ici, étaient généralement élaborées pour traitées des documents bien structurés. Le Projet EPAC, mené par quatre laboratoires académiques (IRIT, LI, LIA, LIUM) vise la proposition de nouvelles méthodes d'extraction d'information et de structuration de données audio non structurées. L'accent sera mis plus précisément sur le traitement de la parole conversationnelle (plus ou moins spontanée). Aussi le corpus EPAC est-il constitué d'émissions radiophoniques francophones présentant des interactions entres locuteurs parmi environ 1500h disponibles (1500 heures plus 100h manuellement annotées).

Le corpus utilisé dans le cadre de mon travail est issu du corpus EPAC puisque ces données sont bien appropriées pour les études des interactions entre locuteurs. La section 3.2 présente une description ces données.

## 3.2 – Les ressources utilisées

Le caractère conversationnel des données du projet EPAC fait qu'elles sont tout indiqué pour nous permettre de mener cette étude sur les expressions clés retrouvées dans les échanges entre des locuteurs.

L'ensemble des fichiers d'enregistrement audio (WAV) est constitué d'émissions où on s'attend à de la parole spontanée et des interactions entres des locuteurs : les émissions de débat, le journal, … Les fichiers peuvent être classés suivant leur contenu en 3 grandes catégories : France Inter, France Culture, RFI.

| | **France Inter** | **France Culture** | **RFI** |
|---|---|---|---|
| **Nombre de fichier** | 38 | 38 | 25 |
| **Volume horaire** | 55h<br>(30 h de Téléphone) | 75h | 125h |

Outre ces fichiers d'enregistrement audio on dispose d'autres ressources. En effet pour la première partie de travail consistant, rappelons-le, à extraire une suite de termes clés utilisés par les présentateurs dans les émissions à interaction, il est nécessaire d'avoir des transcriptions (théoriquement) sans erreur.

*a) Les fichiers de transcription manuelle -TRS*

Etant effectué manuellement, ces fichiers d'extension (.TRS) sont considérés comme la transcription de référence des fichiers audio utilisés dans ce travail.

Les informations des ce fichier utilisées dans le cadre de mon travail sont la transcription de la parole, indications de temps, la durée des tours.

```
<Turn speaker="spk10 spk13" mode="spontaneous" fidelity="high" channel="studio" startTime="1983.14" endTime="1985.460">
<Sync time="1983.14"/>
<Who nb="1"/>
commentaire supplémentaire , monsieur Parent ?
<Who nb="2"/>

<Event desc="i" type="noise" extent="instantaneous"/>
 oh juste une suggestion
</Turn>
...
```

*b) Les fichiers de transcription automatique - TRANSAUTO.XML*

Ces fichiers contiennent la transcription automatique des fichiers WAV. Les données de ces fichiers qui sont utilisé dans le cadre de ce travail seront : le texte généré par le système de transcription automatique (dans les balises <text>), les indications de temps pour les segments de paroles (les attributs *start* et *end* des balises ).

```
...
<segment start="1973.140000" end="1985.230000" bandwidth="S" speaker="S203" generator="auto">
<text generator="auto"><sil/> qui est de deux cent quatre vingt trois euros pour l' enseignement supérieur et si elle n' est pas
rattachée vous pouvez donc lui verser une une pension euh <carillon/> alimentaire <sil/> <i/> qui c' est qui est déductible
dans dans certaines limites <sil/> commentaires supplémentaires <carillon/> jusqu' <carillon/> à l' US <end/> </text>
<graph>
 ...
</graph>
</segment>
...
```

*c) Les fichiers des segmentations et regroupement en locuteurs – MDTM*

```
…
20040326_1855_2005_INTER_ELDA 1 1948.721 34.419 speaker NA male Fredericlannucci.invite.directiondelalegislationfiscale
20040326_1855_2005_INTER_ELDA 1 1983.140 2.320 speaker NA male AlainBedouet.presentateur.societe
20040326_1855_2005_INTER_ELDA 1 1983.140 2.320 speaker NA male BrunoParent.invite.directeurgeneraldesimpots
20040326_1855_2005_INTER_ELDA 1 1985.460 4.364 speaker NA male BrunoParent.invite.directeurgeneraldesimpots
20040326_1855_2005_INTER_ELDA 1 1989.824 0.429 speaker NA male AlainBedouet.presentateur.societe
20040326_1855_2005_INTER_ELDA 1 1990.253 1.431 speaker NA male AlainBedouet.presentateur.societe
20040326_1855_2005_INTER_ELDA 1 1990.253 1.431 speaker NA male BrunoParent.invite.directeurgeneraldesimpots
20040326_1855_2005_INTER_ELDA 1 1991.684 19.666 speaker NA male BrunoParent.invite.directeurgeneraldesimpots
20040326_1855_2005_INTER_ELDA 1 2011.350 22.867 speaker NA male AlainBedouet.presentateur.societe
20040326_1855_2005_INTER_ELDA 1 2034.217 24.807 speaker NA male Fredericlannucci.invite.directiondelalegislationfiscale
…
```

Ces fichiers renseignent sur la répartition des locuteurs sur les enregistrements audio. Ils fournissent les informations sur le début et la durée de chaque intervention ainsi que des informations sur le rôle du locuteur.

- Les fichiers des zones d'interactions – ZI

Les fichiers des zones d'interactions produits dans le cadre du travail de thèse de Bigot renseigne sur les caractéristiques des zones d'interactions détectées dans les contenus audio. Ils sont utilisés pour tirer des informations sur les indices de temps délimitant les échanges de tour de paroles entre locuteurs.

# 4 –Termes clés comme marqueurs des entités nommées

Dans cette partie du document, nous allons étudier en profondeur les termes clés. On rappelle l'intérêt accordé aux termes clés vient du fait que la nature même des conversations (à la radio surtout) fait que dans les échanges de paroles, des termes sont souvent utilisé par le présentateur pour introduire les personnes avec qui il va interagir.

## 4.1 – Les termes clés

Généralement, la notion de terme clé désigne, dans un document (audiovisuel ou textuel), des termes qui affichent un nombre relativement élevé d'occurrences dans le contenu. Ils permettent d'appréhender les thèmes traités dans le document. Alors qu'ici, il s'agit d'expressions typiques introduisant une information très pertinente qu'il est souvent impossible de récupérer avec la détection de termes répétitifs.

Nous considérons les termes clés (désignés aussi par *expressions clés*) comme marqueurs qui permettent de dire souvent avec certitude que les mots suivants (ou précédents, dans certains cas) contiennent une information très importante. Ils constituent des suites de mots introduisant les informations (entités nommées) qui doivent être retrouvées.

Un exemple d'expression clé illustré ici : « ***Bonjour Monsieur*** *Dupont.* ***Bienvenue*** *à…* ». Cette expression dans une interaction entre deux personnes introduit l'autre locuteur. Et ce sera très probablement le cas partout où cette expression sera détectée. Contrairement au concept de « keyphrase » (groupe de mots clés) pris dans le contexte général, les termes « Bonjour Monsieur » et « Bienvenue » peuvent ne se trouver qu'une seule fois dans une source audio et apporter une information pertinente qui échappe à la recherche de termes répétitifs : le nom du prochain locuteur.

Il est donc clair que certaines expressions jouent un rôle important dans un dialogue. Dans ce cas, si on considère la littérature sur les actes de dialogue, on peut les classer suivant leur fonction dans le dialogue entre deux ou plusieurs personnes. [Bunt2007]

## 4.2 – Les actes de dialogue

Les groupes de mots donnant un sens à un tour de parole en termes d'interaction sont désignés par le terme « acte de dialogue ». Ils sont regroupés, suivant leur fonction, en 3 grandes catégories : les actes de gestion de tours, de structuration du dialogue, de gestion des obligations sociales.

a) Gestion des tours

A la fin d'un tour de parole, un locuteur peut exprimer (par un acte de dialogue approprié) qu'il souhaite continuer à s'exprimer, ou passer la parole à un autre locuteur, ou encore mettre fin à son tour de parole tout en étant indifférent à qui doit prendre la parole par la suite. C'est ce qui est respectivement désigné par les termes : *Conservation du tour (Turn keeping), passage de tour (Turn giving) et aussi abandon du tour (turn realease).* Parmi les actes de dialogue de gestion de tour, on compte aussi les expressions par lequel un locuteur montre qu'il accepte de prendre le tour *Acceptation du tour (Turn accept).* Ceci est généralement suivi d'une *prise de tour (Turn taking).* Dans les émissions avec de fortes interactions, des locuteurs prennent parfois la parole sans que celui qui avait la parole précédemment n'ait exprimé un *Passage ou un Abandon de Tour.* Ce cas particulier de tour de parole est désigné par *Turn grapping*. Ce qui peut donner lieu à de la parole superposée.

b) Structuration du dialogue

Les actes de dialogue participant à la structuration sont les actes d'*ouverture (Opening)* et de *fermeture (Closing)*. Le premier signale que le locuteur veut bien s'engager dans une interaction avec un autre locuteur. Ce à quoi ce second locuteur peut répondre. Par exemple, deux tours constitués l'un d'une salutation initiale et l'autre d'une salutation en réaction à la première, peuvent bien être considérés comme un acte d'*ouverture*. On s'attend à retrouver deux tours au début du dialogue. Il en est de même pour l'acte de *fermeture.*

c) Gestion des obligations sociales

On parle d'obligations sociales lorsque le sens des mots du locuteur relèvent plus de la politesse que de l'expression de ses intentions. Font partie de cette catégorie d'actes de dialogue, les *excuses*, les *remerciements*, les *salutations*.

Nous nous somme donc focaliser sur les termes clés en rapport avec ces actes de dialogue.

## 4.3 – Approches de la recherche de termes clé

Nous travaillons avec des enregistrements d'émissions audio. Ceci fait que les différents locuteurs seront en général clairement introduits par le présentateur à un moment donné. Puis la détection du présentateur est très robuste, on peur restreindre la zone de recherche des termes clés sur ses interventions en se basant sur le résultat fiable de la détection du rôle du locuteur.

En suite, les tours de parole qui nous intéressent sont les interventions du présentateur où il ouvre ou ferme le dialogue. C'est pourquoi, il faut encore restreindre la zone de recherche aux extrémités des zones d'interaction.

Notre recherche d'expressions clés sur porte désormais sur cet ensemble de tour de parole des présentateurs qui se situent particulièrement en début ou à la fin d'une échange de parole avec un locuteur. L'objectif à ce stade est d'extraire des suites de mots et des motifs répétitifs qui peuvent servir de marqueurs d'informations pertinentes (à l'identification du locuteur et aussi à la structuration du document audiovisuel).

## 4.4 – Construction de la liste des termes clés

Ma première contribution consiste à construire la liste des termes clés. J'applique l'approche de la section 4.3 sur l'ensemble des données (décrites en 3.2) à ma disposition pour ce travail. Je présente un les étapes de l'application des cette approche un peu plus en détail avant de présenter une partie de la liste des termes clés que j'ai obtenue. La figure ci-dessous présente les différents processus.

*Récupération des Tours des Présentateurs* - Partant du fait que le présentateur est en général très bien détecté et que les informations sur les autres intervenants pourraient se trouver dans les tours de paroles de ceux-ci, l'étude sur les termes clés sera étendue uniquement sur les tours de parole des présentateurs. Les données contenues dans le fichier MDTM (décrit en 3.2) permettent d'extraire du fichier TRS (décrit en 3.2) correspondant les tours de parole des présentateurs.

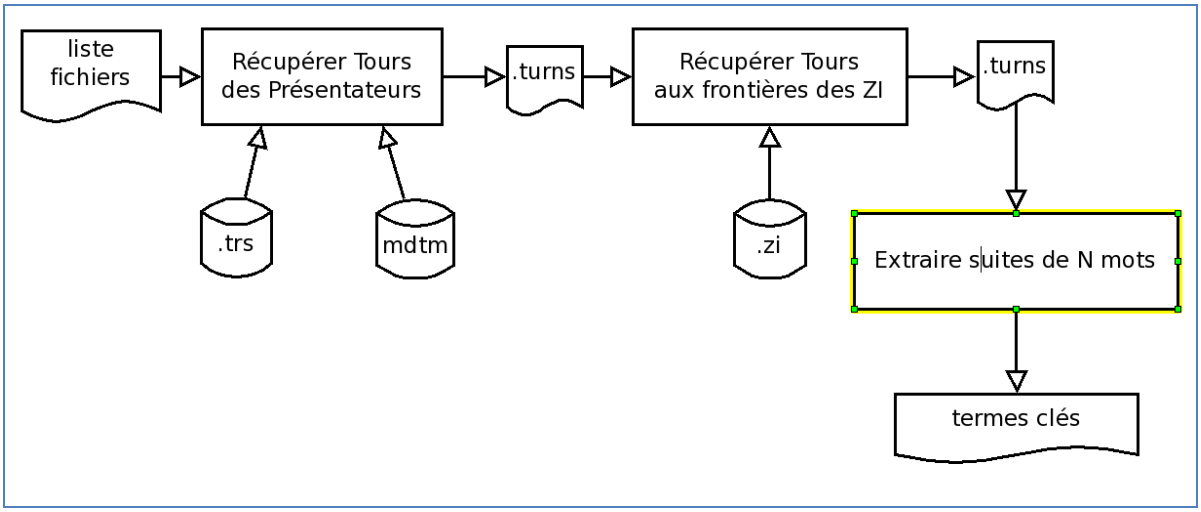


*Figure 9 - Etapes de l'extraction des termes clés*

*Récupération des Tours aux Frontières des ZI* - Les actes de dialogue d'ouverture de dialogue susceptibles de contenir les informations sur les intervenants sont les *actes d'ouverture, de fermeture, de salutation, de remerciement*. Ces actes de dialogue se trouvent en principe dans les extrémités des zones d'interaction. Le fichier ZI (décrit en 3.2) renseigne sur les indices temporels des zones d'interactions. Cette information permet donc de récupérer à partir du résultat du traitement précédent une liste de tours de parole contenant uniquement les tours de parole des présentateurs sur trouvant au début ou a la fin des zones d'échanges entre locuteurs.

*Générer Suite de Mots* – J'ai ensuite développé un outil permettant d'extraire les suite de N premiers mots pour analyser les fréquences des termes qui sont utilisés.

### *Résultats*

Je garde dans ce tableau les termes qui évoquent les ouvertures et fermetures de dialogue, et plus particulièrement ceux qui renseignent sur les noms et statuts des intervenants. (Voir l'annexe D pour la liste complète)

| **Actes dialogue** | **Expression** | **Fréquence** | **Motif** |
|---|---|---|---|
| Ouverture émission ou interaction | oui bonjour !<br>oui bonsoir ! | 6<br>50 | Oui bonjour<br>Oui bonsoir |
| | Bonjour …<br>Bonsoir | 89<br>23 | |
| | Bonjour Ali !<br>Bonjour Nicolas! | 11 | Bonjour {PRENOM} |
| Statut | vous êtes directeur de<br>vous êtes philosophe<br>spécialiste<br>vous êtes … | 4<br>2<br><br>14 | Vous êtes {FONCTION} |
| | | | |
| Fermeture | c'est la fin de | 5 | |
| Gestion tours de parole | on vous écoute(.)( !)<br>nous vous écoutons.<br>on est en ligne (avec)<br>au standard nous avons<br>… | 41<br>2<br>4<br>4 | |
| | Jérôme Jaffré<br>Roland Cayrol ?<br>… | 6<br>7<br>39 | PRENOM NOM |
| | professeur Maraninchi?<br>professeur Dubois? | 5<br>4 | Professeur {NOM} |
| | monsieur Parent?<br>… | 5<br>7 | Monsieur {NOM} |
| Prise de parole | oui tout à fait<br>… | 18<br>41 | |
| | | | |
| Cloture tour | merci à tous<br>merci à vous!<br>… | 19<br>5<br>12 | |

# 5 – Fiabilité et robustesse des termes clé

Cette partie du travail constitue ma deuxième contribution. Les mots clés comme marqueur des entités nommées sont obtenus à partir de la transcription de référence. On vient d'expliquer la constitution de cette liste (section 4.3). Cette étude consiste maintenant à étudier la robustesse de la transcription de ces termes par le traitement automatique dans le but d'évaluer leur robustesse pour la reconnaissance de la parole.

## 5.1 - Approche

La recherche d'informations pertinentes (extraction d'entités nommées) est assez robuste sur la transcription manuelle (10% d'erreurs), mais l'application de ces méthodes sur les transcriptions automatiques affiche un taux d'erreur autour de 40% avec les meilleurs systèmes [Gall2009]. Je vais étudier la robustesse de la transcription des termes clés par la transcription automatique.

*Récupérer Tours (Transcription manuelle) avec Termes clé* - Une première étape consiste à extraire du fichier de transcription manuelle tous les tours de paroles contenant le terme clé donné avec pour objectif de les soumettre à transcription automatique pour mesurer le taux de réussite de la transcription pour ce terme. Les fichiers audio ne sont pas entièrement transcrits manuellement. Les informations sur l'étendu de la transcription manuelle sont dont récupérées pour faciliter l'alignement avec la transcription automatique.

<syn_times **min=923.543 max=6443.043**>
fichier=... terme_cle=... debut=**923.543** fin=**928.453** phrase=... **bonjour** Alain Nicolas !
...
...
fichier=... terme_cle=... debut=**6438.953** fin=**6443.043** phrase=invité des Matins de France Culture aujourd'hui euh Pierre Guyotat . **bonjour** !

*Récupérer Tours (Transcription automatique) avec Termes clé* - Il devient possible d'extraire tous les segments d'un fichier transauto.xml contenant des termes clés. Il s'agit d'une opération similaire à celle consistant à extraire les tours contenant les termes clé de la transcription manuelle, sauf que pour ces fichiers on extrait des segments.

*Mesurer le taux de bonne reconnaissance de la transcription des termes clé* – Le découpage en tour de parole dans la transcription manuelle, ne correspond pas au découpage en segments de parole par le transcription automatique. Les indications temporelles sur les tours et les segments n'étant pas forcément les mêmes, un travail d'alignement s'avère alors nécessaire pour comparer les transcriptions manuelles aux transcriptions automatiques. Cet alignement consiste à relier tous les tours et les segments qui partagent un temps de parole. Un tour est lié à un segment parce que le début de ce tour se trouve dans le segment ou la fin de ce tour est dans le segment. Un segment est lié à un tour suivant que le début du segment se trouve dans le tour ou que la fin du segment se trouve dans le tour. Un exemple d'un fichier retour de ce travail d'alignement permet de mieux illustrer.

```
nbr_turns=28 nbr_sgmts=24 turn=1 segment=1 cas=ft_s mot_cle=bonsoir nbre_occurences_t=1 nbre_occurences_s=1
nbr_turns=28 nbr_sgmts=24 turn=1 segment=1 cas=ds_t mot_cle=bonsoir nbre_occurences_t=1 nbre_occurences_s=1
nbr_turns=28 nbr_sgmts=24 turn=2 segment=1 cas=dt_s mot_cle=bonsoir nbre_occurences_t=1 nbre_occurences_s=1
nbr_turns=28 nbr_sgmts=24 turn=2 segment=1 cas=ft_s mot_cle=bonsoir nbre_occurences_t=1 nbre_occurences_s=1
…
…
nbr_turns=28 nbr_sgmts=24 turn=26 segment=23 cas=dt_s mot_cle=bonsoir nbre_occurences_t=1 nbre_occurences_s=1
nbr_turns=28 nbr_sgmts=24 turn=26 segment=23 cas=ft_s mot_cle=bonsoir nbre_occurences_t=1 nbre_occurences_s=1
nbr_turns=28 nbr_sgmts=24 turn=27 segment=24 cas=dt_s mot_cle=bonsoir nbre_occurences_t=1 nbre_occurences_s=3
nbr_turns=28 nbr_sgmts=24 turn=27 segment=24 cas=ft_s mot_cle=bonsoir nbre_occurences_t=1 nbre_occurences_s=3
nbr_turns=28 nbr_sgmts=24 turn=28 segment=24 cas=dt_s mot_cle=bonsoir nbre_occurences_t=2 nbre_occurences_s=3
nbr_turns=28 nbr_sgmts=24 turn=28 segment=24 cas=ft_s mot_cle=bonsoir nbre_occurences_t=2 nbre_occurences_s=3
```

Pour déterminer le nombre de transcriptions automatiques du terme clé à considérer comme correctes, absentes ou insérées, je fais correspondre des groupes de tours à des groupes de segments.

Le principe pour constitué un groupe minimal de tours étant que : un tour $T_{i+1}$ appartient au même groupe que $T_i$ s'il est lié à un des segments liés à $T_i$ et que si $T_j$ n'appartient pas à un groupe, quelque soit k>j, $T_k$ ne peut appartenir à ce groupe.

Un principe similaire est utilisé pour constituer les groupes de segments. Un segment $S_{i+1}$ appartient au même groupe que $S_i$ s'il est lié à un des tours liés à $S_i$ et que si $S_j$ n'appartient pas à un groupe, quelque soit k>j, $T_k$ ne peut appartenir à ce groupe.

Le dessin ci-dessous illustre le rapport entre les liens des tours et segments et la notions de groupes de tours et de groupes de segments.

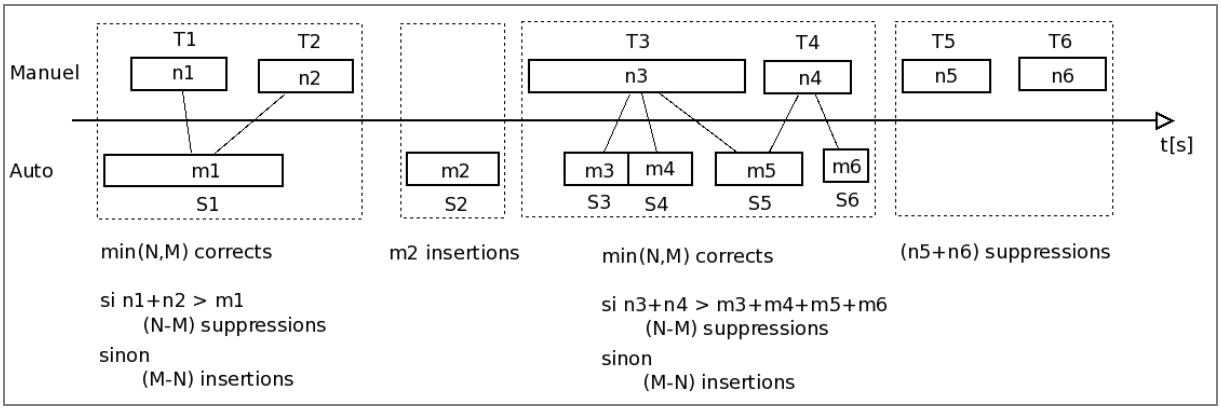


*Figure 10 – Alignement des tours et segments – Les groupes en pointillés.*

Cette façon de procéder fait que un groupe de tours correspond à au plus un groupe de segments. Et que ce groupe de tour partage la même période de temps délimitée par le minimum des temps du premier tour et du premier segment de ces groupes et le maximum des temps des derniers tours et segments de ces groupes. Pour ce qui nous intéresse ici, le nombre total (N) de termes clé retrouver dans ces tours devrait être le même que le nombre total (M) de termes clé retrouvés dans les segments du groupe correspondant.

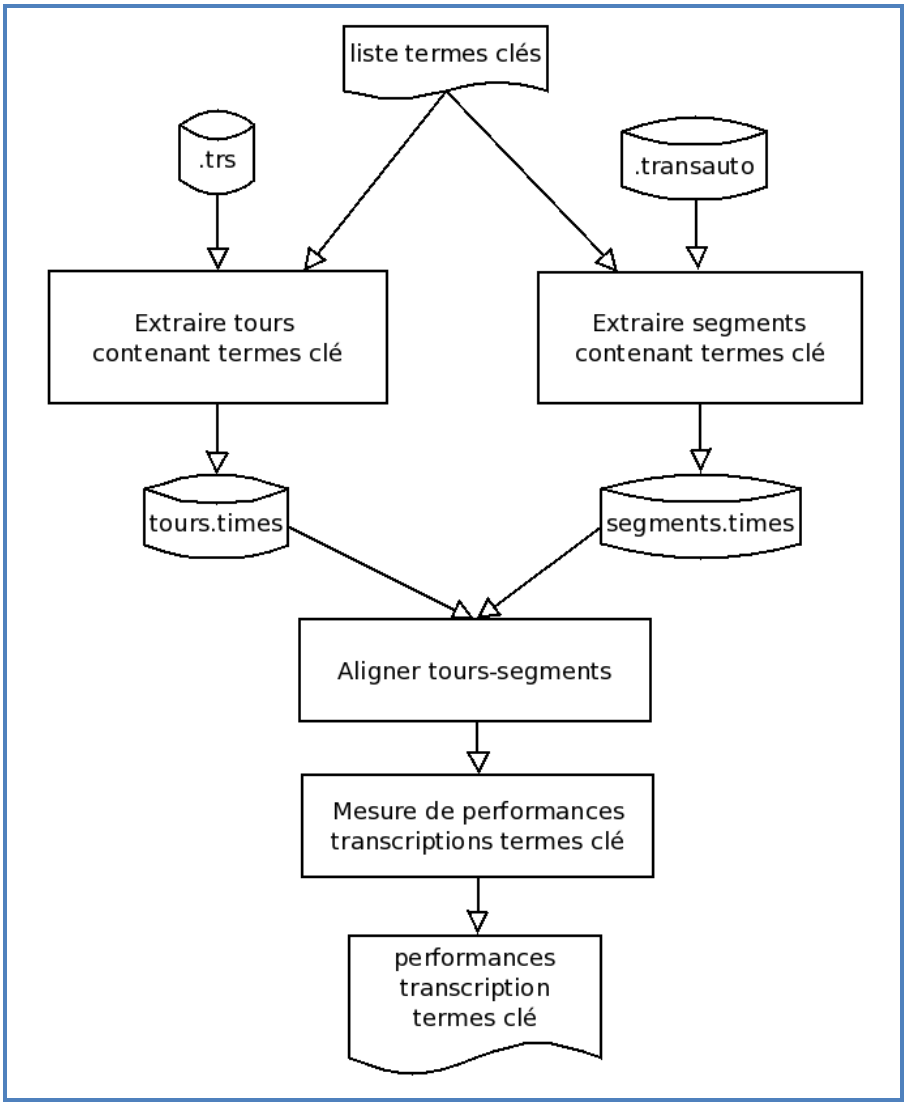


*Figure 11 : Mesure du taux de reconnaissance des termes clés*

### *Résultats*

***Par type de fichier :***

| Type de fichier | INTER | CULTURE | RFI / INFO |
|---|---|---|---|
| Nbre total de termes clés | 780 | 397 | 292 |
| Correctement retrouvés | 724 | 372 | 266 |
| Suppressions | 56 | 25 | 26 |
| *Insertions* | 5 | 31 | 5 |

***Par terme clé :***

*BONJOUR*

| | | |
|---|---|---|
| Nbre total d'occurrences | 343 | |
| Correctement retrouvés | 309 | *90 ,1%* |
| Suppressions | 34 | |
| *Insertions* | *31* | |

*BONSOIR*

| | | |
|---|---|---|
| Nbre total d'occurrences | 503 | |
| Correctement retrouvés | 452 | *89 ,9%* |
| Suppressions | 51 | |
| *Insertions* | *6* | |

*MERCI*

| | | |
|---|---|---|
| Nbre total d'occurrences | 457 | |
| Correctement retrouvés | 436 | *95,4%* |
| Suppressions | 21 | |
| *Insertions* | *4* | |

*MERCI A TOUS*

| | | |
|---|---|---|
| Nbre total d'occurrences | 12 | |
| Correctement retrouvés | 12 | *100%* |
| Suppressions | 0 | |
| *Insertions* | *0* | |

*ON VOUS ECOUTE*

| | | |
|---|---|---|
| Nbre total d'occurrences | 36 | |
| Correctement retrouvés | 36 | *100%* |
| Suppressions | 0 | |
| *Insertions* | *0* | |

*NOUS VOUS ECOUTONS*

| | | |
|---|---|---|
| Nbre total d'occurrences | 19 | |
| Correctement retrouvés | 19 | *100%* |
| Suppressions | 0 | |
| *Insertions* | *0* | |

*PROFESSEUR*

| | | |
|---|---|---|
| Nbre total d'occurrences | 111 | |
| Correctement retrouvés | 110 | *99,1%* |
| Suppressions | 1 | |
| *Insertions* | *0* | |

***Résultat global***

| | | |
|---|---|---|
| Nbre total de termes clés | 1469 | |
| Correctement retrouvés | 1362 | *92 ,7 %* |
| Suppressions | 107 | |
| *Insertions* | *41* | |

### *Amélioration possible*

Les objectifs dans cette partie du travail consistent d'une part à voir s'il est possible, pour récupérer les erreurs avec la reconnaissance automatique, de détecter les termes clés du flux audio à partir du décodage acoustico-phonétique en enrichissant le dictionnaire avec les différentes suites phonétiques possible d'un mot, et de l'autre part, de voir comment il est possible de détecter les termes clés non obtenus avec le système de reconnaissance automatique.

*Extraire WAV Termes clé* – Cette opération extrait de l'ensemble des fichiers audio (WAV) les tours de parole contenant les termes clé retenus. Elle se base sur les informations contenues dans les fichiers (TRS) pour récupérer les temps pour extrait de la source audio la partie contenant uniquement le tour de parole contenant le terme clé.

*Transcrire Termes clés* – Cette opération consiste à transcrire les fichiers WAV extraits. L'objectif étant de pouvoir faire une analyse plus poussées (au niveau phonétique) sur la nature des erreurs rencontrées au niveau de la transcription automatique, on produit pour chacun de ces extraits la suite phonétique correspondant et le treillis (lattice).

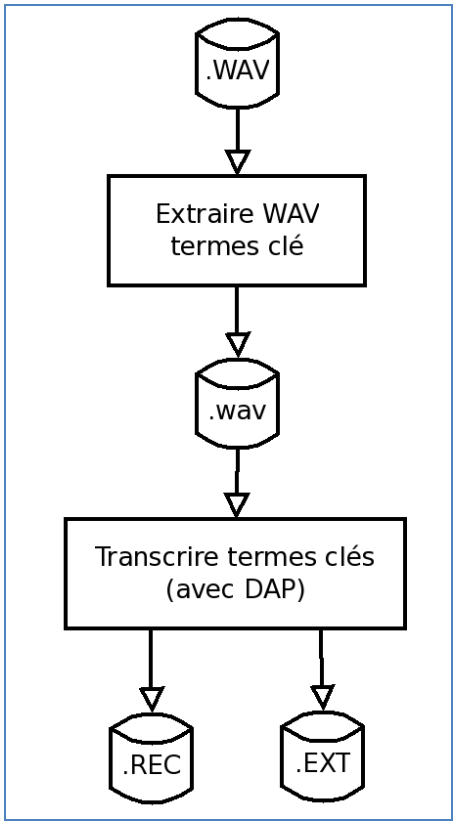


*Figure 12 : Génération suite phonétique et treillis de phonèmes.*

***Transcription phonétique obtenu avec les modèles phonétique HREST.***

Un extrait du fichier : 20040217_1855_2005_INTER_ELDA (plage 654s – 663s).

```
SENT-START pause y n y v j e9 e cl p a v l U~ cl p w s j o~ vcl d e2 y
vcl g U~ S i sil pause e cl k y cl t e v a l E v e v w cl p w e v w E
l e cl p e l y v y cl k @ m H i cl t e9 O f R O N cl p e z E U~ l y N
cl t e9 R a~ U~ m @ cl p pause E e cl p a m E R a~ R a vcl p o~ n a~ w
a cl t i vcl b o v e z o~ n e9 v vcl d @ f R a~ z v @ l J R u vcl p
pause sil w a f H O l m e2 @ H E y H e2 SENT-END
```

Phrase prononcée : Une véritable impression de gâchis. **Ecoutez** Valérie Rodrigo elle est élue du comité d'entreprise d'Air Littoral. Elle est amère en répondant à Thibault Maisonneuve de France Bleu Hérault.

Transcription par l'outil du LIUM : une véritable impression de gâchis <sil/> <i/> **et coûter** Valérie Rodrigo elle est élue du comité d' entreprise d' Air Littoral <sil/> <i/> <sil/> est amer en répondant à Thibault Maisonneuve de France Bleu Hérault

Un autre exemple de ce type d'erreurs est l'exemple avec « et c'est » transcrit « essaie », … Il est facile de récupérer les termes clés substitués par des homophones.

Par contre, les autres types d'erreurs sont difficilement récupérables uniquement avec le dictionnaire. Il est parfois nécessaire d'analyser les treillis des phonèmes quand la suite des phonèmes de meilleur score ne permet pas de détecter le terme prononcé. Une illustration est ce graphe de phonème permettant de voir clairement qu'il existe un chemin permettant de récupérer le terme « bonsoir » par exemple. La figure 13 illustre ce cas.

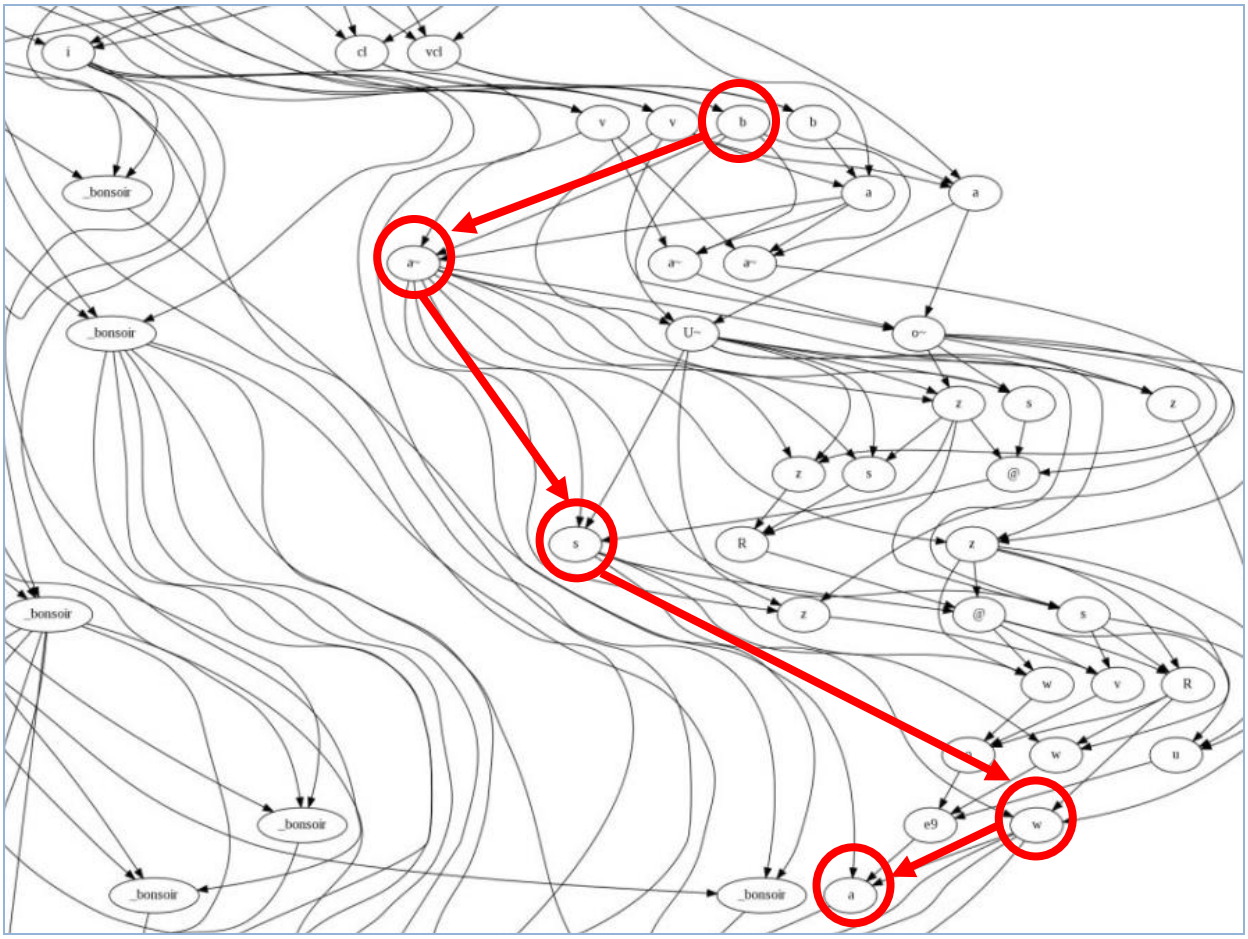


*Figure 13 : Exemple de graphe phonétique (bonsoir)*

Par contre, les autres types d'erreurs sont difficilement récupérables. Ci-dessous, un exemple de transcription phonétique d'un extrait d'une émission Téléphone pour illustrer ce type d'erreurs. On remarque que la suite de phonèmes constituant le mot « professeur » ne se trouve pas dans la transcription.

```
SENT-START pause sil sil e2 l e9 cl p w s e9 E vcl d i v a~ w l a vcl
d i s e cl p a s U~ e2 N cl p o~ N cl p a R l e2 s e v e a i cl t a
vcl b l e9 m a~ pause pause U~ n E vcl d z j a s i s f H z i N sil cl
k O a vcl t i vcl d i j U~ s i N cl p O z s i N vcl b l e9 e2 pause a
R e cl p e cl t e N vcl p a f O R s e m a~ e2 cl p pause e2 N cl t R a
e f O m H J a~ cl t H N cl k a sil pause f sil m N cl p R a cl t i v
vcl d i z a~ R e H y l j E U~ sil u e9 a vcl d i v O R a~ SENT-END
```

**Phrase prononcée :**

```
<Sync time="1929.987"/>
euh le professeur Guy-Grand l'a dit c'est pas de ça euh qu'on parle
c'est véritablement un exercice répété pas forcément euh euh très fort
mais en tout cas une pratique disons régulière
<Turn speaker="spk3 spk5" mode="spontaneous" fidelity="high" channel="studio" startTime="1944.055" end-
Time="1944.814">
<Who nb="1"/>
professeur Guy-Grand
<Who nb="2"/>
euh
</Turn>
```

Transcription par l'outil du LIUM

```
<i/> euh le prince a dit dans la vie c' est pas de ça euh qu' on parle
c' est véritablement <sil/> un exercice physique <sil/> quotidien si
possible <sil/> <sil/> répété pas forcément euh <sil/> <i/> euh très
fort mais en tout cas <sil/> une pratique disant régulière <sil/> l'
un des grands
```

Compte tenu du fait que le système de décodage acoustico-phonétique propose en fait pour chaque trame, le modèle acoustique le plus probable parmi les modèles appris, en fonction éventuellement aussi d'une grammaire, on peut penser que les phonèmes qui permet des retrouver le bon mot pourraient ne pas tous être les phonèmes les plus probables. Bien entendu, le système de reconnaissance automatique de la parole propose de toute façon un mot. Ce mot est parfois très éloigné du mot prononcé effectivement (exemple : dans l'expérience actuelle on a « *le prince* », ou encore, « *l'un des grand* » pour l'expression « *professeur Guy-Grand* »). On a donc pensé à consulter le treillis de phonème de ces extraits où la suite phonétique proposée ne permettait pas de récupérer le mot pour étudier la possibilité de récupérer ces mots-clés en parcourant dans le bon ordre les différents nœuds.

La figure 14 permet de se rendre compte de la complexité de l'analyse des graphes phonétiques générée. Les techniques développées par Olivier Le Blouch sur l'utilisation du décodage acoustico-phonétique dans l'indexation pourrait permettre de récupérer les termes clés dans ces cas [Blou2009]. La durée limitée de mon travail ne m'a pas permis de bien exploiter cette alternative.

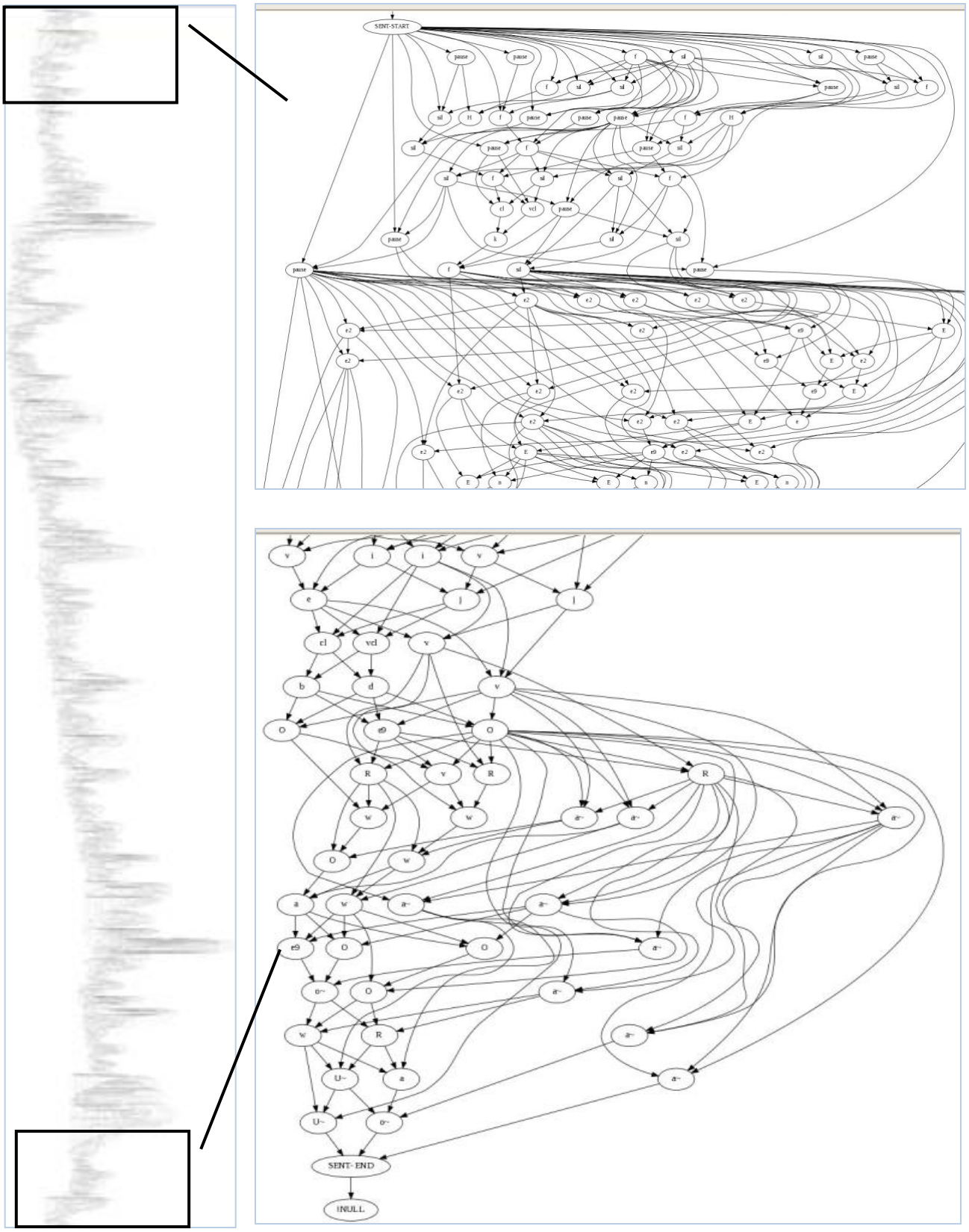


*Figure 14 : Complexité du parcours des nœuds pour une phrase de 15s*

### 5.2 - Interprétation des résultats

*Un bon taux de reconnaissance pour termes composés de plusieurs mots et les « statuts » (professeur, ...)*

*Un taux de suppression pour certains termes clé « d'ouvertures de dialogue » relativement bas.*

Pour les termes clé *Bonsoir* et *Bonjour*, le système de reconnaissance de la parole utilisé présente un taux de suppression d'environ 10%. Ces termes représentent obligations sociales, mais aussi sont parfois des actes d'ouverture de dialogue. Et quand les deux salutations se trouvent pratiquement dans la même période de temps, effectivement on se retrouve dans le cas de parole superposée. La transcription automatique, quand le signal de parole présente ce cas de figure ne transcrit pas les phrases prononcées ou dans le meilleur des cas, la phrase d'un seul des locuteurs.

# 6 – Conclusions et Perspectives

*Conclusion*

Ces travaux m'ont permis de mieux connaître l'environnement de la recherche et d'acquérir des connaissances plus poussées sur le domaine du traitement automatique de la parole. Ils m'ont amené à manipuler et à maîtriser des outils très utilisés dans des domaines liés au traitement automatique de la parole : HTK, Transcriber, Wavesurfer et d'autres méthodes de traitement. Les résultats obtenus, par rapport aux objectifs fixés, sont concluants même si la durée du stage ne permettait pas d'étudier en profondeur un aspect du projet : la détection des expressions clés par les treillis de phonèmes.

L'étude menée permet de dire que certains termes clés sont très robustes par la reconnaissance automatique de la parole. Ils constituent des marqueurs fiables pour la détection des entités nommées et plus particulièrement des informations pertinentes permettant d'affiner les classes de locuteurs assez génériques : *Journaliste* et *Autre*. En cas d'erreurs dans la transcription des termes clés, une voie possible passe par l'exploitation du résultat du décodage acoustico-phonétique. Ceci permet alors de détecter les zones susceptibles de contenir les informations pertinentes : noms des intervenants, leur statut, leur rôle dans l'émission.

Une telle méthode, appliquée à la détection du rôle, permet de raffiner encore plus les classes considérées généralement. Il permettre par exemple d'avoir des sous-classes non seulement suivant leur fonction dans le programme : journaliste chroniqueur, journaliste envoyé spécial, journaliste correspondant ; mais aussi suivant que le locuteur est sur le plateau ou à distance.

Dans cette étude, nous avons mis l'emphase sur une liste de termes clés permettant de localiser les informations plus fines sur rôles et d'avoir les noms des locuteurs. Mais il est tout à fait envisageable de détecter un jeu de termes clés marqueurs d'informations renseignant par exemple sur le nom du programmes, le nom des thèmes développés, des ouvertures et clôtures des thèmes. Ceci peut contribuer à une structuration du document audiovisuel.

La localisation au niveau des extrémités des zone d'interaction des ces actes de dialogue permettant de délimiter les informations sur les locuteurs et récupérer des informations sur la structure du document rend possible la réduction de la zone d'analyse.

*Perspectives*

Nous nous sommes concentrés dans cette étude sur une liste de termes clés pouvant servir de marqueurs pour les informations permettant de raffiner les rôles et d'avoir les noms des locuteurs. La localisation au niveau des extrémités des zone d'interaction des ces actes de dialogue permettant de délimiter les informations sur les locuteurs et récupérer des informations sur la structure du document rend possible la réduction de la zone d'analyse. Mais il est tout à fait envisageable de détecter un jeu de termes clés marqueurs d'informations renseignant par exemple sur le nom du programmes, le nom des thèmes développés, des ouvertures et clôtures des thèmes. Toutes les catégories d'entités nommées pourraient être utilisées. Il serait également possible de récupérer dans la partie vidéo de l'enregistrement, les informations provenant de la détection de visages, de costumes, des sous-titres affichés à l'écran afin de compléter les informations disponibles. Ceci peut contribuer à une structuration du document audiovisuel.

Cette étude est menée sur un corpus constitué essentiellement d'émissions en Français. La façon de procéder ne tenant pas compte de la langue. Il est tout à fait possible d'utiliser une telle méthode tout enregistrement audiovisuel quelle que soit la langue. Il suffit pour cela de chercher à détecter des termes clés qui lui sont propres. En effet, toutes les langues disposent d'expressions clés comme actes de dialogue.

Ces études sur les expressions qui caractérisent les interventions peuvent être utilisées avec d'autres indicateurs pas obligatoirement textuels dans des recherches des intentions du locuteur. L'analyse de la prosodie locale pourrait apporter des informations sur le façon dont les phrases sont dites : le rythme, l'intonation, les emphases. Ceci permettrait de caractériser le discours (s'agit-il de l'expression d'un avis ?, de la manifestation d'une contradiction ? …) et de pouvoir localiser les passages les plus porteur de sens dans les émissions afin d'aider à l'émergence de zones intéressantes voire la construction de résumés automatiques.

# Annexe A

## *Une émission radiophonique interactive*

*Zones d'interaction, Actes de dialogues, Termes clés, Entités nommées*

**Les locuteurs**
L1 - [Pierre Weill] -- présentateur--société
L2 - [Marina] -- auditeur -- auditeur
L3 - [Patrick Tounian] -- invite -- nutritionniste et pédiatre
L4 - [Alain Bazot]
L5 - [Dominique Turk]
L6 - [Gilles-Vincent]
L7 - [Jacques]
L8 - [Maryline]
L9 - [René]
L10 - [Philippe]

**Tours de paroles / Interactions**

ZI 1 (L1 + L2)
L1 : ***Et c'est*** **Marina** qui nous appelle, elle se trouve en Seine-et-Marne. ***Bonsoir*** **Marina** !
L2 : Oui bonsoir
---
ZI 2 (L1 + L3)
L1 : très bien ben ***réponse du*** *professeur* **Patrick Tounian** ; je rappelle que vous êtes euh nutritionniste et pédiatre à l'hôpital euh Armand Trousseau à Paris.
L3 : bonsoir euh
---
ZI 3 (L1 + L4)
L1 : mais ça on aura l'occasion je pense d'en parler () d'en reparler durant l'émission , les problèmes de de prévention et d'éducation . (Fin Z2)
L1 : ***Intervention aussi d'***'**Alain Bazot** , président de l'union fédérale des consommateurs "que choisir" , qui se trouve euh dans les studios de France Bleu Orléans . bonsoir !
L4 : Oui, bonsoir ! Simplement je je voulais également rappeler des chiffres : c'est vrai que …
**---------**
ZI 4 (L1 + L3)

L1 : *Professeur* **Patrick Tounian** euh vous aussi, vous estimez qu'il faudrait carrément interdire la publicité à la radio et à la télévision euh

L3 : Qu'il faille réglementer la publicité destinée aux enfants, comme monsieur le disait notamment

**---------**

ZI 5 (L1 + L5)
L1 : ***Réaction aussi de*** **Dominique Turk**, dans les studios de France Bleu Lille Nord ; vous êtes professeur, je rappelle, de pédiatrie à la faculté de médecine de Lille, et et président du comité nutrition humaine de l'Agence française de sécurité sanitaire des aliments . C'est long hein tout ça!
L5 : Oui c'est un peu compliqué …
---
ZI 6
L1 : ***Au standard nous avons*** **Gilles** qui est en ligne **Gilles-Vincent** , il est à Orléans . Bonsoir !
L6 : Oui bonsoir à vous , merci pour l'émission. …
---
ZI 7
L1 : Réponse euh bah *le professeur* **Patrick Tounian** ***va vous répondre***.
L3 : euh uniquement une cause génétique non
---
ZI 11 -
L1 : ***on va au standard et c'est*** **Jacques** qui est en ligne, il est à Sceaux . bonsoir !
L7 : oui bonsoir ! …
---
ZI 12 -
L1 : eh bien écoutez *on* ***on va demander à*** *à* **Alain Bazot** , qui est président de l'union fédérale des consommateurs "que choisir" euh de vous répondre (TRANSITION ZI 11, ZI12)
L4 : bah vous savez si on était en mesure de de d'empêcher toutes les aberrations euh ce ce serait assez formidable.
---
ZI 13 -
L1 : ***commentaire aussi du*** *professeur* **Patrick Tounian**.
L3 : alors je voudrais souligner tout ce qui vient d'être dit…
---
ZI 14
L1 : ***en ligne*** **Maryline** , elle est à Angoulême . bonsoir !
L8 : bonsoir ! euh voilà Marilyn , donc je suis diététicienne …
---
ZI 16
L1 : ***commentaire aussi de*** **Dominique Turk** , professeur euh de pédiatrie à la faculté de médecine de Lille .
L5 : oui je voulais euh revenir sur la question de l'auditrice sur la restauration scolaire.
---
ZI 18
L1 : ***En ligne de*** *Villeneuve-le-Roi*, **René** ***nous appelle***. vous êtes le bienvenu , bonsoir !
L9 : Oui bonsoir euh j'aurais aimé si () savoir si en France , on a les moyens de lutter contre le lobby de tous les marchands de boissons sucrées , qu'elles soient boissons ou qu'elles soient euh…
---
ZI 21
L1 : **Philippe** ***est en ligne*** , ***bonsoir*** **Philippe** !
L10 : bon() oui bonsoir , merci de prendre mon appel …

---
DERNIRE INTERVENTION -
L1 : eh bien voilà , c'est le mot de la fin : il faut penser au plaisir , mais il faut le contrôler . ***bonsoir , à demain*** !
Speakerine [Journaliste annonceur] : c'était "le téléphone sonne", une émission dirigée par **Pierre Weill**.

# Annexe B

## *Les variations dans la prononciation des mots*

Cette partie de document traite des différentes prononciations des mots à prendre en compte pour mette en place un dictionnaire capable de les détecter. Elle sera consacrée plus particulièrement aux phénomènes de coarticulation mais présentera très brièvement les différents types d'erreurs à considérer dans les systèmes de reconnaissance vocale surtout que, avec l'approche phonétique, il vaut mieux avoir un dictionnaire capable de prendre en compte différentes formes d'un mot avec les erreurs de prononciation, les coarticulations, les liaisons ou toute autre variation connue (les malentendus de prononciations, les erreurs de liaisons, …).

Les différents types de variations :

*d) Variations liées au contexte pour un même mot.*

Un bon exemple pour l'illustrer est le mot *« fils »*. Le contexte sémantique permet de résoudre la prononciation de ce mot.

*e) Variations liées au locuteur*

Un exemple typique est la prononciation du mot « Lait ». Des locuteurs d'une région prononceront « L AI » tandis que d'autre diront « L EI ».

*f) Variations dues à la liaison*

Il est souvent nécessaire, quoique certaines liaisons soient généralement presque toujours respectées, de prendre en compte le fait que parfois (et c'est imprévisible) elles ne le sont pas.

*g) Variations dues à la coarticulation*

Il s'agit d'une déformation (involontaire) dans la prononciation d'un mot en fonction de ceux qui l'entourent. Ou plus généralement une médication dans la prononciation d'un phonème suivant les phonèmes autour de lui et de l'ampleur du changement que la bonne prononciation exigerait au niveau des caractéristiques articulatoires (mouvements des lèvres, distances des cordes vocales, etc…). Ce phénomène est vu par Martinet comme « l'économie des changements linguistique », un compromis entre l'inertie des organes phonateurs à la discrimination des sons et la compréhension du message.

Les phénomènes de coarticulation se présentent sous différentes formes :

h) L'assimilation
i) La dilation
j) La différentiation
k) La dissimilation
l) L'interversion : Permutation de phonèmes contigus.
m) La métathèse : Permutation de phonèmes voisins mais non contigus.
n) L'épenthèse : Un phonème de trop est prononcé.
o) L'élision (ou syncope) : Un phonème n'est pas prononcé

| | ***Phénomène*** | ***Exemple*** |
|---|---|---|
| **Phénomènes d'assimilation** | Assimilation par voisement (régressive) | p) Cap-Vert : K A **P** V AI R → K A **B** V AI R<br>q) Paquebot : P A **K** B O → P A **G** B O<br>r) Israël : I **S** R A AI L → I **Z** R A AI L<br>s) Chef de : CH AI **F** D EU → CH AI **V** D EU |
| | Assimilation par dévoisement (régressive) | t) Absent : A **B** S AN → A **P** S AN<br>u) Je crois : **J** K R W A → **CH** K R W A |
| | Assimilation par nasalisation | v) Pentecôte : P AN **T** K O T → P AN **N** K O T<br>w) Vingt-deux : V IN **T** D EU → V IN **N** D EU |
| | Dilation (sons éloignés) | x) Surtout : S **U** R T OU → S **OU** R T OU<br>y) Disséminer : D I S **EI** M I N EI → D I S **I** M I N EI |
| **Phénomène de différenciation** | Différenciation (sons en contact) | z) Dehors : D **EU** O R → D **EI** O R |
| | Dissimilation | aa) venimeux : V **EU N** I M EU → V **L** I M EU |
| **Changement de l'ordre des sons** | Interversion (sons en contact) | bb) Aéroport : A **EI R** O P O R → A **R EI** O P O R T |
| | Métathèse (sons éloignés) | cc) Séchoir : **S** EI **CH** W A R → **CH** EI **S** W A R |
| **Insertion d'un son** | Epenthèse | dd) Ours polaire : OU R S P O L AI R → OU R S **EU** P O L AI R<br>ee) Arc-boutant : A R K B OU T AN → A R K **EU** B OU T AN |
| **Effacement d'un son** | Elision / Syncope | ff) Deux seconds : D EU **S EU** G ON D → D EU **Z** G ON D<br>gg) Question : K AI S **T** Y ON → K AI S Y ON |
| | Divers, erreurs fréquemment rencontrées | hh) Dommage : D **O** M A J → D **EU** M A J<br>ii) National : N A S Y **O** N A L → N A S Y **EU** N A L |

# Annexe C

## *L'alphabet phonétique international dans le système*

| | Alphabet international | Alphabet utilisé à l'IRIT |
|---|---|---|
| 1 | Ø | e2 |
| 2 | Œ | e9 |
| 3 | ə | @ |
| 4 | ɛ | E |
| 5 | ɥ | H |
| 6 | ɴ | N |
| 7 | ɲ | J |
| 8 | ɔ | O |
| 9 | R | R |
| 10 | ʃ | S |
| 11 | ɛ̃ | U~ |
| 12 | ʒ | Z |
| 13 | A | a |
| 14 | ɑ̃ | a~ |
| 15 | B | b |
| 16 | d | d |
| 17 | e | e |
| 18 | f | f |
| 19 | g | g |
| 20 | i | i |
| 21 | j | j |
| 22 | k | k |
| 23 | l | l |
| 24 | m | m |
| 25 | n | n |
| 26 | o | o |
| 27 | | o~ |
| 28 | p | p |
| 29 | s | s |
| 30 | t | t |
| 31 | u | u |
| 32 | v | v |
| 33 | w | w |
| 34 | y | y |
| 35 | z | z |

# Annexe D

## *Liste des termes clés obtenus*

Ce tableau présente les termes obtenus après les traitements pour la recherche des termes clés pouvant servir de marqueur d'informations pertinentes sur les intervenants d'une émission.

| **Actes dialogue** | **Expression** | **Fréquence** | **Débat** | **Motif** |
|---|---|---|---|---|
| Ouverture émission ou interaction | oui bonjour (!)<br>oui bonsoir ! (.) (…) | 6<br>50 | | Oui bonjour<br>Oui bonsoir |
| | Bonjour (!) (bonjour !)<br>Bonsoir ( !)<br>Bonjour ! vous …<br>Bonjour ! on<br>Bonjour ! je<br>Bonjour ! alors …<br>Bonsoir ! (bonsoir !)<br>(Bon)soir<br>Bonsoir ! oui moi je … | 37<br>23<br>22<br>4<br>3<br>11<br>8<br>2<br>2 | | |
| | Bonjour Ali !<br>Bonjour Nicolas! | 7<br>4 | | Bonjour {PRENOM} |
| | Nicolas Demorand bonjour à … | 5 | | |
| | bonjour et bienvenue dans<br>bonjour et bienvenue à | 2<br>2 | | |
| Annonce radio | vous écoutez RFI il<br>vous écoutez France Culture | 2<br>7 | | Vous écoutez {RA-DIO\|EMISSION} |
| | il est neuf heures. | 4 | | Il est {HEURE} heures. |
| | | | | |
| Statut | vous êtes directeur de<br>vous êtes philosophe spécialiste<br>vous êtes sociologue vous<br>vous êtes secrétaire général<br>vous êtes rédacteur en<br>vous êtes journaliste à<br>vous êtes chercheur au<br>donc vous êtes un … | 4<br>2<br><br>2<br>2<br>2<br>2<br>2<br>4 | | Vous êtes {FONCTION} |
| | | | | |
| | | | | |
| Fermeture | c'est la fin de | 5 | | |

| | | | | |
|---|---|---|---|---|
| Gestion tours de parole | on vous écoute(.)( !)<br>nous vous écoutons.<br>on est en ligne (avec …)<br>au standard nous avons<br>Dominique Turk veut intervenir | 41<br>2<br>4<br>4<br>2 | TS<br>TS<br>TS<br>TS | |
| | Jérôme Jaffré<br>Roland Cayrol ( ?)<br>Christophe Barbier<br>Hélène Jouan?<br>Gaëtan Gorce?<br>Brice Hortefeux?<br>Hervé de la Martinière<br>David Rodrigues?<br>Valérie Chauvin?<br>Pascal Champvert?<br>Frédéric Iannucci?<br>Hugues Portelli<br>Hosham Daoud?<br>Frédéric Iannucci?<br>Faustine | 6<br>7<br>5<br>4<br>4<br>4<br>3<br>3<br>2<br>2<br>2<br>4<br>2<br>2<br>2 | | PRENOM NOM |
| | professeur Maraninchi?<br>professeur Dubois? | 5<br>4 | | Professeur {NOM} |
| | monsieur Parent?<br>monsieur Guy-Grand<br>monsieur Lebreton ?<br>monsieur Dubois? | 5<br>3<br>2<br>2 | | Monsieur {NOM} |
| Prise de parole | oui tout à fait (.)<br>oui oui tout à (fait)<br>ah tout à fait<br>oui oui(.) (oui \| oui oui)<br>oui mais | 18<br>4<br>2<br>27<br>8 | <br>TS<br><br>TS<br>TS | |
| | | | | |
| | | | | |
| Cloture tour | merci à tous<br>merci à vous!<br>merci à tous les (deux)<br>merci pour ce témoignage(.)<br>à bientôt!<br>au revoir ! | 15<br>5<br>4<br>4<br>2<br>2 | <br><br>TS<br>TS | |
| Transition | à tout à l'heure<br>à tout de suite | 6<br>6 | | |
| | | | | |
| | | | | |
| Convention sociale | je vous en prie. | 4 | | |

# Bibliographie


[Herl2006] C. Herley. ARGOS : Automatically extracting repeating objects from multimedia streams. *IEEE Transactions on Multimedia,* Feb.2006

[Musc2008] A. Muscariello, G. Gravier, and F. Bimbot. Audio keyword extraction by unsupervised word discovery. In *Proc. Interspeech,* 2008

[Musc2010] A. Muscariello, G. Gravier, and F. Bimbot. *Découverte non supervisée de mot(if)s dans le signal de parole*. Mai 2010

[Grav2004] G. Gravier, J.F. Bonastre, S. Galliano, E. Geoffrois, K. Mc Tait and K. Choukri. *ESTER, une campagne d'évaluation des systèmes d'indexation d'émissions radiophoniques*, Proc. Journées d'Etude sur la Parole, Avril 2004.

[Bigot2011] B. Bigot, I. Ferrané, J. Pinquier, R. André-Obrecht. *Speaker Role Recognition to help Spontaneous Conversational Speech Detection (regular paper).* Dans : *International workshop on Searching Spontaneous Conversational Speech SCSS (SCSS 2010), Firenze, Italy,25/10/2010-29/10/2010,* ACM, p. 5-10, octobre 2010.

[Gall2005] E. Galliano et al., "The ESTER Phase II Evaluation Campain for the Rich Transcription of French Broadcast News", dans *Proceedings of the 9th European Conférence on Speech Communication and Technology,* Interspeech 2005, Lisbonne (Portugal).

[Gall2009] E. Galliano et al., "The ESTER 2 Evaluation Campain for the Rich Transcription of French Broadcast News", dans Interspeech 2009, Royaume-Uni.

[Adda2004] Adda-Decker, M. & alii, « Une étude des disfluences pour la transcription automatique de la parole spontanée et l'amélioration des modèles de langage », JEP 2004, Fès (Maroc), 19-22 avril 2004

[Thier2008] T. Bazillon, Y. Estève, D. Luzzati. *Transcription manuelle vs assistée de la parole préparé et spontanée*, Revue TAL, 2008

[Kuma2008] N. Kumar, K. Srinathan, « Automatic Keyphrase Extraction from Scientific Documents Using N-gram Filtration Technique. » 2008

[Coli1996] N. Colineau, J. Caelen . Une approche lexicale pour la reconnaissance d'actes de dialogue. In *Séminaire lexique en traitement de la parole,* Toulouse, France, 1996

[Bres1996] J. Bres, *Dialogisme et polyphonie: approches linguistiques*. De Boeck Université. 2005

[Dolm2000] J.-M. Dolmazon, F. Bimbot, G. Adda, J. Caerou, J Zeiliger, M. Adda-Decker. *Première campagne AUPELF d'évaluation des systèmes de Dictée Vocale : organisation et résultats*. 2000

[Blou2009] O. Le Blouch. *Décodage acoustico-phonétique et applications à l'indexation.* Travail de thèse. 2009

[Bunt2007] H. Bunt, A. Schiffrin. *Interoperable concepts for dialogue act annotation.* Paper presented at the 7th International Workshop on Computational Semantics, Tilburg. 2007

[Hato2006] J-P Halton, C. Cerisara, D. Fohr, Y. Laprie, K. Samaïli, *Reconnaissance automatique de la parole, du digital à son interprétation,* Dunod, Paris, 2006

[Mari2002] Joseph Mariani, *Reconnaissance de la parole,* 2002

[Juan2005] B. H. Juang, L. R. Rabiner. *Automatic Speech Recognition—A Brief History of the Technology*, dans Elsevier Encyclopedia of Language and Linguistics, Second Edition, 2005.

[Cett2005] M. Cettolo, M. Vescovi, and R. Rizzi, *Evaluation of BIC-based algorithms for audio segmentation*, Computer Speech and Language, vol. 19, pp. 147–170, 2005.

[Siva2001] P. Sivakumaran, J. Fortuna, and A. M. Ariyaeeinia, *On the use of the Bayesian information criterion in multiple speaker detection*, dans Proc. Eur. Conf. Speech Communication and Technology, Aalborg, Denmark, Sep. 2001, pp. 795–798.

[Tuba1989] J. P. Tubach, *La parole et son traitement automatique*, Collection technique et scientifique des télécommunications, 1989

[Khou2010] *Unsupervised Video Indexing based on Audiovisual Characterization of Persons.* Thèse de doctorat, Université de Toulouse, juin 2010 [BIC Based Speaker Segmentation]

[Sala2010] H. Salamin, G. Mohammadi, K. Truong, A. Vinciarelli. *Automatic Role Recognition Based on Conversational and Prosodic Behaviour.* Dans Proceedings of the ACM International Conference on Multimedia, pages 847–850, 2010.