Campagne ESTER : une première version d'un système complet de transcription automatique de la parole grand vocabulaire

Martine de Calmès, Jérôme Farinas, Isabelle Ferrané et Julien Pinquier

Institut de Recherche en Informatique de Toulouse
Université Paul Sabatier, Toulouse, France
{decalmes, jfarinas, ferrane, pinquier}@irit.fr

Abstract

Dans cet article, nous décrivons les différents systèmes développés à l'Institut de Recherche en Informatique de Toulouse (IRIT) au sein de l'équipe SAMoVA. Ces systèmes ont été évalués lors de la campagne ESTER qui s'est tenue en janvier 2005. Les ressources distribuées dans le cadre de cette campagne, ainsi que l'infrastructure mise en place pour l'évaluation des systèmes, nous ont permis non seulement de mesurer les performances de systèmes déjà existants relatifs notamment à la tâche de segmentation en événements sonores, mais également développer une première version d'un système de transcription. Ce système, évalué dans le cadre de la tâche de transcription, nous a permis de doter notre laboratoire d'une première version d'un système complet de transcription automatique de la parole grand vocabulaire. Ceci est un premier pas qui nous permettra d'aller vers des applications d'indexation de documents audio vidéo et de recherche d'éléments porteurs d'informations sémantiques comme les entités nommées.

1. Introduction

Le programme français **Technolangue** et plus particulièrement le projet **Evalda** d'évaluation des technologies de la langue relatives au traitement du français oral ou écrit, ont pour objectifs de stimuler les travaux de recherche dans différents domaines. La campagne d'évaluation **ESTER** est centrée sur le traitement automatique de la parole et plus particulièrement sur l'évaluation de systèmes de transcription enrichie d'émissions radiophoniques. Partir du document brut pour aboutir à une transcription fiable à partir de laquelle on puisse extraire des informations pertinentes et fournir ainsi une transcription enrichie du document traité, est en soi une tâche très vaste et très complexe. Elle ne peut se réaliser que par la combinaison de systèmes complémentaires opérant à différents niveaux. Les trois tâches évaluées dans le cadre de la campagne ESTER, *segmentation*, *transcription* et *extraction d'information* [1], correspondent aux phases essentielles qui peuvent rendre possible un tel traitement.

L'organisation de la campagne d'évaluation ESTER, assurée conjointement par l'AFCP, la DGA et l'ELDA, a permis de fournir aux participants un environnement unique en France, que ce soit du point de vue développement de ressources acoustiques et textuelles, vitales pour le développement et l'évaluation des systèmes, ou du point de vue logistique pour la mise en œuvre de la phase d'évaluation proprement dite.

Dans ce contexte très motivant, nous avons pu évaluer des systèmes existants dans notre laboratoire comme la segmentation Parole/Musique/Bruit, ou bien nous lancer dans le développement d'une première version d'un système de transcription automatique de la parole grand vocabulaire. Dans cet article nous décrivons d'abord les caractéristiques des deux systèmes qui ont été présentés dans le cadre de la tâche **SES** (Segmentation en Evénements Sonores), l'un des systèmes proposés présentant des résultats contrastifs par rapport à l'autre. Nous décrivons ensuite la première version du système présenté dans le cadre de la tâche **TRS** (transcription). Ce système effectue la transcription des segments de parole détectés grâce à l'utilisation d'une autre version du système de segmentation que celui présenté dans la cadre de la tâche SES (lissage spécifique pour la détection de pause). Compte tenu du matériel à notre disposition, nous avons également voulu voir quelles étaient les performances de ce même système du point de vue temps réel. C'est pourquoi nous avons également soumis des résultats concernant la tâche **TTR** (transcription temps réel). Enfin, nous conclurons en évoquant les résultats obtenus et les nombreuses perpectives d'évolution de nos systèmes.

2. Participation à la tâche SES

Nous avons développé deux systèmes : le premier désigné comme système contrastif et le second comme système primaire.

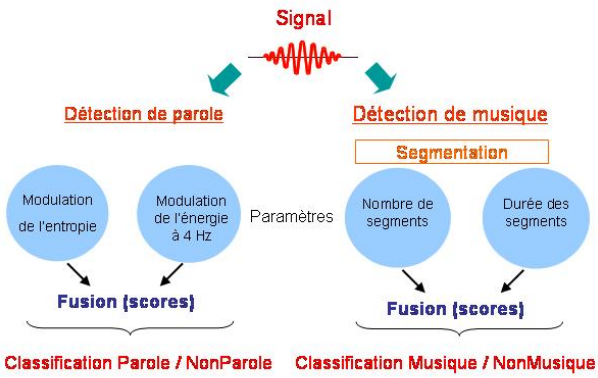

Figure 1 : système SES contrastif.

2.1. Description du système SES dit "contrastif"

Le système se décompose en deux systèmes de classification correspondant aux deux détections disjointes de la parole et de la musique.

Il est fondé sur l'extraction de quatre paramètres (cf. figure 1) : la modulation de l'énergie à 4 Hertz, la modulation de l'entropie, et après l'application d'un algorithme de segmentation, le nombre de segments par seconde et la durée de ces segments [2]. La décision est prise en comparant les scores (vraisemblances) issus de la modélisation de chacun des paramètres considérés.

2.1.1. Paramètres utilisés

- **Modulation de l'énergie à 4 Hertz**

Le signal de parole possède un pic caractéristique de modulation en énergie autour de la fréquence syllabique 4 Hertz [3]. En effet, ces modulations correspondent au rythme syllabique. La parole possède une modulation de l'énergie à 4 Hertz plus forte que la musique.

- **Modulation de l'entropie**

Des observations menées sur le signal ainsi que sur le spectrogramme font apparaître une structure plus « ordonnée » du signal de musique que de parole. Pour mesurer ce « désordre », nous avons calculé un paramètre fondé sur l'entropie du signal [4]. La modulation de l'entropie est alors plus élevée pour la parole que pour la musique.

- **Paramètres de segmentation**

La longueur des segments quasi stationnaires est différente pour la parole et la musique. En utilisant une segmentation du signal en zones quasi stationnaires, nous cherchons à mettre en évidence cette information. La segmentation est issue de l'algorithme de « Divergence Forward-Backward » (DFB) [5] qui est fondé sur une étude statistique du signal dans le domaine temporel **.**

- **Nombre de segments**

Le nombre de segments présents durant chaque seconde de signal est calculé. Les signaux de parole présentent une alternance de périodes de transition (voisées/non-voisées) et de périodes de relative stabilité (les voyelles en général) [6]. Au niveau de la segmentation, cela se traduit par de nombreux changements. La musique, étant plus tonale (ou harmonique), ne présente pas de telles variations. Le nombre de segments par unité de temps (ici la seconde) est donc plus important pour la parole que pour la musique **.**

- **Durée des segments**

La durée des segments, obtenue après segmentation automatique (DFB), est fortement corrélée au nombre de segments par seconde. Afin de limiter la corrélation de ces deux paramètres de segmentation, la durée moyenne des segments sur une seconde est calculée sur les 7 segments les plus longs de la seconde. Le nombre de segments caractéristiques est fixé expérimentalement. Les segments sont généralement plus longs pour la musique que pour la parole.

2.1.2. Commentaires sur les résultats obtenus

Le système est globalement assez performant avec une « F-measure » de 0,93 bien que la détection de musique soit très pénalisante : 0,49. La détection de parole est très bonne : 0,98. Notre système contrastif est fondé sur des seuils qui ont été appris sur des données extérieures à la campagne ESTER : MULTEXT [7] pour la parole et base personnelle pour la musique. Sa particularité est donc sa robustesse mais il atteint ses limites ici, lorsque la parole est sur un fond musical assez faible.

2.2. Description du système SES dit « primaire »

Le second système décrit dans cette section, a été présenté lors de la campagne comme système primaire et met en fait en présence deux systèmes différents.

2.2.1. Fusion de deux systèmes

Nous avons fusionné deux systèmes de segmentation en événements sonores (toujours par maximisation des scores de vraisemblance). Le premier est le système contrastif décrit précédemment (cf. §2.1). Le second est un système classique basé sur des Modèles de Mélanges de lois Gaussiennes (GMM) et une approche spectrale. L'apprentissage de ce second système a été effectué sur une partie seulement du corpus d'apprentissage de la phase 1 (soit 14 heures d'enregistrement, ressources LIA - segmentation vérifiées manuellement sur l'apprentissage de la phase 1 [8]). Ainsi, pour la détection de parole on utilise les coefficients cepstraux (12 MFCC, Énergie et dérivées), la modulation de l'entropie et la modulation de l'énergie à 4 Hertz. Pour la détection de musique, les coefficients spectraux (28 coefficients et Énergie) et les deux paramètres issus de la segmentation automatique sont combinés (cf. figure 2).

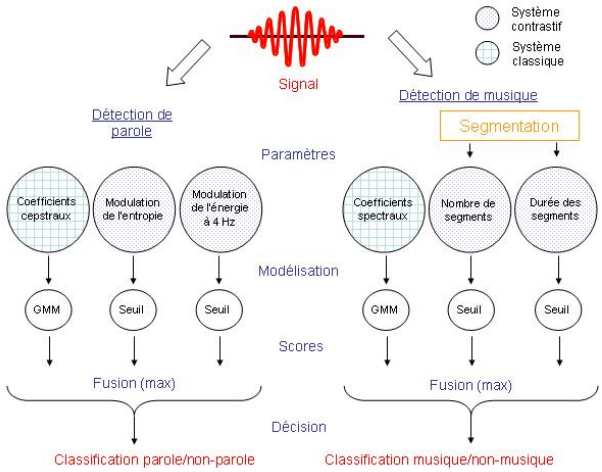

Figure 2 : système SES primaire.

2.2.2. Commentaires sur les résultats obtenus

L'apport des GMM est très bon pour la détection de parole (la F-measure atteint 0,99) car il permet de détecter en plus la moitié des segments de parole non détectés par le système contrastif. Dans le cas de la musique, l'augmentation de la F-measure qui atteint 0,529 est moins significative car seulement une petite partie des segments de musique non détectés par le système contrastif sont récupérés. Le rôle des GMM pour la musique n'est pas concluant car le score de confiance associé aux résultats des paramètres de segmentation (nombre de segments et durée des segments) est trop important ici, les coefficients spectraux n'influent pas alors que seuls ils donnent une F-measure de 0,85 !

2.3. Premier bilan concernant la tâche SES

La détection de parole est satisfaisante mais la détection de musique est mauvaise. Des travaux sont en cours afin de normaliser les scores de confiance de notre système contrastif (notamment pour la détection de musique). Une autre approche visant à définir des GMM dédiés à la parole sur fond musical devrait permettre de mieux identifier les segments de ce type.

3. Participation aux tâches TRS et TTR

L'ensemble des ressources diffusées dans le cadre de la campagne ESTER (enregistrements sonores, transcriptions et corpus de textes) nous a permis de développer notre première version du système de transcription automatique

3.1. Description du système TRS

Pour développer ce système nous avons utilisé un ensemble de ressources logicielles qui nous ont permis de constituer les différents modèles nécessaires pour effectuer la tâche de transcription : modèles acoustiques, lexique phonétique et modèles de langage statistiques. Un ensemble de pré-traitements doit être réalisé au préalable.

3.1.1. Ressources logicielles

- Boîte à outils **HTK** [9] version 3.2 : pour l'apprentissage des modèles acoustiques.
- Boîte à outils **CMU SLM** [10] version 2 : estimation des modèles de langage.
- Moteur de reconnaissance de la parole continue grand vocabulaire : **JULIUS** [11] version 3.4.2 (Options de compilation : réglage standard, N-gram, WordsInt, ShortWordTree, StrictIWCD2, WordPairApprox, ShortPauseSegment).

3.1.2. Prétraitements

La détection des zones de parole à été réalisée en utilisant une version des programmes de segmentation utilisant des GMM. Un lissage de 200 ms a été ensuite effectué sur le résultat. Pour des raisons de performance du moteur de reconnaissance, la durée maximale d'un tour de parole a été limitée à 8 minutes. En complément, la détection de pauses courtes, intégrée au moteur de reconnaissance, a été activée. Dans la mesure où un modèle de pause de plus de 10 fenêtres est détecté au cours du premier décodage phonétique, le fichier à traiter est sectionné à ce niveau, la suite du décodage étant alors reprise après la pause détectée.

Le décodage phonétique est effectué en fonction des caractéristiques des paramètres des vecteurs d'observation. Il s'agit de fenêtres de 16 ms avec un recouvrement 8 ms. On utilise un fenêtrage de hamming. Les vecteurs sont constitués de 39 paramètres soit 12 MFCC, l'énergie, leur dérivée première et leur dérivée seconde. Les caractéristiques suivantes sont également prise en compte : normalisation de l'energie, pré-emphase de 0.97, fréquence minimum 300 Hz et maximum 8000 Hz. Ces paramètres sont calculés en utilisant la boîte à outils HTK.

3.1.3. Modèles acoustico-phonétiques

La modélisation acoustique est réalisée en utilisant des modèles de Markov cachés gauche-droite à 3 états. Chaque état est décrit par un mélange de 32 lois gausiennes.

L'initialisation et les réestimations des modèles ont été réalisées grâce aux transcriptions phonétiques alignées distribué par l'IRISA/ENST [12].

La modélisation acoustique compte 35 phonèmes ainsi que deux modèles représentant respectivement les pauses courtes et les pauses longues.

La réestimation de ces modèles par l'algorithme de Baum-Welch a été effectuée sur le corpus d'apprentissage fourni lors de la phase 1 de la campagne (soit environ 31 heures d'enregistrement).

3.1.4. Lexique phonétique

Pour constituer le lexique de l'application, nous avons d'abord sélectionné l'ensemble des formes présentes dans la transcription manuelle du corpus d'apprentissage de la phase 2 (environ 35 000 formes graphiques). Cet ensemble a ensuite été complété par les mots les plus fréquents apparaissant au moins douze fois dans le corpus du Monde 1987-2002 et/ou dans le corpus du Monde 2003 (environ 26 500 formes graphiques). A noter que des groupes de mots, représentant des locutions et expressions courantes, ont été considérés en tant qu'entrée lexicale, comme par exemple la locution adverbiale « tout_à_fait » où la contrainte phonologique portant sur la consonne latente de l'élément « tout » est levée.

Pour phonétiser le lexique obtenu, nous avons utilisé les ressources lexicales phonologiques développées à l'IRIT, notamment la ressource BDLex [13] pour ce qui concerne les noms communs, complétées par environ 6000 mots nouveaux (en majorité des noms propres) dont la phonétisation a été vérifiée manuellement.

Ce lexique est constitué de 61 225 mots sous forme orthographique, ce qui se traduit ensuite par 119 297 formes phonétiques, un même mot pouvant avoir plusieurs prononciations. Les variantes de prononciation résultent en effet des transformations phonologiques ou des variantes libres de prononciation prises en compte dans les ressources de type BDLex, telles que le schwa dans « samedi », la consonne de liaison z'' dans « les », l'élément optionnel (l) dans « fusil » ou le groupe à prononciations multiples {6R} dans « starter » – ces groupes à prononciations multiples sont souvent rencontrés dans les mots d'origine étrangère. La représentation phonétique des mots utilisant initialement le jeu d'unités phonétiques disponible dans BDLex a été convertie de façon à être compatible avec le jeu d'unités retenu dans la phase d'apprentissage des modèles acoustico-phonétiques.

Le taux de couverture lexicale par rapport au corpus de développement a été évalué à 1358 mots hors vocabulaire sur environ 96 000 mots, soit un taux de mots hors vocabulaire de 1,39%.

3.1.5. Modèles de langage

L'utilisation du moteur de reconnaissance Julius cité précédemment, nécessite différentes ressources linguistiques. Pour décrire les enchaînements possibles des formes orthographiques des mots, et effectuer un décodage en deux passes, il est nécessaire de définir deux modèles de langage N-grammes : un modèle **bigramme** pour la première et un modèle **trigramme** inversé pour la seconde.

Ces modèles ont été construits avec les outils du CMU SLM dédiés à la construction de modèles de langage statistiques introduit plus haut. Les modèles obtenus sont des modèles définis sur un **vocabulaire ouvert** c'est-à-dire prévoyant une classe pour les mots hors vocabulaire. Une méthode de prélèvement (ou **discount**) a été appliquée pour prélever une masse de probabilité déterminée à partir des N-grammes observés, masse qui sera ensuite redistribuée suivant la technique du repli (ou **back-off**) sur les N-grammes non présents dans le corpus d'apprentissage mais présents dans le corpus de test.

Plusieurs versions des modèles nécessaires au moteur de reconnaissance ont été construites. Certaines à partir d'une seule source, d'autres en mélangeant les bigrammes (resp. trigrammes) de différentes sources distribuées dans le cadre de la campagne : (1) Transcription des fichiers audio représentant 90 heures d'enregistrement et utilisés pour l'apprentissage de la phase 2 ; (2) les textes du journal Le Monde publiés entre 1987 et 2002 ; (3) les textes du journal Le Monde publiés en 2003. Nous n'avons pas utilisé la quatrième source disponible, à savoir les transcriptions des débats du Conseil Européen.

Pour la constitution des modèles de langage, un premier travail de normalisation a été effectué pour que les mots présents dans les corpus écrits (2) et (3) aient la même représentation que ceux présents dans le corpus issu de la transcription des énoncés oraux (1) : valeurs numériques, unités de mesures, énoncés de sites web, …

Bien que la version actuelle de la boîte à outils SLM du CMU propose 4 méthodes différentes de prélèvement, nous n'avons utilisé que deux d'entre elles : Good Turing ou Witten-Bell. Pour chacune d'elles, plusieurs jeux de modèles bigrammes et trigrammes inversés ont été générés avec des valeurs de cutoff différentes (0, 1 ou 2 pour les bigrammes (B) et 0, 1, 2 ou 3 pour les trigrammes (T)).

Le tableau 1 présente les résultats de l'évaluation des modèles générés en utilisant la première méthode de prélèvement (Good-Turing). Ces modèles ont été évalués grâce à l'outil disponible dans la boîte à outils du CMU appliqué à la transcription du corpus de développement.

Le même travail a été effectué pour une seconde série de modèles générés en appliquant la seconde méthode de prélèvement (Witten-Bell). Cependant, son évaluation complète en terme de calcul de la perplexité, n'a pu être terminée à temps pour la fin de la campagne. Aujourd'hui disponible, ce résultat montre que, pour les mêmes modèles (même type, même source et mêmes cutoff) la perplexité varie de moins de 4%, deux fois sur trois en faveur d'un modèle construit avec la méthode de prélèvement Witten-Bell. Or, comme rappelé dans [14], une réduction de la perplexité inférieure à 5% par rapport à un modèle de référence n'est pas significative.

Source	Modèle Bigramme		Modèle Trigramme	
	Cutoff	Perplexité	Cutoff	Perplexité
Ester (1)	0	201.56	0 (B) 0 (T)	169.81
			0 (B) 1 (T)	168.84
	1	220.99	1 (B) 2 (T)	187.38
	2	239.94	2 (B) 2 (T)	203.24
			2 (B) 3 (T)	206.24
Ester et Le Monde 2003 (1) + (2)	0	181.80	0 (B) 0 (T)	124.02
			0 (B) 1 (T)	124.40
	1	188.60	1 (B) 2 (T)	133.14
	2	196.52	2 (B) 2 (T)	139.05
			2 (B) 3 (T)	142.19
Ester et Le Monde 87-03 (1)+(2)+(3)	0	190.27	0 (B) 0 (T)	113.32
			0 (B) 1 (T)	113.96
	1	194.11	1 (B) 2 (T)	118.63
	2	197.19	2 (B) 2 (T)	120.62
			2 (B) 3 (T)	122.67

Tableau 1 : Perplexité des différents modèles de langages générés (méthode de prélèvement = Good Turing)

Bien que l'étude de la perplexité en soi ne permette pas uniquement de déterminer la qualité d'un modèle, la sélection du modèle utilisé s'est faite en fonction de ce seul critère. Disposer du taux d'erreur sur les mots aurait permis d'affiner cette sélection et d'obtenir des résultats plus ou moins contrastifs, mais le temps de traitement et la proximité de l'échéance de la fin de la campagne ne nous ont pas permis d'effectuer ce travail à temps pour pouvoir en exploiter les résultats.

Les résultats soumis lors de la campagne ont donc été obtenus avec les modèles construits sur l'ensemble des sources, sans cutoff et en appliquant la méthode de prélèvement Good Turing. (perplexité en gris dans le tableau 1).

3.1.6. Temps de traitement

Dans un but constrastif, les calculs ont été effectués sur deux machines différentes :

- sur une machine multi-processeur SGI Altix 3300 : 12 processeurs Itanium2 à 900 Mhz, 2 Go de mémoire vive par processeur partageable, système linux Red Hat Linux Advanced Serveur 3 modifié par Silicon Graphics ;
- sur un ordinateur personnel Dell Précision 360, processeur Pentium 4 3 Ghz, 2 Go de mémoire vive, système Red Hat Linux Entreprise 3.

Dans le premier cas, le calcul réparti sur 10 processeurs de l'Altix 3300 a pris 11h de calcul (utilisation mémoire d'environ 1.2 Go max).

Dans le second cas, c'est-à-dire sur le Dell, cela a pris 39h11min pour tout calculer, soit environ 3.92xRT.

3.2. Description du système TTR

Les ressources matérielles dont nous disposons à l'IRIT nous ont conduit à proposer des résultats dans le cadre de la tâche TTR. Les composantes du système de reconnaissance utilisé sont exactement les mêmes que celles du système évalué dans la cadre de la tâche TRS (modèles acoustiques, lexique et modèles de langage). Une autre version de Julius compilée avec la configuration rapide a été utilisée. Aucune autre modification ou aucun autre réglage n'a été effectué pour optimiser le temps processeur disponible.

3.2.1. Ressources matérielles et logicielles

- Julius compilé avec une configuration rapide et avec les options : N-gram, WordsInt, WordPairApprox, ShortPauseSegment.
- Matériel : Dell Précision 360 PIV 3 Ghz 2Go RAM.

3.2.2. Temps de traitement

Les temps de calculs sont effectivement inférieurs au 1xRT. Il a fallu 6h33 pour réaliser tous les calculs, soit 0.65xRT. Aucun gain en terme de taux de mots corrects n'était à attendre de cette expérimentation.

3.3. Premier bilan concernant les tâches de transcription

Bien sûr, les performances de notre système de transcription ne sont pas à la hauteur de ce qu'on pourrait attendre d'un participant à une campagne d'évaluation. Le taux d'erreur sur les mots est de l'ordre de 64,3% pour la tâche TRS et de 70,6% pour la tâche TTR. Les performances en terme de temps d'exécution pourraient être améliorées par une meilleure utilisation de la valeur du cutoff. En effet, dans les modèles utilisés pour l'évaluation, cette valeur, que ce soit pour les bigrammes ou les trigrammes était nulle. Si le nombre de bigrammes est de l'ordre de 16,5 millions, le nombre de trigrammes est beaucoup trop élevé. En effet, pour le même nombre de bigrammes, la différence entre le modèle sans cutoff au niveau des trigrammes (ligne 11 du tableau 1) et celui avec un cutoff de 1 (ligne 12 du tableau 1) se traduit par une diminution de plus de deux tiers du nombre de trigrammes et ce pour une très faible variation de la perplexité : augmentation de 0.64. Un choix plus judicieux aurait pu être fait au moment de la campagne en utilisant des modèles moins lourds et par conséquent moins gourmands en temps de traitement et espace mémoire. A court terme, une nouvelle évaluation pourra être menée de façon à mesurer l'impact réel de ce changement.

4. Objectifs à court/moyen terme

Notre participation à la campagne ESTER a été très stimulante. Les bonnes performances en ce qui concerne la détection des segments de musique, vient contrebalancer celles de la transcription. Comme nous l'avons évoqué, les améliorations sont à mener sur plusieurs fronts.

4.1.1. Du point de vue segmentation en événements sonores

Les résultats obtenus sur les zones comportant à la fois parole et musique sont à améliorer. En effet, dans les systèmes actuels (primaire et contrastif), une partie des segments comportant de la parole et de la musique sont détectés comme des segments de parole. C'est le cas des segments comportant un léger fond musical comme lors de l'annonce ou du résumé des titres du journal par exemple. Ceci est dû notamment à la prise en compte du nombre de segments (cf. § 2.1.1). Dans le cas de notre système complet, où seul le résultat de la détection de parole est utilisé la parole sur fond musical si elle n'est pas traitée spécifiquement, fait chuter les performance du système de transcription. Un des objectifs à court terme est de réestimer les modèles de façon à disposer de modèles spécifiques permettant de traiter les différents types de parole : parole pure, parole téléphonique et parole sur fond musical.

4.1.2. Du point de vue transcription.

Chaque composante du système de transcription peut sans conteste être améliorée. Un gros travail reste à faire du point de vue de chaque modélisation.

- **Modélisation acoustique**

Les modèles acoustiques utilisés pour l'évaluation, n'ont été appris uniquement que sur le corpus d'apprentissage distribué pour le test à blanc, soit environ 31 heures d'enregistrements.

La transcription phonétique de la seconde partie du corpus d'apprentissage nous a été fournie par l'IRISA. L'adaptation des modèles acoustiques est en cours. Elle nous permettra de prendre en compte la totalité des 90 heures d'apprentissage disponibles pour la phase 2 de la campagne.

A court terme, le passage à des modèles de phonèmes en contexte sera envisagé ainsi que l'augmentation du nombre de gaussiennes.

- **Modélisation lexicale**

D'un point de vue lexical, les améliorations peuvent se porter sur la couverture lexicale et sur la représentation phonétique des mots du dictionnaire. Le taux de mots hors vocabulaire obtenu sur le corpus d'apprentissage (évoqué en fin de section 3.1.4) pourrait être diminué par la prise en compte de noms propres présents dans l'actualité.

Comme déjà évoqué, le code phonétique utilisé pour les besoins de l'apprentissage, est celui de l'IRISA. Un retour aux codes de BDLex, voire l'utilisation d'une autre ressource lexicale disponible à l'IRIT MHATLex peuvent être envisagés. Plus adaptée à la reconnaissance de la parole MHATLex permet de modéliser la prononciation avec ses variabilités libres et contextuelles [15].

- **Modélisation du langage**

En ce qui concerne les modèles de langage, il est impératif de procéder à court terme à l'interpolation des modèles relatifs à chacune des sources utilisées. Dans l'urgence, le modèle utilisé a été construit en mélangeant les 3 sources disponibles (cf. § 3.1.5). Cependant il est clair que compte tenu de la quantité de données représentées par les corpus de textes du Monde, comparativement à la quantité de données issues de la transcription des fichiers audio relatifs à la partie apprentissage, les connaissances linguistiques plus spécifiques de l'oral et donc plus proches du type de document à transcrire

automatiquement sont très certainement « noyées » par les connaissances caractérisant mieux le langage écrit.

5. Conclusion

Nous n'avons pas participé au test à blanc lancé début 2004 car nous ne disposions alors d'aucun système de transcription. Ce système n'a été opérationnel que quelques jours avant la fin de la campagne, en janvier 2005. Les ressources distribuées dans le cadre de la campagne ESTER ainsi que l'infrastructure mise en place pour l'évaluation nous a permis non seulement de tester les performances de deux systèmes de segmentation en événements sonores mais surtout de développer une première version d'un système de transcription. Les résultats obtenus lors de l'évaluation du système développé ne sont pas comparables à ceux des autres participants mais, ce qu'il faut voir dans cette première version, c'est la mise en place des briques de base permettant à notre équipe de disposer d'un système complet de transcription automatique.

Le champ des améliorations possibles est très vaste, puisqu'il touche aussi bien les modèles acoustiques, le lexique que les modèles de langage. En un sens, avoir comme perspectives d'obtenir des résultats plus honorables et plus proches de ceux obtenus par la majorité des participants à cette campagne est en soi une très bonne motivation à poursuivre ces travaux. De plus, disposer d'un système complet avec de meilleures performances nous donnera accès à d'autres types d'applications comme l'indexation de documents audio vidéo et la détection, dans le résultats des transcriptions, d'éléments porteurs d'informations plus sémantiques comme les entités nommées.

6. Références

[1] G. Gravier, J.-F. Bonastre, E. Geoffrois, S. Galliano, K. Mc Tait, K. Chouckri, « *ESTER, une campagne d'évaluation des systèmes d'indexation automatique d'émissions radiophoniques en français* », Journée d Etude sur la Parole 2004, Fes, Maroc.

[2] J. Pinquier, J.L. Rouas et R. André-Obrecht, « *Fusion de paramètres pour une classification automatique parole/musique robuste* ». Technique et Science Informatique (TSI), 22(7-8), 2003, p. 831-852.

[3] T. Houtgast et J.M. Steeneken, « A review of the MTF Concept in Room Acoustics an dits Use for Estimating Speech Intelligibility in Auditoria », Journal of the Acoustical Society of America, 77(3), 1985, p. 1067-1077.

[4] R. Moddemeijer, « *On Estimation of Entropy and Mutual Information of Continuous Distributions* », Signal Processing, 16(3), 1989, p. 233-246.

[5] R. André-Obrecht, « *A New Statistical Approach for Automatic Speech Segmentation* », IEEE Transactions on Audio, Speech, and Signal Processing, 36(1), 1988, p. 29-40.

[6] Calliope, « *La parole et son traitement* », 1989, Paris, France : Masson.

[7] J. Campione et J. Véronis, « A Multilingual Prosodic Database », ICASSP'98, p. 3163-3166, Sydney, Australia.

[8] http://www.afcp-parole.org/ester/private/Ester_SegAcManuel.tar.gz

[9] S. Young "The HTK Hidden Markov Model Toolkit : Design and Philosophy", Rapport technique 152, Cambridge University Engineering Department, UK, 1994.

[10] P. Clarkson, R. Rosenfeld, *Statistical Language Modeling using the CMU-Cambridge toolkit.* In Proc. Eurospeech '97, September 1997.

[11] A. Lee, T. Kawahara, K.Shikano *"Julius- an Open Source real Time Large vocabulary Recognition Engine"* Eurospeech 2001, 3- septembre 2001, Aalbord, Danemark.

[12] M. de Calmès, G. Pérennou, *BDLEX : a Lexicon for Spoken and Written French*, Dans : 1st Int. Conf. on Langage Resources & Evaluation, Grenade, 28-30 mai 1998. p. 1129-1136.

[13] http://www.afcp-parole.org/ester/private/align-irisa-enst.tar.gz.

[14] A. Allauzen « Modélisation linguistique pour l'indexation automatique de documents audiovisuels », Université Paris XI, Orsay, décembre 2003.

[15] G. Pérennou, M. de Calmès. *MHATLex: Lexical Resources for Modelling the French Pronunciation*. Dans : 2nd Int. Conf. on Language Resources and Evaluation, Athens, Greece, 31 mai-2 juin 2000 p. 257-264.