Indexation en locuteur : utilisation d'informations lexicales

J. Mauclair, S. Meignier, Y. Estève

LIUM, Université du Maine Le Mans, France
{julie.mauclair,sylvain.meignier,yannick.esteve}@lium.univ-lemans.fr

ABSTRACT

The automatic speaker indexing consists in splitting the signal into homogeneous segments and clustering them by speakers. However the speaker segments are specified with anonymous labels. This paper propose to identify those speakers by extracting their full names pronounced in the show. With a semantic classification tree, the full names detected in the segment transcription are associated to this segment or to one of its neighbors. Then, a merging method associates a full name to a speaker cluster instead of the anonymous label. The experiments are carried out over French broadcast news from the ESTER 2005 evaluation campaign. About 70% show duration is correctly processed for evaluation corpus.

1 Introduction

Les transcriptions manuelles d'enregistrements audio sont très coûteuses, particulièrement lorsqu'on cherche à indexer des informations spécifiques telles que le thème principal, les mots-clés, le nom du locuteur... Seules les méthodes automatiques génèrent des transcriptions enrichies à moindre coût, mais leur taux d'erreur doit être suffisamment faible pour pouvoir les exploiter. Ici, nous nous intéressons uniquement au problème de l'identité du locuteur.
La première étape pour obtenir automatiquement des transcriptions enrichies consiste à segmenter le signal puis à regrouper les segments en locuteur. Les principales méthodes reposent uniquement sur des paramètres acoustiques [1, 2]. L'étape suivante consiste à transcrire automatiquement les segments résultants.
L'indexation attribue uniquement des étiquettes anonymes aux segments, alors que connaître la véritable identité du locuteur serait judicieux pour la recherche documentaire de documents sonores. Les méthodes existantes permettant d'associer le nom complet (prénom, nom) d'un locuteur aux segments issus de l'indexation sont de deux sortes. Les premières méthodes reposent sur le traitement d'informations purement acoustiques, avec généralement l'utilisation d'un système de reconnaissance automatique du locuteur requérant des échantillons de voix pour apprendre les modèles [3]. Les autres reposent sur des informations lexicales et extraient l'identité du locuteur directement à partir la transcription de l'émission. Le nom du locuteur et sa localisation sont présents dans les mots prononcés durant une émission de radio et ces informations peuvent être utilisées pour identifier le locuteur avec son véritable nom. Aucun échantillon de voix n'est ici nécessaire. En effet, les intervenants se présentent ou présentent l'intervenant suivant, ils félicitent le précédant ou le suivant, concluent le reportage par leur nom ... Dans de récents travaux menés sur des émissions radiophoniques en anglais [4], le LIMSI utilise des règles linguistiques extraites manuellement pour identifier le locuteur du segment avec son véritable nom. En fonction des annonces de qui parle, de qui parlera ou de qui vient de parler, un nom détecté dans la transcription permet d'étiqueter avec la véritable identité du locuteur le segment courant, le suivant ou le précédent. Le taux d'erreur de ce processus basé sur des règles manuelles est d'environ 13% à partir de transcriptions manuelles (18% à partir de transcriptions automatiques).
Dans ce papier, nous proposons une association automatique du locuteur avec son nom complet grâce à l'utilisation d'un arbre de classification sémantique qui apprend automatiquement ce genre de règles. Cette méthode fournit seulement une décision locale pour le segment courant et les segments contigüs. L'identité du locuteur est ensuite propagée sur la totalité de l'émission de radio.
Pour réaliser l'étude préliminaire présentée ici, nous utilisons en entrée pour le système les indexations en locuteur ainsi que les transcriptions manuelles de référence.

Nous gardons à l'esprit que les erreurs provenant de l'indexation et de la transcription automatique réduisent les performances (voir résultats de [4]). Les corpus proviennent de la campagne d'évaluation française ESTER [5]. Cependant, le processus entièrement automatique utilisé nous permet d'adapter facilement la méthode à d'autre langues.

2 Informations sur le locuteur

2.1 Identité cliente

Les participant à des émissions radiophoniques sont principalement des personnes publiques comme des journalistes, des politiciens, des artistes ou des sportifs. Cette population est facilement identifiable : leurs noms et prénoms sont bien connus, ils sont présents dans plusieurs émissions, et ils correspondent aux locuteurs principaux en termes de temps de parole. Ces locuteurs sont identifiés par la concaténation de leurs noms et prénoms dans les transcriptions d'ESTER. Ce sont les locuteurs à identifier dans la tâche proposée (locuteurs clients). 1007 noms complets différents ont été extraits des corpus utilisés dans nos expériences en ne conservant que les noms des personnes publiques. Le procédé de détection du nom du locuteur repose sur cette liste fermée. Nous avons choisi d'employer le nom complet pour éviter les fausses détections introduites par la méthode de détection : l'ambiguïté présentée par l'utilisation de noms partiels (seulement le prénom ou le nom) amène des problèmes que nous ne résoudrons pas ici.

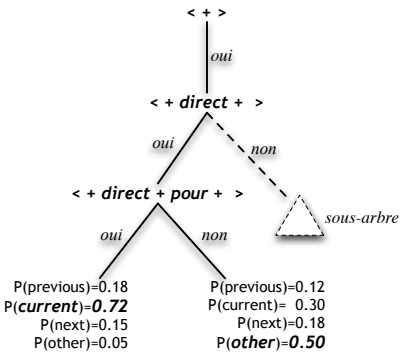

Figure 2: Exemple d'une partie d'un arbre de classification sémantique : à chaque feuille, une probabilité est associée à chaque étiquette.

2.2 Étiquetage des occurrences de noms

Un nom complet détecté dans un segment peut être associé à une des 4 étiquettes suivantes : *current*, *next*, *previous* et *other*. Elles sont attribuées respectivement si le nom détecté semble se rapporter au locuteur du segment de parole courant, du segment suivant ou du segment précédent (voir figure 1). Si ce n'est pas le cas, l'étiquette *other* est attribuée au nom détecté.

3 Méthode employée

Pour associer un nom complet à une étiquette anonyme issue de l'indexation en locuteur (figure 1, partie ①), nous proposons les deux étapes suivantes :
1- **Analyse du contexte lexical pour chaque un nom complet** (figure 1, partie ②) : cette étape traite chaque nom complet détecté dans la transcription d'un segment de parole. Elle détermine si ce nom se rapporte au locuteur précédent, courant, suivant ou à un autre locuteur. Seuls les segments proches d'un nom détecté dans la transcription peuvent être associés à ce nom. D'ailleurs, des segments peuvent être associés à différents noms : les processus d'association sont faits sans coopération et peuvent fournir des résultats antagonistes sur un même segment.
2- **Dénomination du locuteur** (figure 1, partie ③) : la deuxième étape consiste à fusionner les hypothèses précédentes afin d'associer un nom complet à une étiquette anonyme d'un locuteur et de répercuter ce nom à tous les segments étiquetés avec cette même étiquette anonyme. (figure 1, partie ④).

3.1 Analyse du contexte lexical

Quand un nom est détecté, le contexte lexical de la transcription est analysé pour associer à ce nom l'étiquette la plus pertinente. Cette analyse est faite en employant un arbre de classification sémantique (SCT) [6].

Arbre de classification sémantique Les SCTs reposent sur l'utilisation d'expressions régulières. Des couples composés d'une occurrence de nom complet et de son contexte lexical sont classés en utilisant ces expressions régulières. Notre but est de classer ces couples parmi les 4 étiquettes *previous*, *current*, *next* et *other* Pour un nom complet et son contexte lexical, chaque feuille de l'arbre peut donner une probabilité pour chaque étiquette possible. La figure 2 donne un exemple d'expression régulière utilisée pour construire l'arbre.

Décisions locales Pour une occurrence o de nom complet détectée dans un contexte lexical $W_s(o)$ (voir section 4.1) associé à un segment de parole s, un SCT peut donner la probabilité $P(t|W_s(o))$ pour chaque étiquette possible t de l'ensemble des étiquettes $T = \{previous, current, next, other\}$. Soit l'étiquette $\delta(o) \in T$ associée à une occurrence d'un nom complet du segment de parole s telle que :

$$\delta(o) = \underset{t}{\operatorname{argmax}}\, P(t|W_s(o))$$

Dans notre approche, parmi les 4 étiquettes possibles pour $W_s(o)$, seule l'étiquette $\delta(o)$ est prise en considération pour la suite du processus. En outre, si plus d'une étiquette obtient une probabilité égale à $\max_t P(t|W_s(o))$, aucune décision locale n'est prise.
Définissons la valeur $\Gamma(o)$ qui servira par la suite telle que $\Gamma(o) = P(\delta(o)|W_s(o))$.

3.2 Dénomination du locuteur

Soit ψ un locuteur anonyme d'un segment de parole : il s'agit de trouver le vrai nom $N(\psi)$ de ce locuteur. Chaque segment de parole est associé à un locuteur anonyme (par exemple dans la figure 1, le segment 1 est associé à SPK1, ainsi que les segments 9 et 11). De plus, en utilisant un arbre de classification sémantique sur les noms complets détectés dans la transcription des segments, on obtient une liste de noms correspondant aux locuteurs possibles pour quelques segments (figure 1, partie ②).

Fusion des décisions prises par le SCT Soit K, l'ensemble de tous les noms complets des locuteurs clients. Soit ν_ψ, l'ensemble des différents noms complets associés à au moins un segment prononcé par ψ grâce à une décision locale du SCT : ν_ψ est la liste des noms complets candidats pour ψ et on a $\nu_\psi \subset K$. Soit la fonction $\nu(o)$ qui associe une occurrence de nom complet o à ce nom complet n. Enfin, soit l'ensemble Ω_ψ des occurrences o qui réfèrent aux segments prononcés par ψ grâce aux décisions locales prises par le SCT.
Pour trouver le nom complet $N(\psi)$ d'un locuteur ψ, nous proposons la formule suivante :

$$N(\psi) = \underset{n\in K}{\operatorname{argmax}} \frac{\sum_{(\nu(o)=n)\wedge(o\in\Omega_\psi)} \Gamma(o)}{\sum_{o\in\Omega_\psi} \Gamma(o)} \quad (1)$$

$$= \underset{n\in K}{\operatorname{argmax}} \sum_{(\nu(o)=n)\wedge(o\in\Omega_\psi)} \Gamma(o) \quad (2)$$

Ainsi, le nom complet qui sera associé à une étiquette anonyme de locuteur est le nom dont les occurrences maximisent la somme des scores données par le SCT quand ces occurrences se rapportent à des segments associés à cette étiquette. Comme expliqué dans la section 3.1, seules les valeurs associées aux décisions locales valides sont conservées.

4 Expériences et résultats

4.1 Données

Corpora Les méthodes proposées sont développées et évaluées avec les données de la campagne ESTER 2005. ESTER est une campagne d'évaluation sur les systèmes de transcription d'émissions radiophoniques en Français [5]. Les données (Table 1) comportent six radios différentes dont les émissions durent de 10 à 60 min et sont décomposées en 3 corpora. Le corpus d'apprentissage (*Train*) contient 81h de données

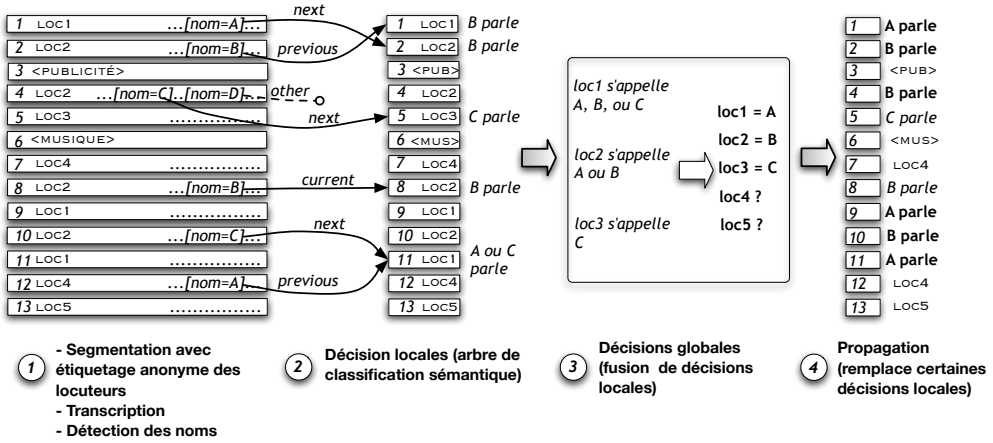

Figure 1: Identification du locuteur

Table 1: Détails sur les corpus : Apprentissage, Développement & Évaluation provenant de la campagne d'évaluation ESTER et statistiques sur les $\neq$ étiquettes applicables aux noms.

	Train	***Dev***	***Eva***
# Radios	5	5	6
durée (h)	81	12.5	10
# Segments	8547	2294	1417
Previous (%)	14.3	12.6	18.6
Current (%)	7.2	7.1	5.3
Next (%)	46.0	45.3	49.3
Other (%)	32.5	35.0	26.8

(8547 segments) dans lesquels 3297 noms complets sont détectés. Le corpus de développement (*Dev*) [1] (*Dev*) contient 12,5h (2294 segments) et 920 noms complets. Le corpus d'évaluation (*Eva*) contient 10h (1417 segments) et 507 noms complets, il correspond au corpus d'évaluation officiel d'ESTER.Le tableau 1 montre également les probabilités *a priori* des 4 étiquettes sur ces corpus.

Préparation des différents corpus Les références (transcriptions enrichies) doivent être transformées et adaptées pour être utilisées par un arbre de classification sémantique et également pour évaluer les résultats expérimentaux. Les adaptations effectuées sont :
• La définition des 4 étiquettes de noms complets suppose que les locuteurs voisins sont différents du locuteur courant. Les segments contigus contenant le même locuteur sont donc fusionnés pour obtenir une segmentation basée sur les changements de locuteur.
• Les informations concernant les 4 étiquettes doivent être accessibles durant les différentes phases. Nous avons donc étiqueté automatiquement la référence en extrayant du flux audio les noms et prénoms des locuteurs. Chaque nom complet extrait est comparé au nom de locuteur associé au segment ainsi qu'aux noms de locuteurs associés aux segments voisins. Cette tâche étant automatisée, nous supposons qu'il n'y a pas d'erreur d'identification du locuteur.
• Pour généraliser les exemples d'apprentissage pour la construction de l'arbre, chaque nom de locuteur est remplacé par une étiquette générique.
• Le SCT apprend les expressions régulières en tenant compte des mots appartenant aux contextes gauche et droit de l'occurrence du nom détecté. Au plus 20 mots pour le contexte gauche et 20 pour le droit sont conservés, ce nombre ayant été fixé sur le corpus de développement *Dev* pour maximiser le nombre de bonnes détections locales sur les 4 étiquettes.

[1] fusion des corpus de développement officiels des phase I et II d'ESTER

4.2 Etiquetage des segments

Table 2: Scores des décisions locales obtenus grâce au SCT sur les différents corpus.
- Etiquetés : % de noms complets détectés pour lesquels une règle de décision locale prorpose un étiquetage other, current, previous ou next.
- Previous (resp. pour chacune des étiquettes) : % de noms complets détectés correctement étiquetés par previous.

	Train	***Dev***	***Eva***
# Noms complets détectés	3297	920	507
Etiquetés (%)	94.51	94.78	97.23
Correctement Etiquetés (%)	88.25	76.49	68.76
Previous (%)	88.98	71.67	82.98
Current (%)	94.76	90.14	85.71
Next (%)	89.32	80.67	75.29
Other (%)	84.87	68.94	50.32

L'arbre de classification sémantique qui donne les résultats du tableau 2 a été construit à partir du corpus d'apprentissage. Le tableau montre les résultats des décisions locales prises sur chaque segment contenant un nom complet sur les corpus *Train*, *Dev* et *Eva*. La première colonne montre les résultats sur les données d'apprentissage utilisées comme des données de test.

94% des noms complets détectés sur *Dev* et 97% sur *Eva* sont associés à une de nos 4 étiquettes. Le taux d'étiquetage correct est d'environ 76.4% sur *Dev* et de seulement 68.7% sur *Eva* : ces valeurs peuvent être considérées comme étant la précision de la décision locale sur chaque corpus.
Le taux moins élevé de*Eva* (environ 8% de moins que *Dev*) peut être expliqué par la présence de deux nouvelles radios non présentes dans les autres corpus et ces données ont en plus été enregistrées 15 mois après les autres. Environ 6% des noms détectés ne sont pas étiquetés comme pour le corpus d'apprentissage.

Les résultats pour l'étiquette *other* sont les plus mauvais. Cela peut s'expliquer par le fait que cette étiquette est confrontée à une plus grande diversité de contextes lexicaux que les autres. Néanmoins, puisque l'étiquette *other* n'intervient pas directement

dans le processus global de prise de décision, l'impact de ces mauvais résultats est minime.

En choisissant toujours l'étiquette ayant la plus forte probabilité *a priori* (tab. 1), le meilleur score serait d'environ 49.3% d'étiquetage correct sur le corpus *Eva*. Avec la méthode proposée, nous atteignons le score d' environ 68% d'étiquetage correct pour *Eva*. Ces résultats montrent que l'utilisation d'un arbre de classification sémantique s'insère efficacement dans un processus de dénomination du locuteur.

4.3 Dénomination du locuteur

Table 3: Dénomination du locuteur : résultats détaillés pour les différents corpus (les taux sont calculés en termes de durée). *- Loc. : correpond aux 2 catégories de locuteurs de la référence, ceux qui sont les locuteurs clients (C) de l'application (locuteurs publiques avec un nom complet) et les autres (NC).*
- Dénom. : correspond aux dénominations correctes et incorrectes. "Non nommé" correspond au cas où le processus ne propose pas de nom.

Loc.	Dénom.	***Train***	***Dev***	***Eva***
C	Cor.(%)	63.68	64.82	66.35
C	Inc.(%)	3.19	5.48	14.36
C	Non nommé(%)	15.68	18.19	11.91
NC	Cor.(non nommé)(%)	15.50	7.54	3.59
NC	Inc.(%)	1.95	3.98	3.79
	Total(%)	100	100	100

Les décisions locales sur les segments sont ensuite fusionnées pour associer un nom complet à tous les segments de parole prononcés par le même locuteur (voir section 3.2). Les résultats détaillés de cette seconde étape sont reportées dans le tableau 3.

Méthode d'évaluation L'entrée du système est basée sur les transcriptions manuelles de référence : il n'y a ni d'erreurs d'indexation , ni de segmentation parole/non parole ni de transcriptions. Les frontières des segments de référence et d'hypothèse sont les mêmes, seuls les noms de leurs locuteurs diffèrent. Ici, seuls les locuteurs qui sont des personnes publiques (avec un nom complet dans la référence) sont les locuteurs clients. L'identité des autres ne peut pas être trouvée. Il y a donc erreur quand le processus donne à un locuteur non-client un nom complet et quand il ne donne pas de nom à un locuteur client (Table 3 lignes 2 & 5). De plus, le processus ne propose pas de nom à un locuteur client dans les cas où :
• aucune décision locale n'atteint un segment de ce locuteur client. Soit aucune décision n'est prise pour toutes les occurrences détectées de ce locuteur, soit les décisions locales attribuées sont fausses ;
• ce nom n'est pas détecté dans la transcription.
Pour les locuteurs clients, quand les noms complets hypothèses et ceux de référence sont les mêmes, la dénomination est correcte (Table 3 ligne 1). Pour les locuteurs non-clients, quand le processus ne propose pas de noms, il semble raisonnable de considérer cela comme correct (Table 3 ligne 4).
Tous les résultats proposés sont calculés en terme de durée comme c'est le cas lors des évaluations NIST [7].

Commentaires Le processus de dénomination de locuteur atteint 72% de décision correcte en terme de durée (64.82% + 7.54%) sur le corpus *Dev* et environ 70% (66.35% + 3.59%) sur *Eva* (voir tableau 3).

5 Conclusion

Dans le contexte de la transcription enrichie, nous proposons une méthode totalement automatisée qui permet d'identifier les locuteurs par leur véritable identité extraite directement de la transcription.
Le procédé proposé est basé sur l'utilisation d'un arbre de classification sémantique qui permet d'étiqueter les occurrences de noms détectées : cette première étape consiste à prendre des décisions locales qui associent une telle occurrence à un segment de parole. Ensuite, les résultats obtenus sont fusionnés pour associer un nom complet à tous les segments d'un même locuteur qui était anonymement annoté par l'indexation en locuteur.
Les expériences sont menées sur des émissions radiophoniques Françaises, fournies par la campagne d'évaluation ESTER. Environ 70% de la durée totale des émissions est correctement indexée en locuteur pour chacun des corpus de développement et d'évaluation. Sur le corpus d'évaluation, 18,15% de la durée totale des émissions est incorrectement indexée et aucune décision n'est prise pour 11.91%.
Le but principal est atteint : les résultats valident la méthode proposée de dénomination du locuteur à partir d'une indexation et d'une transcription manuelle. Les perspectives de travail s'orientent vers l'utilisation d'une indexation en locuteur et d'une transcription automatiques dans lesquelles des erreurs interviendront.

Remerciements

Merci à Frédéric Béchet du LIA pour la mise à disposition sous licence GPL de LIA_SCT. LIA_SCT est un classifieur basé sur un arbre de décision disponible sur le site web du LIA.

Bibliographie

[1] C. Barras, X. Zhu, S. Meignier, and J.-L. Gauvain, "Improving speaker diarization," in *DARPA RT04 Fall*, Palisades, NY, USA, 2004.

[2] S. Meignier, D. Moraru, C. Fredouille, J.-F. Bonastre, and L. Besacier, "Step-by-step and integrated approaches in broadcast news speaker diarization," *Computer Speech and Language*, 2005.

[3] F. Bimbot, J.-F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-Garcia, D. Petrovska, and D. A. Reynolds, "A tutorial on text-independent speaker verification," *EURASIP Journal on Applied Signal Processing, Special issue on biometric signal processing*, vol. 4, pp. 430–451, 2004.

[4] L. Canseco-Rodriguez, L. Lamel, and J.-L. Gauvain, "A comparative study using manual and automatic transcriptions for diarization," Jeju, Oct 2005.

[5] S. Galliano, E. Geoffrois, D. Mostefa, K. Choukri, J.-F. Bonastre, and G. Gravier, "The ESTER phase II evaluation campaign for the rich transcription of french broadcast news," in *Proceedings of European Conference on Speech Communication and Technology – Interspeech 2005*, Lisboa, Sep 2005, pp. 1149–1152.

[6] R. Kuhn and R. De Mori, "The application of semantic classifciation trees to natural language understanding," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 17, no. 5, pp. 449–460, 1995.

[7] NIST, "Fall 2004 rich transcription (RT-04F) evaluation plan," `http://www.nist.gov/speech/tests/rt/rt2004/fall/docs/rt04f-eval-plan-v%14.pdf`, August 2004.