Structuration de l'information parlée

Sommaire

Le chapitre 2 était dédié à la recherche d'information au travers d'une description de la recherche documentaire (section 2.1) et du résumé automatique (section 2.2). Les méthodes présentées sont pour la plupart issues du traitement de la langue naturelle écrite avant d'être adaptées aux problématiques de la parole. Cette adaptation implique l'extraction de descripteurs structurels et sémantiques dans le flux de parole. Nous allons maintenant nous intéresser aux différentes étapes nécessaires pour structurer ce type de flux.

La figure 3.1 illustre les différentes étapes de la structuration. Une chaîne de structuration a été développée au LIA : elle regroupe un module de segmentation en macro-classes acoustiques, un module d'indexation en locuteur et un module de transcription orthographique. Cette chaîne, antérieure aux travaux décrits dans ce document, est présentée rapidement par la section 3.1. Les différentes composantes sont évaluées en section 3.2, au travers de la campagne ESTER. Nous avons développé d'autres éléments de structuration, la segmentation en phrases et l'extraction d'entités nommées, pour compléter cette chaîne de structuration. Ces éléments feront l'objet du chapitre 4.

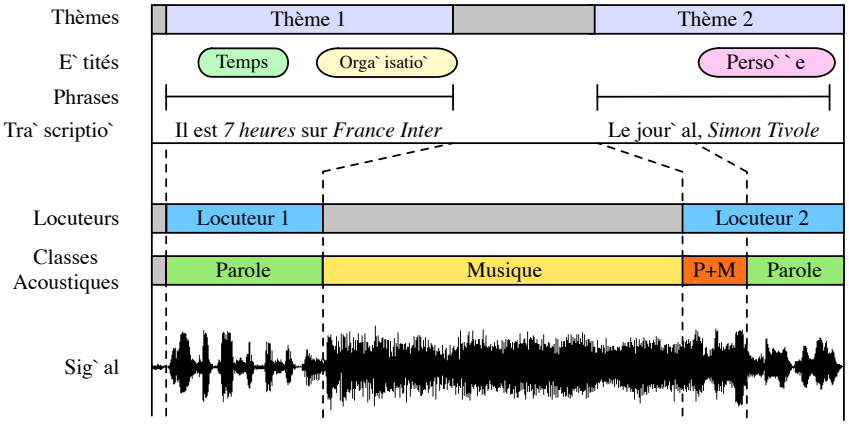

FIG. 3.1: *Principe de la structuration du flux de parole par tâches de plus en plus proches de la sémantique. Le signal sonore est découpé en classes acoustiques pour détecter les portions contenant de la parole ; puis les tours de parole et éventuellement l'identité des locuteurs sont retrouvés. Ces informations permettent d'améliorer la transcription orthographique du discours parlé. Cette dernière facilite l'extraction de descripteurs sémantiques de plus haut niveau (entités nommées, thèmes...).*

3.1 La chaîne de structuration Speeral

Le système de structuration de parole du LIA effectue la transcription du contenu parlé d'un document audio et génère une segmentation en locuteurs tout en étiquetant les zones de silence et de musique. La plupart des outils ont été développés au LIA et utilisent des techniques classiques d'apprentissage artificiel. La figure 3.2 détaille le fonctionnement de la chaîne étape par étape.

3.1.1 Paramétrisation acoustique

La parole est stockée sous la forme d'un signal numérique généralement quantifié sur 16 bits à une fréquence de 16000 échantillons par seconde. La plupart des tâches de structuration reposent sur une reconnaissance de forme dans cet espace. Pour rendre cette reconnaissance possible, les paramètres d'un modèle de production et/ou de perception sont représentés sous forme de vecteurs dans un « espace acoustique ». Les paramètres les plus répandus sont : *Linear Predictive Cepstral Coefficient* (LPCC, Rahim et Lee, 1996), *Mel Frequency Ceptstrum Coefficents* (MFCC, Davis et Mermelstein, 1980), ou encore *Perceptual Linear Predictive* (PLP) analysis (Hermansky, 1990).

Dans la chaîne présentée, le signal de parole est découpé en vecteurs acoustiques (d'une portée de 25 ms, avec un décalage de 10 ms) représentant les fréquences caractéristiques de la parole, ainsi que leur dynamique. Les paramètres extraits prennent en compte à la fois la production et la perception de la parole (12 coefficients PLP et l'énergie, leurs dérivées et dérivées secondes, soit 39 dimensions pour la transcription (Lévy et al., 2004) ; les autres tâches reposent sur des jeux de paramètres similaires). La

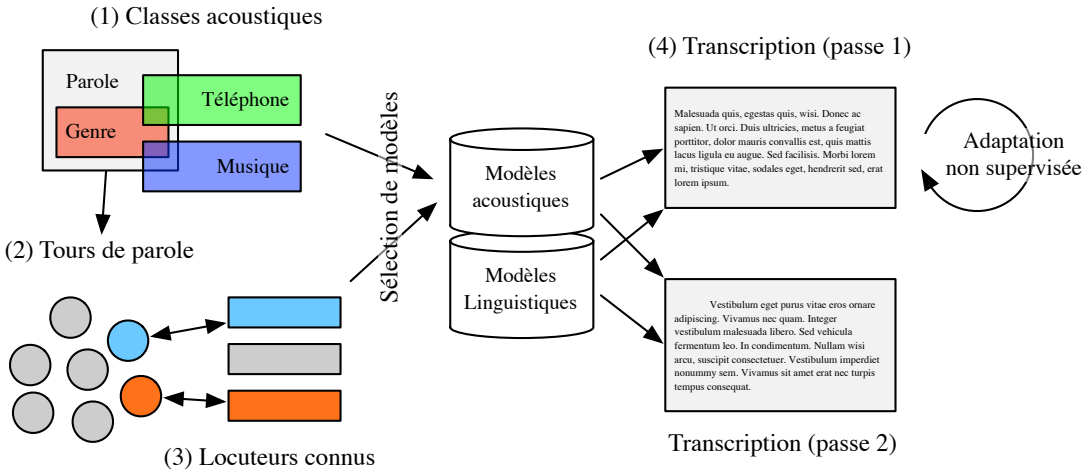

FIG. 3.2: *La chaîne de structuration audio Speeral procède en 4 étapes : segmentation en classes acoustiques (1), segmentation en tours de parole (2), identification de locuteurs connus (3), transcription (4) à l'aide de modèles adaptés aux caractéristiques de la parole détectées dans les étapes précédentes, seconde passe de transcription après adaptation non supervisée des modèles acoustiques.*

figure 3.3 illustre la paramétrisation acoustique.

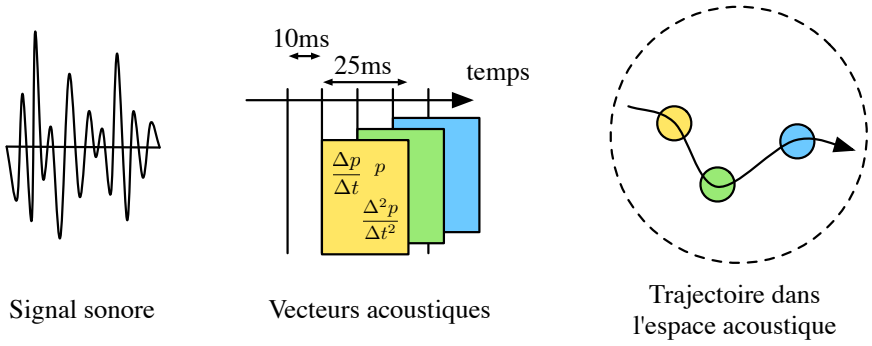

FIG. 3.3: *Passage d'une représentation numérique du signal sonore en un espace acoustique dédié à l'application de méthodes mathématiques et statistiques classiques pour les tâches de segmentation et d'identification.*

3.1.2 Segmentation en classes acoustiques

Un segmenteur en classes acoustiques permet de séparer les différents types de signaux (parole, silence, musique, bruit) et de classifier le type de parole (genre du locuteur, téléphone ou studio, parole sur musique) afin d'utiliser des modèles adaptés à l'environnement acoustique durant les phases suivantes. Cette approche est implémentée dans la chaîne de structuration sous forme d'un *Hidden Markov Model* (HMM) ergodique dont les états sont modélisés par des *Gaussian Mixture Models* (GMM, Fredouille

et al., 2004). Ces densités de probabilités sont estimées par l'algorithme Estimation-Maximisation (EM, Dempster et al., 1977) ; les probabilités de transition entre les états sont estimées par l'algorithme Baum-Welch (Baum et al., 1970). Cette approche est illustrée dans la figure 3.4. La séquence la plus probable est trouvée par programmation dynamique à l'aide de l'algorithme Viterbi (Viterbi, 1967).

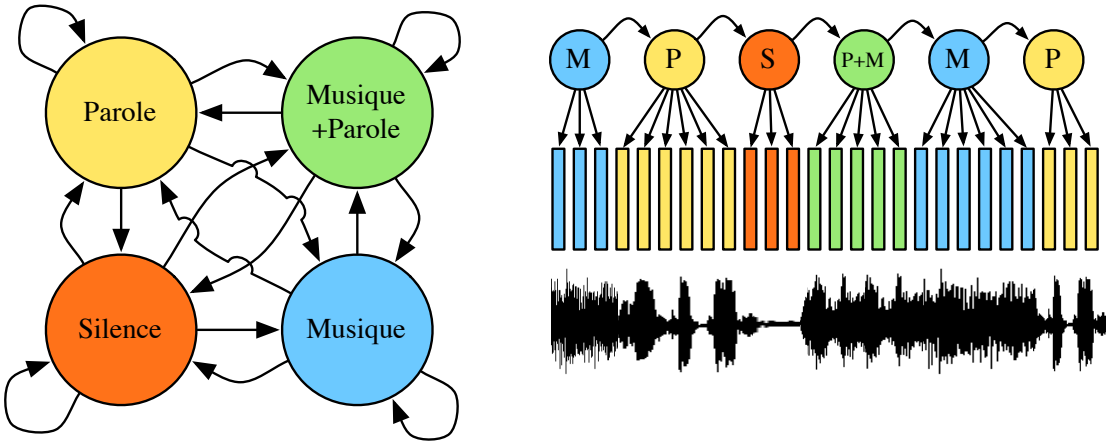

FIG. 3.4: *Illustration de la segmentation en classes acoustiques d'un signal sonore à l'aide d'un HMM ergodique dont les données générées par les états sont modélisés par des GMM.*

Ce type de modélisation a été étendu à d'autres tâches : Atrey et al. (2006) appliquent les techniques de modélisation GMM à la détection d'événements sonores (bruits de pas, course, pleurs, bruits de chute) pour la surveillance multimédia ; Dufaux et al. (2000) utilisent des techniques similaires, pour détecter des événements sonores dans un environnement bruité.

3.1.3 Indexation en locuteurs

L'indexation en locuteurs consiste en une étape de segmentation en tours de parole, suivie du regroupement des tours de parole en locuteur et d'une identification des locuteurs connus (suivi de locuteur). Connaître l'identité des locuteurs permet l'emploi de modèles adaptés pour les locuteurs fréquents lors de la phase de transcription (comme, par exemple, les présentateurs d'émissions radiophoniques).

Dans notre cas, la segmentation et le suivi de locuteur sont réalisés grâce aux outils LIA_SpkSeg et LIA_SpkDet, fondés sur Alize[1] (Istrate et al., 2005). La segmentation est générée par un HMM dynamique auquel est ajouté un état à chaque fois qu'un nouveau locuteur prend la parole. Les locuteurs connus sont ensuite recherchés parmi les regroupements de tours de parole. La décision de classification provient du rapport de vraisemblance entre un modèle de locuteur et un modèle générique (UBM). Comme

[1] disponible sous une licence GPL sur `http://lia.univ-avignon.fr/heberge/ALIZE/`, visité en septembre 2006

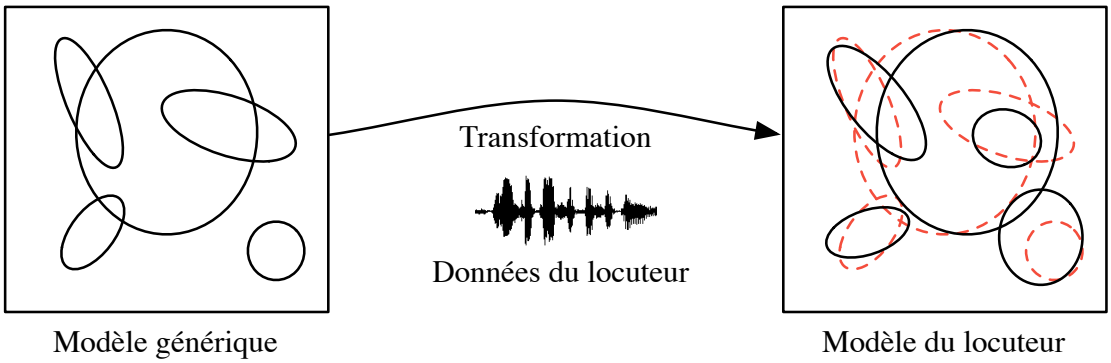

FIG. 3.5: *Modélisation de l'occupation de l'espace acoustique par un locuteur en estimant une transformation d'un modèle générique à partir des données observées de ce locuteur.*

peu de données sont disponibles sur l'occupation de l'espace acoustique par un locuteur, le modèle de ce locuteur est estimé à l'aide d'une adaptation du modèle générique aux données observées (voir figure 3.5).

3.1.4 Transcription automatique

La transcription orthographique consiste en une reconnaissance de la séquence de mots prononcée dans un flux de parole. Les systèmes de transcription actuels sont indépendants du locuteurs, traitent de la parole continue et reconnaissent un vocabulaire étendu (*Large Vocabulary Continuous Speech Recognition*, LVCSR).

Dans la chaîne de structuration présentée, la transcription automatique est effectuée en 2 passes, dont la première sert à générer rapidement une séquence de mots approximative. Celle-ci permet une adaptation en aveugle des modèles acoustiques. Les modèles ainsi adaptés sont utilisés en deuxième passe. Le système de transcription, Speeral (Nocéra et al., 2004), est un moteur de reconnaissance de la parole grand vocabulaire, multi-locuteurs utilisant une reconnaissance HMM des phonèmes, un lexique de phonétisation et des modèles linguistiques n-grammes. Le meilleur chemin dans le graphe d'hypothèses est déterminé grâce à une version modifiée d'A* basée sur une estimation à moindre coût de la fin de la transcription (acoustique et prédiction linguistique) et une méthode d'élagage de l'arbre d'hypothèses.

Des ouvrages comme (Huang et al., 2001), (De Mori, 1998), ou (Haton et al., 2006) expliquent plus en détail le fonctionnement d'un système de transcription de la parole.

3.1.5 Traitements de plus haut niveau

La segmentation en macro-classes acoustiques, l'indexation en locuteurs et la transcription orthographique représentent les éléments de structuration fournis par la chaîne au début des travaux présentés dans ce document. Nous y avons ajouté des traitements

spécifiques au résumé automatique : une segmentation en phrases et une détection d'entités nommées. Ces modules sont décrits en détail dans les sections 4.1 et 4.2. Nous n'abordons pas la segmentation et le suivi de thèmes qui pourraient aussi être importants selon les traitements de haut niveau envisagés. Walls et al. (1999) proposent une détection de thème dans des émissions radiodiffusées. Shriberg et al. (2000) étudient des paramètres prosodiques pour une détection de frontières thématiques.

3.2 Évaluation lors de la campagne ESTER

La campagne d'Évaluation des Systèmes de Transcription d'Émissions Radiophoniques (ESTER) a été organisée par l'Association Francophone de la Communication Parlée (AFCP), la Délégation Générale pour l'Armement (DGA) et *Evaluation and Language resources Distribution Agency* (ELDA), dans le cadre du projet EVALDA (évaluation des technologies de la langue en français), un volet de l'action Technolangue, financée par le Ministère de la Recherche et de l'Industrie. Entre 2003 et 2005, différentes phases de la campagne ont dynamisé les sous-domaines du traitement de la parole, en fournissant les moyens d'évaluer les systèmes issus de la recherche sur des tâches bien définies et reconnues au niveau international.

3.2.1 Présentation des données et des tâches

Un premier corpus de 90 heures de radio en français, transcrit et annoté, a été fourni aux participants. Ce corpus est divisé en 2 parties aux fonctions différentes : la première, « ensemble d'entraînement » (*train*), sert à l'apprentissage des paramètres et des modèles utilisés dans les systèmes automatiques ; la seconde, « ensemble de développement » (*dev*), donne une estimation des performances des systèmes automatiques lors de leur développement. À ce corpus est ajouté un corpus de test, distribué sans annotation ni transcription quelques temps avant la date de soumission des résultats de l'évaluation. Ce corpus permet de comparer les performances des participants de façon raisonnablement équitable. Des ressources annexes facilitent le travail des participants : une grande quantité de corpus de texte pour l'apprentissage de modèles de langage (10 ans du journal Le Monde, soit 400 millions de mots) et un corpus audio non transcrit pour les approches non supervisées (1700 heures de radio).

Les données, dont la répartition par source et par corpus est détaillée dans la table 3.1, sont en majorité des journaux radio diffusés et des émissions radiophoniques impliquant des invités et des interventions d'auditeurs. De nombreuses difficultés sont présentes dans les données et peuvent détériorer la qualité de l'annotation automatique :

- des locuteurs parlant une variante nord-africaine du français avec notamment des prononciations de noms propres en arabe ;
- des tours de parole dans une langue étrangère avec éventuellement un doublage ;
- de nombreuses difficultés d'élocution comme des hésitations, des coupures, des reprises, et des lapsus ;

Source	Entr.	Dév.	Test	Non-Trans
France Inter	33h03	2h00	2h00	346h24
France Info	8h01	2h00	2h00	660h10
RFI	23h00	2h00	1h59	457h14
RTM	18h28	1h58	2h04	-
France Culture	-	-	1h01	260h29
Radio Classique	-	-	1h00	-
Total	82h	8h	10h	1724h

TAB. 3.1: *Répartition des données de la campagne ESTER entre les corpus d'entraînement (Entr.), de développement (Dév.), de test (Test) et non transcrit (Non-Trans).*

– de la parole sur un fond musical, systématiquement au moment des titres du journal ;
– des bruits de fond, comme des éternuements et des jingles courts structurant les émissions ;
– des segments où plusieurs locuteurs parlent en même temps ;
– des coupures de fréquences dues à l'enregistrement.

Certaines de ces difficultés sont des cas particuliers trop peu nombreux pour que les systèmes soient capables de les traiter (par exemple, les tours de parole de locuteurs dans une langue étrangère). Ces difficultés sont annotées dans le corpus pour être ignorées lors de l'évaluation. Les spécificités du corpus de tests sont détaillées dans la table 3.2 ; la table 3.3 donne la répartition des conditions acoustiques et du genre des locuteurs sur les segments évalués du corpus de test. Un exemple de l'annotation de référence ESTER est décrit dans La figure 3.6.

Durée	10h07
Nombre de mots	103203
Nombre de locuteurs	343
Transcriptions ignorées en éval.	5.99%
Segmentations ignorées en éval.	2.47%
Parole simultanée	0.43%
Musique et bruits	4.95%

TAB. 3.2: *Spécificités du corpus de test ESTER.*

Les tâches de l'évaluation ESTER, décrites dans (Galliano et al., 2005), sont de 3 formes : transcription, segmentation et extraction d'information.

Parmi les tâches de transcription, la première est la transcription orthographique du contenu parlé sans limites de ressources, alors que la seconde est limitée en temps de calcul (le système doit transcrire le corpus dans un temps équivalent à sa durée). La mesure d'évaluation de la tâche de transcription est le taux d'erreur de mots (*Word Error Rate*, WER).

Une première tâche de segmentation est focalisée sur le suivi d'événements sonores sous la forme d'une détection des classes acoustiques suivantes : parole, musique et

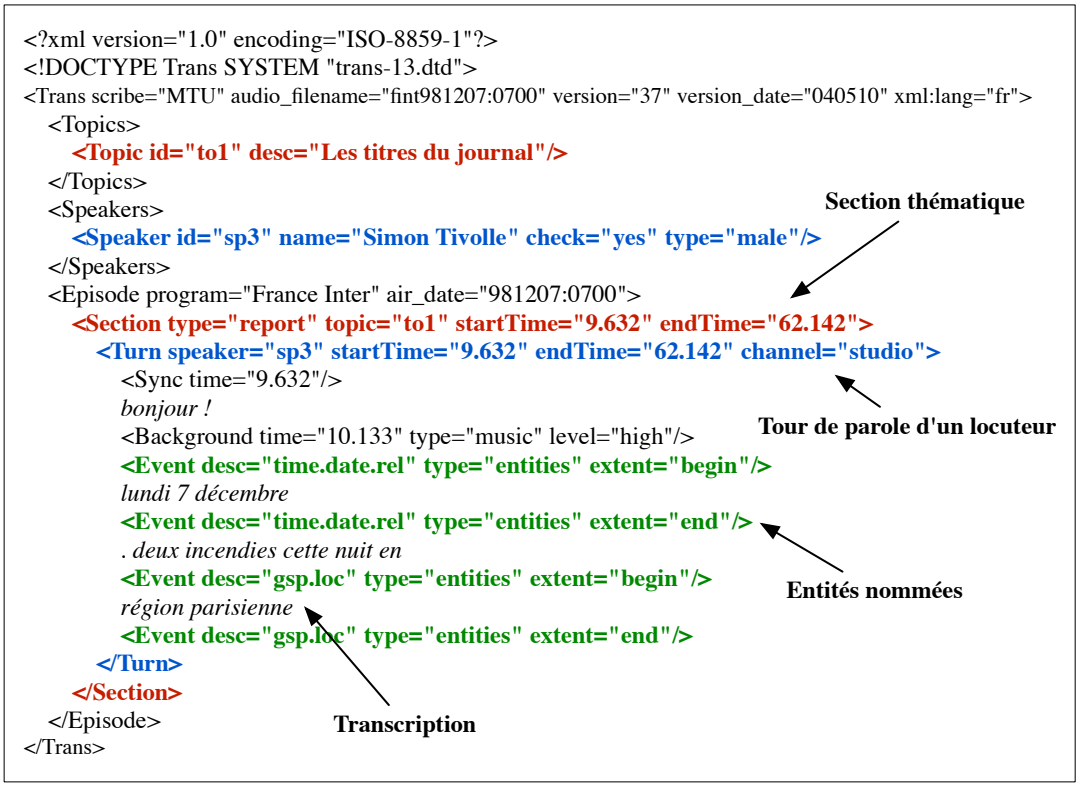

FIG. 3.6: *Exemple de structuration de référence des données ESTER : l'annotation est réalisée par un expert humain à l'aide du logiciel Transcriber et sauvegardée dans un format XML. Ce format inclut des définitions de Thèmes, de Locuteurs, puis des sections thématiques, des tours de parole, la transcription accompagnée d'événements sonores, de synchronisation et des débuts et fins d'entités nommées. L'annotation est synchronisée sur le flux audio en utilisant des attributs et des balises dédiés.*

parole sur musique. Puis la seconde tâche a pour but de segmenter le flux audio en tours de parole sur les changements de locuteurs, tout en identifiant les contributions consécutives d'un même locuteur. Enfin, une tâche d'identification de locuteurs connus parmi les locuteurs précédents complète la tâche de segmentation en locuteurs.

Une dernière tâche d'extraction d'information consiste en l'annotation des entités nommées (noms de personnes, lieux, organisations...). Cette tâche fait l'objet de la section 4.2.

3.2.2 Mesures d'évaluation

La segmentation en classes acoustiques est évaluée selon le nombre de frontières bien classées avec une tolérance de 0.25 seconde. La précision (P) est calculée comme le rapport entre le nombre de frontières correctes et le nombre de frontières de l'hypothèse ; et le rappel (R) comme le rapport entre le nombre de frontières correctes et le

Conditions	Répartition (%)	WER
Studio	61.1	23.5
Téléphone	4.2	24.8
Fond musical	7.6	23.2
Acoustique dégradée	3.6	38.3
Locuteurs non natifs	14.1	34.9
Autre	9.0	36.7
Féminin	27.9	22.9
Masculin	72.0	28.3
Inconnu	0.1	73.7
Total	100.0	26.7

TAB. 3.3: *Répartition des conditions acoustiques et genre des locuteurs dans le corpus de test ESTER. Les taux d'erreur de mots (WER) du système lors de la campagne sont donnés pour illustrer la dégradation des performances selon les conditions (des modèles adaptés sont utilisés en fonction des classes acoustiques détectées).*

nombre de frontières de la référence. La F_β-mesure est une moyenne harmonique du rappel et de la précision. La valeur de β est fixée[2] en fonction de l'application pour donner plus de poids au rappel ou à la précision (équation 3.1).

$$\begin{aligned} P &= \frac{\text{nb}(correct)}{\text{nb}(hyp)} \\ R &= \frac{\text{nb}(correct)}{\text{nb}(ref)} \\ F_\beta &= \frac{(1+\beta^2)PR}{\beta^2 P + R} \end{aligned} \tag{3.1}$$

Le suivi de locuteur est évalué par une F-mesure similaire à celle utilisée pour la segmentation en classes acoustiques. Des mesures de performances alternatives comme les courbes *Detection Error Tradeoff* (DET) sont présentées dans les résultats officiels de la campagne (Galliano et al., 2005).

Le taux d'erreur de segmentation en locuteurs (E_{seg}) est détaillé par l'équation 3.2 dans laquelle : $dur(s)$ représente la durée du segment s ; $nb_{ref}(s)$ est le nombre de locuteurs de la référence parlant dans s ; $nb_{hyp}(s)$ est le nombre de locuteurs de l'hypothèse dans s ; $nb_{correct}(s)$ est le nombre de nombre de locuteur de l'hypothèse parlant réellement dans s.

$$E_{seg} = \frac{\sum_s dur(s)\left(max(\text{nb}_{ref}(s), \text{nb}_{hyp}(s)) - \text{nb}_{correct}(s)\right)}{\sum_s dur(s)\,\text{nb}_{ref}(s)} \tag{3.2}$$

La qualité de la transcription est évaluée selon le taux d'erreur de mots (*Word Error Rate*, WER) explicité dans la formule 3.3.

[2]Dans ESTER, β=1.

$$WER = \frac{I + S + D}{R} \tag{3.3}$$

où I est le nombre d'insertions, S est le nombre de substitutions, D est le nombre de suppressions et R est le nombre de mots de la référence. Les types d'erreurs sont déterminés par alignement dynamique de la référence et de l'hypothèse.

3.2.3 Résultats du système LIA

Les performances de la chaîne de traitement du LIA sont détaillées dans la table 3.4. Les résultats sur chaque tâche sont comparés à la meilleure soumission lors de l'évaluation et aux performances correspondant aux améliorations des systèmes depuis l'évaluation. Il est à noter que des problèmes dans la détection du genre et la stratégie d'utilisation des classes acoustiques ont dégradé les performances en détection du locuteur (Istrate et al., 2005). De plus, le système de transcription a surtout été amélioré au niveau de la vitesse d'exécution (sa vitesse a été multipliée par 20 pour des performances identiques). Les différences entre le système LIA et le meilleur système sont principalement dues à la différence entre la quantité de données d'apprentissage utilisées, la taille de l'espace de recherche (65000 mots contre 200000 mots) et l'expérience dans le domaine de la transcription d'émissions radio.

Tâche	**Perf.**	**Post.**	**Meill.**	**Unité**
Détection de parole	99.2	-	99.2	f_1-m
Détection de parole sur musique	92.7	-	94.2	f_1-m
Détection de musique	54.8	-	54.8	f_1-m
Segmentation en locuteurs	19.2	-	11.5	%err
Suivi de locuteurs	66.0	75.5	84.3	f_1-m
Transcription	26.7	22.7	11.9	WER
Transcription (temps limité)	36.3	-	16.8	WER

TAB. 3.4: *Résultats du LIA sur les différentes tâches d'ESTER phase II (Perf.), post-évaluation (Post.) et du meilleur participant lors de l'évaluation (Meill.) selon l'unité correspondante : le taux d'erreur de mots (WER) pour la transcription, la f_1-mesure et le taux d'erreur de segmentation (%err) pour les tâches de segmentation.*

Une analyse des erreurs de transcription fait ressortir qu'une grande quantité d'entre elles provient de noms propres mal reconnus et d'homophonies dont la résolution est hors de portée des modèles tri-grammes. Des exemples extraits de la soumission illustrent cette observation :

Réf. : les grands titres de l' actualité **** **Maude Bailleux** bonjour
Hyp. : les grands titres de l' actualité **émaux de Bayeux** bonjour

Réf. : Nicolas **** **Pierron signait** la troisième édition de **Classique** Matin
Hyp. : Nicolas **Pierre ont signé** la troisième édition de **Classiques** Matin

L'amélioration des systèmes de structuration audio passe d'abord par une augmentation de la quantité de données d'apprentissage afin de couvrir un maximum d'événements mais aussi pour mieux estimer les paramètres des algorithmes statistiques employés. L'intégration des objectifs de la tâche finale et une analyse fine des erreurs donneront les voies de recherche à privilégier.

3.3 Conclusion

L'extraction de descripteurs sémantiques est indispensable pour la construction de résumés parlés. Nous avons présenté dans ce chapitre les différentes tâches impliquées dans une chaîne de structuration des données acoustiques. Les méthodes les plus répandues pour la résolution de ces tâches sont fondées sur l'apprentissage artificiel, formulées comme des problèmes de segmentation et d'identification. Leurs performances sont directement liées à la quantité de données d'apprentissage et à l'adéquation de ces dernières aux conditions d'utilisation. La chaîne de structuration présentée, développée au LIA, est une concrétisation de ces différentes tâches. Ses performances, validées sur la campagne ESTER sont suffisantes pour envisager l'emploi des descripteurs sémantiques et structuraux ainsi extraits dans une méthode de résumé automatique de parole. Toutefois, des éléments complémentaires à cette chaîne et indispensables pour le résumé sont présentés dans le chapitre suivant.

Chapitre 4

Compléments à l'extraction de descripteurs structurels et sémantiques

Sommaire

Le chapitre 2 a présenté les nombreuses méthodes de recherche d'information parlée et il s'avère que ce type d'information nécessite une extraction spécifique de descripteurs sémantiques à partir de l'acoustique. Par la suite, le chapitre 3 a introduit les différentes tâches de structuration et leur mise en œuvre dans la chaîne de structuration Speeral. L'objectif de ces travaux est de faciliter l'accès à l'information audio à l'aide du résumé de parole et les éléments de structuration présentés au chapitre précédent ne sont pas suffisants pour obtenir un résumé de qualité. Nous nous concentrons maintenant sur la présentation de deux compléments à la structuration pour le résumé automatique de parole. Tout d'abord, une segmentation en phrases de qualité est nécessaire pour résumer la parole avec une approche par extraction. En effet, du point de vue de l'utilisateur, cet aspect de la forme d'un résumé parlé est déterminant car une coupure inopportune au milieu d'une phrase peut fortement dégrader la compréhension. La méthode proposée pour la segmentation en phrases s'appuie sur un étiquetage

de séquence dans le cadre des *Conditional Random Fields* (section 4.1). Le second point de contribution réside dans l'extraction d'entités nommées dans le flux de parole. Ces entités liées au domaine (personnes, organisations, lieux...) dirigent la projection dans l'espace sémantique lors de la génération du résumé. L'approche développée pour cette tâche consiste en une recherche des entités nommées dans l'ensemble des hypothèses de transcription au lieu d'être restreinte à la meilleure hypothèse (section 4.2).

4.1 Segmentation en phrases par étiquetage de séquence

Il a été remarqué dans la section 2.2 que la segmentation en phrases demandait une attention particulière dans le cadre du résumé de parole par extraction (Rappelons que Mrozinski et al. (2006) ont observé une forte réduction de la qualité des résumés de parole fondés sur une segmentation automatique par rapport à une segmentation manuelle).

Dans la littérature, le problème de segmentation en phrases est généralement reformulé en un problème d'identification de frontières de phrases (étiquetage de séquence). La transcription automatique est employée pour générer une suite de mots et des frontières (événement binaire B) sont recherchées entre les mots. La décision est généralement issue d'une combinaison de paramètres prosodiques (événement S) et linguistiques (événement L). Trouver des frontières de phrases est loin d'être facile, en attestent par exemple Stevenson et Gaizauskas (2000), qui évaluent les performances d'annotateurs humains sur la reponctuation d'un texte, et qui observent qu'il est beaucoup plus facile de reponctuer un flux de mots contenant les majuscules d'origine (F_1-mesure de 0.95) qu'en l'absence de ces marqueurs (F_1-mesure de 0.80), comme dans le cas d'une transcription automatique.

La majorité des approches sont fondées sur des modèles probabilistes tentant de prédire la séquence B en fonction de S et L. Gotoh et Renals (2000) constituent un modèle pour chacune des modalités (S et L) sur des ensembles de données séparés. La probabilité linguistique $P(B, L)$ qu'une frontière de phrase précède un mot est modélisée à partir de données textuelles disponibles en masse ; l'implication de la prosodie $P(B, S)$ est modélisée à partir des durées de pauses sur un corpus acoustique de plus petite taille. Les deux modèles sont fusionnés grâce à une heuristique[1]. Shriberg et al. (2000) étudient les différentes caractéristiques prosodiques en profondeur : les pauses, le rythme phonétique ou syllabique, la pente de fréquence fondamentale (f_0) et sa continuité, les sauts de f_0, l'écart à la moyenne de la f_0, et la qualité de voix. Les valeurs sont fonction du locuteur ou d'un locuteur moyen lorsque les données sont insuffisantes. En plus de ces paramètres, la décision repose sur la durée des phrases et les changements de locuteurs (segmentation manuelle en locuteurs). Un arbre de décision donne une sélection des paramètres les plus pertinents et ces derniers servent à construire un modèle de séquence génératif. Les paramètres les plus efficaces sur des données radio-diffusées semblent être les pauses et les changements de locuteurs. Liu et al. (2005) continuent ces

[1] $P(B, L, S) \simeq P(B, S)^{\alpha} P(B, L)$, $\alpha > 10$ donnant les meilleurs résultats.

travaux en comparant des approches HMM, maximum d'entropie et CRF pour l'étiquetage de la séquence : ce dernier modèle s'avérant être le plus efficace (une fusion des trois apporte un gain complémentaire). Il est intéressant de noter que la décision prosodique sur la frontière est prise avant l'inclusion dans le modèle de séquence. Des travaux similaires de (Kim et al., 2004) intègrent des arbres de décision avec un système de détection de difficultés de prononciation.

La tâche de détection de frontières de phrase (en anglais, *Sentence Unit Boundary Detection*, SUBD) a été évaluée lors des éditions 2002 à 2004 des campagnes *Rich Transcription* « automne » (RT-fall), organisées par NIST. Les données de référence reposent sur un guide d'annotation (Strassel, 2003)[2] précisant que la notion de phrase à l'oral (nommée « unité syntagmatique ») est différente de l'écrit. Les différences sont avant tout grammaticales ; les unités sont classées selon leur type (déclarations, questions, éléments phatiques et unités incomplètes). La mesure de performance NIST est le taux d'erreur sur les frontières (nombre de frontières oubliées, ajoutées ou de mauvais type, divisé par le nombre de frontières dans la référence : équation 4.1 dans laquelle $\mathrm{nb}(\cdot)$ est le cardinal d'un ensemble de frontières).

$$SB_{err} = \frac{\mathrm{nb}(\text{oubli}) + \mathrm{nb}(\text{ajout}) + \mathrm{nb}(\text{mauvais type})}{\mathrm{nb}(\text{référence})} \tag{4.1}$$

Sur des données radio-diffusées, Liu et al. (2005) aboutissent à un taux d'erreur de 0.54 (sans prendre en compte les erreurs de type). Cette valeur correspond à une F_1-mesure d'environ 0.70, proche des performances annoncées par les autres auteurs.

La détection de frontières de phrases que nous avons mise en place pour le résumé de parole est similaire à l'approche de Liu et al. (2005). En restant dans le cadre de l'étiquetage bi-classe de la séquence de mots, nous appliquons un modèle CRF sur des caractéristiques prosodiques et linguistiques. Ces dernières sont issues de la chaîne de structuration Speeral. Les frontières de phrases sont recherchées dans les émissions de radio en français de la campagne ESTER.

4.1.1 Conditional Random Fields

Définition

Conditional Random Fields (CRF, Lafferty et al., 2001) est un cadre probabiliste discriminant pour l'étiquetage de séquences. Au lieu de modéliser la probabilité jointe d'apparition des séquences d'observation et des séquences d'étiquettes comme le fait une approche générative telle que HMM, CRF repose sur la probabilité conditionnelle de l'étiquetage sachant l'ensemble de la séquence. Les méthodes à maximum d'entropie de Markov (MEMM) recherchent aussi à maximiser cette probabilité conditionnelle, mais de façon locale. Ceci pose des problèmes au niveau des hypothèses partielles débouchant sur un petit nombre de successeurs car ils sont systématiquement préférés

[2] disponible en ligne sur http://projects.ldc.upenn.edu/MDE/Guidelines/SimpleMDE_V5.0.pdf, visité en novembre 2006

aux chemins de plus grande entropie. Cet effet est décrit sous le nom d'effet du biais des étiquettes par Lafferty et al. (2001).

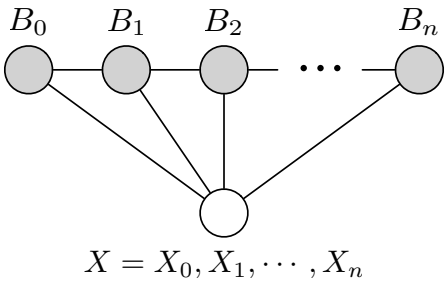

FIG. 4.1: *Détection de frontières de phrases par modélisation CRF. La séquence d'événements représentant la présence ou l'absence de frontières ($B = B_0, \cdots, B_n$) est globalement conditionnée par la séquence d'observations phonétiques et linguistiques ($X = X_0, \cdots, X_n$).*

Appliquons CRF à une tâche de segmentation en phrases : B est une séquence d'étiquettes ($B = 1$ pour une frontière de phrase, $B = 0$ pour une absence de frontière) ; X est une séquence d'observations prosodiques et linguistiques. Le modèle conditionne la séquence B sur l'ensemble de la séquence X (figure 4.1). La meilleure hypothèse d'étiquetage est celle qui maximise la probabilité $P(B|X)$. Cette probabilité est estimée par une distribution de forme exponentielle satisfaisant des caractéristiques sur des données d'apprentissage (équation 4.2).

$$\begin{aligned} \hat{B} &\operatorname*{argmax}_B P(B|X) \\ P(B|X) &\simeq \frac{1}{Z(X)} e^{\sum_k \lambda_k f_k(B,X)} \\ Z(X) &= \sum_B e^{\sum_k \lambda_k f_k(B,X)} \\ f_k(B,X) &\geq 0 \end{aligned} \tag{4.2}$$

Dans cette équation, les λ_k sont les paramètres du modèle ; $Z(X)$ sert à la normalisation de la distribution ; $f_k(B,X)$ sont les fonctions caractéristiques sur les arcs et les sommets du modèle graphique associé au problème. Ces fonctions sont des relations entres les B_i et X et entre des B_i voisins.

L'inférence des paramètres λ_i se fait par maximisation de la vraisemblance conditionnelle sur un ensemble de données étiquetées. Le maximum de cette fonction log-concave est découvert par des méthodes de maximisation classiques, comme *Generalized Iterative Scaling* (GIS, Darroch et Ratcliff, 1972), *Improved Iterative Scaling* (IIS, Della Pietra et al., 1997), ou *Limited-memory Broyden-Fletcher-Goldfarb-Shanno* (LBFGS, Liu et Nocedal, 1989), qui s'avère être la plus rapide. Ces méthodes sont comparées dans (Malouf, 2002). La dépendance des étiquettes sur l'ensemble de la séquence d'observation rend l'apprentissage beaucoup plus coûteux que pour un maximum d'entropie local classique. L'étiquetage d'une séquence nouvelle se fait par programmation dynamique.

La boîte à outils CRF++

L'ensemble de nos expériences sur la détection de frontières de phrases repose sur CRF++[3], une boîte à outils pour l'étiquetage de séquences fondée sur CRF. CRF++ implémente un apprentissage dont l'optimisation repose sur une méthode de quasi-Newton (LBFGS) et un décodage grâce à l'algorithme Viterbi. Cette boîte à outils a été utilisée avec succès pour de nombreuses tâches de traitement automatique du langage naturel comme la désambiguïsation sémantique, la décomposition en groupes grammaticaux, l'étiquetage morpho-syntaxique ou encore la détection d'entités nommées (Kudo et al., 2004).

4.1.2 Traits acoustiques et linguistiques

Nous suivons les approches classiques pour la segmentation en phrases en recherchant des frontières potentielles uniquement entre les mots et en fixant l'événement $B = 1$ si une frontière a précédé un mot et $B = 0$ dans le cas contraire. La prédiction de la présence d'une frontière de phrase avant un mot dépend de caractéristiques linguistiques et acoustiques que nous allons décrire (voir table 4.1). Au niveau linguistique, les mots et leurs catégories morpho-syntaxiques modélisent les phénomènes grammaticaux de la séquence. La catégorie morpho-syntaxique des mots est trouvée grâce à lia_tagg[4]. Cet étiqueteur repose sur un dictionnaire d'étiquettes possibles par mots et effectue l'étiquetage dans un cadre HMM. Alors que certains couples syntaxiques, comme « le déterminant et le nom », qui ne doivent pas être séparés par une frontière de phrase, sont plutôt bien capturés par cette modélisation, d'autres groupes comme « le verbe et son complément » sont moins faciles à détecter sans une modélisation plus approfondie de la grammaire. Si les éléments linguistiques sont utiles pour reponctuer un texte, ils peuvent être faussés par les erreurs de transcription, d'étiquetage morpho-syntaxique et le manque relatif de grammaire de la langue parlée. Pour y remédier, il faut associer des caractères acoustiques aux indices linguistiques, comme les changements de locuteur et quelques éléments de prosodie. Les changements de locuteurs sont issus, comme la séquence de mots, de la chaîne de transcription et employés tels quels sans prendre en compte les identités retrouvées. En terme de prosodie, les pauses sont explorées à deux niveaux : avant le mot et à l'intérieur du mot pour essayer d'éviter de prendre les hésitations pour des fins de phrase. De plus, comme il est difficile de profiter des informations apportées par la courbe de fréquence fondamentale ($f0$), nous utilisons seulement sa pente globale, sur trois horizons temporels différents (le mot, une fenêtre allant de 4 secondes avant le début du mot jusqu'à sa fin et une fenêtre allant de 8 secondes avant le début du mot jusqu'à sa fin). Bien que cette approche ne soit pas optimale, elle permet tout de même de modéliser les grands phénomènes macro-prosodiques de la phrase. Toutefois, certaines caractéristiques sont perdues, comme les effets du rythme prosodique ou syllabique connus pour ralentir en fin de phrase. Les

[3]Disponible sur http://chasen.org/~taku/software/CRF++, visité en août 2006.

[4]Étiqueteur morpho-syntaxiaque du LIA, disponible sous licence GPL, sur http://www.univ-avignon.fr/chercheurs/bechet/download_fred.html, visité en octobre 2006.

frontières des mots de la référence sont extraites grâce à un alignement forcé sur le signal en utilisant un outil dérivé du système de transcription (gvalign).

Mot	Étiquette	P1	P2	Loc.	F1	F2	F3	Ponct.
avait	V3S	C0	C0	SPK	C0	C0	C0	point
le	DETMS	C0	C0	n	C0	C0	C0	x
salut	NMS	C0	C4	n	C4	C0	C0	x
à	XSOC	C0	C0	n	C0	C8	C3	x
tous	AINDMP	C0	C0	n	C4	C0	C0	x
ceux	PDEMMP	C0	C0	n	C5	C4	C0	x
en	PREP	C0	C0	n	C5	C5	C0	x
bonne	AFS	C0	C0	n	C2	C4	C0	point
journée	NFS	C0	C0	n	C5	C3	C0	x
euh	ADV	C0	C0	n	C4	C8	C1	x

TAB. 4.1: *Exemple des paramètres extraits pour la segmentation en phrases. Au niveau linguistique : le mot de la transcription et son étiquette morpho-syntaxique. Au niveau prosodique : la durée de pause avant le mot (P1) et à l'intérieur du mot (P2), un éventuel changement de locuteur avant le mot (Loc.), la pente de F0 à divers horizons temporels (F1=le mot, F2=-4s, F3=-8s). La ponctuation qui précède le mot est prédite grâce à ces paramètres (Ponct.). Les valeurs numériques sont quantifiées uniformément selon les classes C0 à C9 (sur une fenêtre glissante de 300 valeurs, avec un jeu de classes par paramètre).*

CRF++ facilite la génération des fonctions caractéristiques en utilisant des patrons de conjonction d'événements de X et B. Dans notre implémentation, une frontière de phrase potentielle est conditionnée par des séquences n-grammes de chaque type de caractères linguistiques et acoustiques autour du mot à étiqueter et par la conjonction des séquences précédentes (illustrées par la figure 4.2). La boîte à outils est cependant limitée dans sa version actuelle à des caractéristiques symboliques. Cette limitation implique la quantification des valeurs continues comme la durée des pauses ou la pente de fréquence fondamentale. La quantification se fait sur une fenêtre glissante en utilisant une répartition uniforme en n classes[5]. Cette approche permet de normaliser les valeurs lors de changements de locuteurs et d'environnement.

4.1.3 Performances

Les performances en segmentation en phrases sont calculées sur la base du nombre de frontières bien placées par rapport au nombre de frontières erronées, en rappel, précision, et f-mesure (un exemple est donné par la table 4.2). Les expériences sont réalisées sur le corpus ESTER qui n'a malheureusement pas fait l'objet de directives d'annotation pour les frontières de phrases. Le guide d'annotation précise que « la ponctuation est facultative, mais peut être utilisée pour faciliter la tâche de transcription ». Cette dernière varie donc beaucoup d'un annotateur à l'autre ; les phrases peuvent être

[5] $n = 10$ dans les expériences qui suivent. La fenêtre glissante fait 300 valeurs. Ces valeurs sont fixées empiriquement, mais ne semblent pas avoir un impact important sur les performances.

Mot	Étiquette	P1	P2	Loc.	F1	F2	F3	Ponct.
avait	V3S	C0	C0	SPK	C0	C0	C0	point
le	DETMS	C0	C0	n	C0	C0	C0	x
salut	NMS	C0	C4	n	C4	C0	C0	x
à	XSOC	C0	C0	n	C0	C8	C3	x
tous	AINDMP	C0	C0	n	C4	C0	C0	x
ceux	PDEMMP	C0	C0	n	C5	C4	C0	x
en	PREP	C0	C0	n	C5	C5	C0	x
bonne	AFS	C0	C0	n	C2	C4	C0	point
journée	NFS	C0	C0	n	C5	C3	C0	x
euh	ADV	C0	C0	n	C4	C8	C1	x

FIG. 4.2: *Illustration des groupes de paramètres utilisés pour la prédiction de la présence ou absence d'une frontière de phrase. En plus de ces événements, le modèle prend en compte les unigrammes dans une fenêtre de deux mots autour du mot courant et la conjonction de chacun des événements précédents sur l'ensemble de la séquence. Les données sont celles de la figure 4.1*

très longues, jusqu'à faire un tour de parole complet, contenant un grand nombre de virgules, alors que dans d'autres cas, chaque pause du locuteur a été annotée par une fin de phrase. Ce problème de fiabilité du corpus implique une nécessaire prudence dans l'interprétation des résultats d'évaluation.

Référence	*	*	*	p	*	p	*	*	*	*	*	p
Hypothèse	*	p	*	*	*	p	*	*	*	p	*	p

TAB. 4.2: *Performances de la segmentation en phrases pour un exemple fictif. « p » représente une frontière de phrase et « * » une absence de frontière. Il y a 3 frontières à trouver dans la référence, 4 frontières ont été trouvées dans l'hypothèse, dont 2 bien placées. La précision est de $P = 2/4 = 0.5$, le rappel est de $R = 2/3 = 0.66$ et la F_1-mesure est de $F_1 = 2 * PR/(P + R) = 0.57$. Le taux d'erreur NIST est égal au nombre d'erreurs ($hyp_i \neq ref_i$) par rapport au nombre de frontères à trouver : $SB_{err} = 3/3 = 100\%$.*

Il est intéressant de noter que les journalistes des radios francophones du corpus ont tendance à utiliser une architecture prosodique très spéciale qui détériore la cohérence des événements caractérisant une frontière de phrase. En effet, pour captiver l'attention de l'auditeur, les journalistes reprennent leur souffle en milieu de phrase, pour provoquer un effet « d'attente ». Cet effet diminue la cohérence de l'annotation par l'insertion de pauses. Ces pauses ont les caractéristiques acoustiques d'une fin de phrase et les caractéristiques linguistiques d'un milieu de phrase.

Comparatif en structuration automatique et manuelle

Les données d'entraînement utilisées dans ces expériences correspondent aux 80 heures d'entraînement (environ 874000 mots) du corpus ESTER, alors que les performances sont rapportées pour les 10 heures de la partie développement du corpus (en-

Données	Rappel	Précision	F_1-mesure
Étiquetage : points			
M+M	0.42	0.80	0.55
M+A	0.34	0.84	0.49
A+M	0.62	0.74	0.67
A+A	0.61	0.77	0.68
Étiquetage : points et virgules			
M+M	0.41	0.64	0.50
M+A	0.30	0.72	0.42
A+M	0.50	0.65	0.56
A+A	0.49	0.74	0.59
Étiquetage : points et virgules fusionnés			
M+M	0.55	0.78	0.64
M+A	0.37	0.81	0.51
A+M	0.59	0.70	0.64
A+A	0.55	0.81	0.66

TAB. 4.3: *Performances de la segmentation en phrases selon le type d'étiquetage recherché et les données utilisées en apprentissage et en test. Par exemple, « M+A » représente un apprentissage sur les données extraites à la main (M) et un test sur les données transcrites et segmentées automatiquement (A).*

viron 88000 mots). La table 4.3 présente des comparatifs entre l'utilisation de données structurées automatiquement ou manuellement en apprentissage et en test, pour les tâches d'étiquetage sur les points («.») comme frontières de phrases, les points et les virgules sous forme d'un problème 3-classes («.», «,» et ∅) et la fusion des points et des virgules («,»=«.»).

Globalement, les tests sur les données structurées manuellement montrent que l'approche admet un faible rappel et une forte précision sur les frontières retrouvées. La différence est moins prononcée lors de l'utilisation de données structurées automatiquement lors du test. De plus, la méthode la plus performante consiste en l'utilisation de données structurées automatiquement en apprentissage et en test. En revanche, les données de référence mènent à de moins bonnes performances générales. Il semblerait que ceci soit dû à une différence dans la notion de pause entre l'algorithme d'alignement automatique et le système de transcription. Pour ce qui est des différents étiquetages possibles, étant donné qu'aucun guide d'annotation en frontières de phrases n'a été fourni lors de la création des données de référence, nous avons essayé de réduire les incohérences virgule-point en fusionnant ces 2 types de frontières et en les annotant séparément. Les performances ne sont néanmoins jamais au niveau de celles obtenues par l'annotation des « points ».

Finalement, nous déduisons de ces résultats que l'approche permet d'établir des performances de l'ordre de ce qui est donné dans la littérature (une f_1-mesure d'environ 0.70). De plus, il est bon de noter que la méthode a une bonne précision et une tendance à sous-générer les frontières de phrases. Ce type de comportement est béné-

fique pour le résumé automatique car le type d'erreur le plus pénalisant dans ce cadre reste l'insertion de frontières de phrases là où elles n'ont pas lieu d'être.

4.1.4 Améliorations envisagées

Nous avons proposé une détection des frontières de phrases par étiquetage d'une séquence d'« inter-mots » à l'aide de CRF. L'approche peut être améliorée en utilisant des caractéristiques continues (et non symboliques) — en prenant en compte les scores de confiance du système de transcription — et en calculant une courbe de $f0$ plus fine, normalisée pour chaque locuteur. Une des limitations de CRF est que cette approche ne peut tenir compte de paramètres au niveau global de la phrase, comme sa longueur ou sa cohérence syntaxique et sémantique. Une solution à ce problème peut être semi-CRF (Sarawagi et Cohen, 2005) qui tente de remettre en cause de l'hypothèse Markovienne[6] du processus sur un segment temporel de taille raisonnable de l'ordre de la phrase. D'autres pistes doivent être envisagées, comme un test sur la fiabilité du corpus afin de détecter et d'écarter les phrases mal annotées, ou une intégration complète de la segmentation en phrases dans la transcription du contenu parlé pour retarder la prise de décision sur les frontières de mots.

4.2 Extraction d'entités nommées dans le flux de parole

Les entités nommées sont des entités du monde « réel », dont la forme linguistique est une représentation directe dénuée d'ambiguïté. Notamment, lorsqu'une de ces entités se retrouve dans le discours de plusieurs personnes, il est considéré que ces différentes références ont le même antécédent. Bien que cette affirmation soit loin d'être vraie dans le cas général, les types d'entités recherchés doivent s'en approcher le plus possible. Par exemple, « une table » est un concept qui se réfère à un objet dans un contexte donné. Dans un autre contexte, le locuteur se référera généralement à une autre entité. En revanche, dans un domaine journalistique, les noms propres se réfèrent à des objets considérés comme uniques, dont la forme linguistique peut être séparée de son contexte sans rendre la référence ambiguë. Ce type de comportement est très intéressant dans le cadre de l'analyse sémantique indispensable pour le résumé car la projection depuis la linguistique devient transparente.

Dans le cadre de l'extraction de descripteurs sémantique de journaux radio diffusés, les entité nommées sont étendues à certaines quantités fortement porteuses d'information dans ce domaine. Les entités recherchées sont de deux types : entités uniques basées sur des noms propres (personnes, lieux, organisations...) et entités basées sur des séquences de noms communs (dates, quantités monétaires, distances...). Les majuscules des noms propres sont de bons indicateurs de la présence d'entités du premier type et les valeurs numériques sont de bons indicateurs de la présence du second type d'entité.

[6] L'hypothèse Markovienne est vérifiée pour un processus si et seulement si la distribution conditionnelle de probabilité des états futurs, étant donné l'instant présent, ne dépend que de ce même état présent et pas des états passés.