Reconnaissance Automatique de Locuteurs à l'aide de Fonctions de Croyance

Simon Petitrenaud, Vincent Jousse, Sylvain Meignier, Yannick Estève

Reconnaissance Automatique de Locuteurs à l'aide de Fonctions de Croyance

Simon Petitrenaud Vincent Jousse Sylvain Meignier Yannick Estève

Laboratoire d'Informatique de l'Université du Maine

Avenue Laennec, 72085 Le Mans Cedex, France,
{simon.petit-renaud, vincent.jousse, sylvain.meignier, yannick.esteve} @lium.univ-lemans.fr

Résumé

Le thème de cet article est l'extraction automatique de l'identité du locuteur (prénom et patronyme) présente dans des enregistrements sonores. À partir des résultats d'un système de transcription de la parole, nous proposons d'améliorer une méthode récente visant à extraire l'identité des locuteurs de la transcription et à l'assigner aux différents tours de parole. Les identités des locuteurs détectés par la transcription sont prises comme candidats potentiels pour ces assignations. L'ensemble de ces informations souvent conflictuelles est décrit puis combiné à l'aide du formalisme des Fonctions de Croyance, qui apporte une cohérence à la représentation des connaissances du problème. Le système est évalué sur des enregistrements radiophoniques provenant d'une campagne d'évaluation francophone de systèmes de reconnaissance automatique de la parole.

Mots Clef

Identification nommée du locuteur, reconnaissance du locuteur, classification, fonctions de croyance.

Abstract

In this paper, we consider the extraction of speaker identity (first name and last name) from audio records of broadcast news. Using an automatic speech recognition system, we present improvements for a method which allows to extract speaker identities from automatic transcripts and to assign them to speaker turns. The detected complete names are chosen as potential candidates for these assignments. All this information, which is often contradictory, is described and combined in the Belief Functions formalism, which makes the knowledge representation of the problem coherent. Experiments are carried out on French broadcast news records from a French evaluation campaign of automatic speech recocognition.

Keywords

Speaker named identification, speaker recognition, clustering, belief functions.

1 Introduction

Pour faciliter la recherche et l'accès à l'information, les grandes collections de données audio ont besoin d'être indexées. Le système présenté dans ces travaux s'intéresse au cas de l'annotation des documents avec l'identité des locuteurs. Cette identité est composée du prénom et du patronyme du locuteur, que nous appellerons "nom complet" dans la suite de l'article. La première étape de l'annotation automatique de documents sonores consiste, à partir du signal acoustique, à segmenter le signal sonore en tours de parole, qui débutent lorsqu'un locuteur commence à parler et finissent lorsqu'un autre locuteur prend la parole ou qu'un intermède (jingle, chanson, publicité...) débute. Les différents tours de parole sont ensuite regroupés en classes contenant les segments produits par un même locuteur. Ces classes sont identifiées par des labels anonymes (locuteur 1, locuteur 2...). À ce stade, aucune connaissance *a priori* sur les locuteurs n'est utilisée. Toutefois, une détection du genre (homme ou femme) de chaque locuteur est réalisée. L'étape suivante consiste à transcrire automatiquement les tours de parole en mots et est complétée par une annotation en "entités nommées" ou catégories, certains mots étant en particulier identifiés comme des "PERSONNES". Une approche récente et prometteuse permettant d'attribuer un nom complet à un locuteur consiste à extraire les noms complets des locuteurs du système de reconnaissance de la parole [1, 11, 7, 6, 8]. Le principe général consiste à déterminer si une entité nommée de type "PERSONNE" se rapporte à un locuteur du document ou à une personne qui ne parle pas dans le document. L'approche se base sur un système en deux phases. Une première phase affecte les noms complets aux tours de parole proches. Puis dans une seconde phase, ces informations sont propagées au niveau des locuteurs.

C'est dans ce cadre que se place notre article. Le système que nous avons développé dans [6, 8, 7] utilise les noms complets prononcés pour les affecter aux locuteurs anonymes à partir des tours de parole identifiés. Le principe est d'attribuer une étiquette *"tour courant"*, *"tour précédent"*, *"tour suivant"* ou *"autre"* à chaque nom complet détecté. Mais cette méthode ne tenait pas compte du

conflit éventuel d'informations concernant les locuteurs au sein même d'un tour de parole. Dans cet article, nous proposons d'améliorer la cohérence du système, et de mieux combiner les différentes informations sur les locuteurs potentiels. Le formalisme des fonctions de croyance est apparu particulièrement adapté à la gestion de ces conflits et la combinaison de ces informations.

Dans un premier temps, nous présentons brièvement le système de transcription utilisé avant de décrire le système de référence pour l'identification nommée de locuteur. Ensuite, nous discutons des imperfections du modèle et de l'amélioration de notre modèle utilisant les fonctions de croyance. Enfin, nous proposons et discutons des métriques pour évaluer de tels systèmes, puis commentons les résultats obtenus sur la campagne d'évaluation ESTER I [4].

2 Reconnaissance du locuteur basée sur la transcription

2.1 Système de transcription

L'hypothèse de travail majeure proposée initialement dans [1] suppose qu'un nom complet détecté dans un tour de parole permet d'identifier le tour courant ou un des tours de parole directement contigus (tour de parole suivant ou précédent). Cependant, certains noms complets identifient des tours de parole plus éloignés ou des personnes n'intervenant pas dans le document. La figure 1 illustre les quatre types d'affectations possibles pour un nom complet détecté dans un tour de parole.

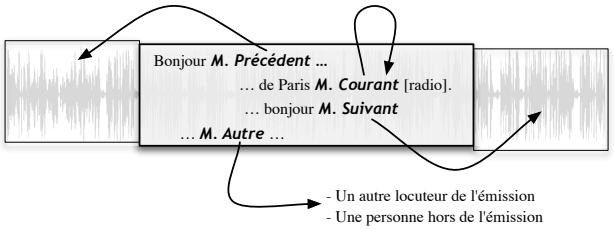

Figure 1: Principe de base des systèmes d'identification nommée basés sur la transcription.

La méthode d'identification nommée utilisée s'appuie sur des documents préalablement transcrits et enrichis. Cette transcription nécessite de découper le document en segments qui seront ensuite classifiés en locuteurs. Ces segments, groupés en tours de parole, sont transcrits et les entités nommées sont annotées. La figure 2 illustre ces trois étapes [7]. Le système de transcription utilisé est celui du Laboratoire d'Informatique de l'Université du Maine décrit dans [3]. Lors de la campagne d'évaluation ESTER 1 phase II en 2005 [4], pour la tâche de transcription, le système a été classé deuxième. Ce système permet d'obtenir 20.5% de taux d'erreur sur les mots sur ce corpus d'évaluation.

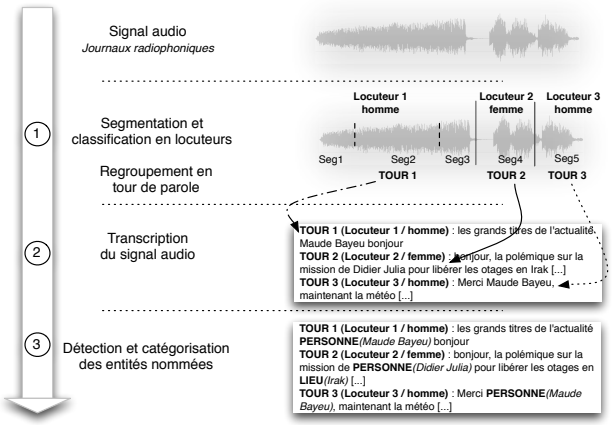

Figure 2: Description du système de transcription enrichie

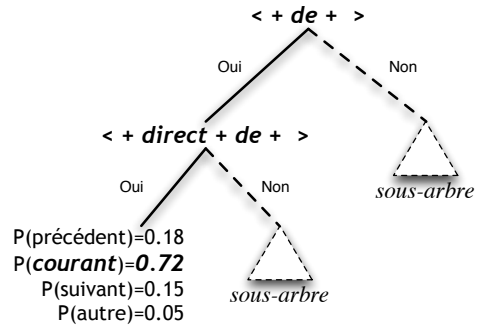

Figure 3: Arbre de classification sémantique

2.2 Arbre de classification sémantique

La méthode d'identification de locuteur utilise un arbre de décision binaire reposant sur le principe des arbres de classification sémantique [5], qui apprend automatiquement des règles lexicales à partir des noms complets détectés dans le corpus d'apprentissage. Cet arbre permet d'associer à chaque nom complet détecté la probabilité qu'il corresponde à l'une des 4 hypothèses envisagées: *"tour courant"*, *"tour précédent"*, *"tour suivant"* ou *"autre"*. Ces probabilités sont établies lors de l'apprentissage de l'arbre et reflètent les cas observés dans le corpus d'apprentissage. La figure 3 représente un arbre de classification avec deux exemples d'expressions régulières ainsi que les probabilités associées aux étiquettes dans une des feuilles.

2.3 Méthode de combinaison de référence

Grâce à l'arbre de classification, à chaque nom complet détecté dans la transcription est associé une liste d'étiquettes (tour courant ,...) valuée par un score probabiliste. Mais le but final du système est d'affecter à chaque locuteur un nom complet détecté. Nous rappelons ici la méthode d'agrégation de l'information fournie par les arbres proposée dans [6].

Soit $\mathcal{E} = \{e_1, \ldots, e_I\}$ correspondant à l'ensemble des noms complets candidats pour nommer un locuteur. Ces candidats sont issus d'une liste des locuteurs possibles connue du système. L'ensemble $\mathcal{O} = \{o_1, \ldots, o_J\}$ correspond aux occurrences successives des noms complets détectés dans les transcriptions, $\mathcal{T} = \{t_1, \ldots, t_K\}$ désigne l'ensemble des tours de parole dans l'ordre chronologique, et $\mathcal{C} = \{c_1, \ldots, c_L\}$ l'ensemble des locuteurs à nommer. Le but est donc d'attribuer un nom complet issu de $\mathcal{E}$ aux locuteurs de $\mathcal{C}$. Chaque locuteur c_l peut intervenir une ou plusieurs fois dans une émission, ce qui correspond donc à un ou plusieurs tours de parole: $c_l = \{t \in \mathcal{T} \ | c_l \text{ est le locuteur de } t\}$. Chaque tour de parole appartient à un et un seul locuteur, c'est-à-dire que les tours de parole sont successifs et ne peuvent être simultanés. Pour chaque occurrence d'un nom complet o_j (pour $j = 1, \ldots, J$) détecté dans un tour de parole t_k, on désigne par $P(o_j, t_k)$ la probabilité que o_j soit le locuteur du tour de parole t_k. Ainsi, $P(o_j, t_{k-1})$ et $P(o_j, t_{k+1})$ représentent la probabilité que o_j soit le locuteur du tour respectivement précédent et suivant celui où il a été détecté. Par hypothèse, la probabilité que o_j soit un autre locuteur est donc: $1 - \sum_{r \in \{-1,0,1\}} P(o_j, t_{k+r})$.

A ce stade, un filtre est réalisé: il s'agit de la comparaison des genres. Si le genre du nom complet e_i et du locuteur c_l sont differents, les probabilités précédentes ne seront pas prises en compte. On note $g(e_i)$ et $g(c_l)$ les genres respectifs (féminin, masculin ou inconnu) de e_i et c_l. Le genre des locuteurs est détecté par le système de segmentation et de classification acoustique avec une grande fiabilité et le genre des noms complets est déterminé par le prénom (généralement sans ambiguité s'il est reconnu) à partir d'une base linguistique de prénoms.

Dans [6], pour l'assignation d'un nom complet e_i à un locuteur c_l donné, nous avons calculé un "score" pour chaque nom complet e_i, dénoté $s_l(e_i)$, qui n'est plus une probabilité, mais qui correspond simplement au cumul des probabilités concernant les tours de parole du locuteur c_l, en tenant compte des contraintes de genre:

$$s_l(e_i) = \sum_{\{(o_j,t)|o_j=e_i,\ t\in c_l, g(e_i)=g(c_l)\}} P(o_j, t) \quad (1)$$

Ensuite, un processus de décision détermine le nom complet attribué à chaque locuteur.

2.4 Processus de décision associé

Le but est maintenant d'attribuer à chaque locuteur c_l un nom complet e_i. Soit $f : \mathcal{C} \to \mathcal{E}$ la fonction d'assignation des noms complets aux locuteurs. Les locuteurs étant censés être tous différents, une caractéristique importante que l'on voudrait imposer pour f est l'injectivité: $f(c_l) = f(c'_l) \Rightarrow c_l = c'_l$. Chaque nom complet ne peut être attribué à plus d'un locuteur. Bien sûr, certains noms complets peuvent rester non attribués et inversement, certains locuteurs peuvent rester sans étiquette. Le principe de notre solution, proposée dans [7], est de réorganiser le partage des noms complets entre les locuteurs anonymes. Il s'agit en fait d'une simple mise en correspondance des noms complets et des locuteurs. Soit $\mathcal{D} = \{c_l \in \mathcal{C} | \ \forall e_i \in \mathcal{E}, s_l(e_i) = 0\}$, l'ensemble des locuteurs n'ayant pas de candidats potentiels. Plusieurs stratégies peuvent être utilisées pour trier les locuteurs c_l en concurrence pour un nom complet e_i donné. Prendre le score maximum $s_l(e_i)$ semble être la solution la plus naturelle. Soit la règle $\mathbf{R_1}$:

$$\begin{gathered} \forall c_l \in \mathcal{C} \setminus \mathcal{D}, e_i^* = \arg\max_{e_i \in \mathcal{E}} \mathrm{s_l}(e_i) \Rightarrow f(c_l) = e_i^* \\ \forall c_l \in \mathcal{D}, f(c_l) = \text{Anonyme} \end{gathered} \quad (2)$$

Mais ce score seul ne tient pas compte de la compétition relative de différents candidats pour un même locuteur. Soit le coefficient β_{il} définissant la proportion de score allouée à e_i pour l'affectation à c_l, relativement aux autres noms complets potentiels:

$$\beta_{il} = \begin{cases} \dfrac{s_l(e_i)}{\sum_{q=1}^{I} s_l(e_q)} & \text{si } c_l \notin \mathcal{D} \\ 0 & \text{si } c_l \in \mathcal{D}. \end{cases} \quad (3)$$

Un exemple concret est donné dans le tableau 1. Le nom complet "Jacques Derrida" a été assigné à trois locuteurs différents à partir de la règle de décision de l'équation 2. Dans cet exemple, c_{13} a le meilleur score, et "Jacques Derrida" devrait donc être affecté à c_{13} ; mais le score ne représente que 35% des scores totaux parmi tous les candidats possibles pour c_{13}, alors que le score pour c_{15} représente 80% des scores totaux. Nous avons alors proposé d'utiliser pour la décision le produit du score $s_l(e_i)$, par le coefficient β_{il} (règle $\mathbf{R_2}$):

$$\mathrm{SC}_l(e_i) = s_l(e_i)\beta_{il} \quad (4)$$

Finalement, dans le même exemple, "Jacques Derrida" est assigné à c_{15} et d'autres noms complets seront attribués aux locuteurs c_{13} et c_{14}.

Table 1: Exemple d'une assignation initiale multiple

Locuteur	nom complet e_i^*	$s_l(e_i^*)$	β_{il}	$\mathrm{SC}_l(e_i^*)$
c_{13}	Jacques Derrida	$\mathbf{8,58}$	35%	$3,00$
c_{14}	Jacques Derrida	$1,67$	56%	$0,94$
c_{15}	Jacques Derrida	$4,94$	$\mathbf{80}\%$	$\mathbf{3,95}$

L'algorithme est donc le suivant: tous les noms complets possibles sont pris en compte *a priori* et triés en fonction de leur score $\mathrm{SC}_l(e_i)$. Premièrement, le nom complet avec le score maximum (noté e_i^*) est choisi, et si plusieurs locuteurs sont associés au même e_i^*, alors ce nom complet sera assigné au locuteur dont le score $SC_l(e_i^*)$ est maximum. Ensuite, tous les noms complets choisis sont supprimés de la liste relative aux locuteurs qui n'ont pas encore été nommés. Lors de l'itération suivante, les noms

complets restants sont examinés de la même manière pour les locuteurs restants et ainsi de suite, jusqu'à ce que tous les locuteurs soient nommés ou que la liste des noms complets à attribuer soit vide. Le tableau 2 montre les résultats obtenus pour l'exemple précédent.

Table 2: Exemple du processus de décision avec deux itérations (décision en gras, scores entre parenthèses).

Locuteur	e_i^* (1ère itération)	2ème itération
c_{13}	J. Derrida $(3,00)$	**N. Demorand** $(0,25)$
c_{14}	J. Derrida $(0,94)$	**A. Adler** (0,56)
c_{15}	**J. Derrida** $(3,95)$	-
c_{16}	**O. Duhamel** $(1,15)$	-

2.5 Critiques de la méthode de combinaison

Plusieurs critiques de natures diverses peuvent être émises quant à la pertinence de la méthode de combinaison exposée précédemment en section 2.3, même si elle a donné de bons résultats [6, 7]. Premièrement, la notion de score est difficile à interpréter, ces quantités obtenues ne représentent pas de degré de confiance, ni de probabilité que tel nom complet soit tel locuteur. Elles conduisent à un manque de lisibilité de la décision. L'équation 4 sur laquelle s'appuie la décision représente un compromis qu'il est difficile de justifier.

Mais la critique principale concerne la méthode-même de combinaison: le conflit d'informations au sein d'un tour de parole donné n'est pas pris en compte. L'information disponible n'est pas combinée dans sa globalité de façon satisfaisante. La méthode de combinaison proposée est susceptible de propager des imperfections pour l'étape suivante de la décision. En effet, l'affectation d'un nom complet à un locuteur ne tient pas compte des liens entre les différentes informations fournies par l'arbre de classification, en particulier quand un même locuteur prononce plusieurs noms complets et peut donc aboutir à des résultats erronés. Le tableau 3 présente un exemple de tour de parole t_k où 8 noms complets sont détectés. Les probabilités indiquées correspondent au locuteur du tour suivant t_{k+1}, qui est de genre masculin. L'un des noms complets (féminin) est donc éliminé. Certaines occurrences sont redondantes, car elles correspondent à un même nom complet (Jean-Claude Pajak et Jacques Chirac) et une seule occurence a une probabilité élevée. Il reste donc 4 noms complets en compétition, ce qui représente une incompatibilité importante. Or, la contribution de ce tour produit des scores élevés pour Jean-Claude Pajak $(1,25)$ et Jacques Chirac $(0,87)$, scores semblables à celui qu'on obtiendrait si on disposait d'une information sans ambiguité, comme par exemple un tour ne contenant qu'une seule occurrence de probabilité élevée sans concurrence. Cet exemple met en évidence le fait que cette méthode ne prend pas en compte la contradiction des informations délivrées par certains tours de parole. Il conviendrait donc de réajuster le calcul des scores en tenant mieux compte de l'incertitude de ces informations. Un formalisme probabiliste basé par exemple sur des probabilités conditionnelles pourrait être envisagé pour ce type de situations, mais le manque d'informations *a priori* ont rendu ce type de modélisation difficile. Bien que les sorties de l'arbre soient de nature probabiliste, la théorie des croyances nous a paru mieux adaptée et moins contraignante, grâce en particuler à la souplesse de son utilisation.

Table 3: Contribution du score dans un tour de parole (le genre du locuteur t_{k+1} est masculin).

Occurrence o_j	sexe	$P(o_j, t_{k+1})$	score
Oscar Temaru	M	$0,29$	$0,29$
Hamid Karzaï	M	$0,29$	$0,29$
Jacques Chirac	M	$0,29$	$0,87$
Jacques Chirac	M	$0,29$	
Jacques Chirac	M	$0,29$	
Jean-Claude Pajak	M	$0,29$	
Jean-Claude Pajak	M	**0,96**	$1,25$
Véronique Rebeyrotte	*F*	*0,29*	–

3 Fonctions de croyances pour la reconnaissance de locuteur

L'apport de cet article réside essentiellement dans le processus de combinaison des différentes informations, en particulier des sorties de l'arbre de classification.

3.1 Fonctions de croyances

Dans cette section, nous rappelons très brièvement quelques éléments de la théorie des fonctions de croyances [9, 10]. Dans cet article, nous adoptons l'interprétation proposée par Smets: le modèle des croyances transférables [10]. Le but de ce modèle est de déterminer la croyance concernant différentes propositions, à partir d'un ensemble d'informations disponibles. Soit $\mathcal{E}$ un ensemble fini, appelé cadre de discernement de l'expérience. La représentation de l'incertitude se fait grâce au concept de la fonction de croyance, définie comme une fonction m de $2^{\mathcal{E}}$ dans $[0,1]$ telle que:

$$\sum_{A \subseteq \mathcal{E}} m(A) = 1. \tag{5}$$

La quantité $m(A)$ représente la part de croyance allouée exactement à la proposition A. Les sous-ensembles A de $\mathcal{E}$ tels que $m(A) > 0$ sont les *éléments focaux* de m. Une structure de croyance est à support simple si elle est focalisée au maximum sur $\mathcal{E}$ et un seul sous-ensemble A. Elle est alors définie par:

$$\begin{cases} m(A) = w, \quad w \in [0,1] \\ m(\mathcal{E}) = 1 - w \end{cases} \tag{6}$$

L'ignorance totale est représentée par la fonction de croyance vide telle que $m(\Omega) = 1$. Si tous les éléments focaux

de m sont des singletons, m devient alors une mesure de probabilité (dite masse bayésienne).
La théorie des croyances possède plusieurs outils d'agrégation permettant de combiner des fonctions de croyance définies sur un même cadre de discernement [10]. En particulier, la combinaison de deux structures m_1 et m_2 définies de façon "indépendante" sur $\mathcal{E}$ utilisant l'opérateur binaire conjonctif non normalisé $\cap$ définit une fonction résultante m' par la formule suivante [10]:

$$\forall A \subseteq \mathcal{E},\ m'(A) = \sum_{B \cap C = A} m_1(B) m_2(C). \qquad (7)$$

On peut alors définir la combinaison de n structures $m_1, \ldots, m_n$ sur $\mathcal{E}$ par : $m = m_1 \cap \ldots \cap m_n$, telle que

$$m(A) = \sum_{A_1 \cap \ldots \cap A_n = A} \prod_{i=1}^{n} m_i(A_i) \quad \forall A \subseteq \Omega. \qquad (8)$$

Si on obtient une fonction de croyance m non normalisée (i. e. telle que $m(\emptyset) \neq 0$), on peut la convertir par la procédure de normalisation de Dempster [9] en une structure normalisée m^* en répartissant proportionnellement la masse de l'ensemble vide sur les autres éléments focaux :

$$m^*(A) = \begin{cases} \dfrac{m(A)}{1 - m(\emptyset)} & \text{si } A \neq \emptyset \\ 0 & \text{si } A = \emptyset. \end{cases} \qquad (9)$$

Une fonction de croyance m décrit l'état d'une croyance sur un phénomène. Une fois qu'une structure m est définie, il est possible de la transformer en distribution de probabilité, en particulier pour des aspects décisionnels. Une de ces distributions, appelée probabilité *pignistique*, consiste à répartir équitablement la masse d'un sous-ensemble de Ω entre ses éléments. Cette distribution, notée P_m, est définie pour tout $\omega \in \Omega$ par [10]:

$$P_m(\{\omega\}) = \sum_{A \subset \Omega} \frac{m^*(A)}{|A|} \delta_A(\omega), \qquad (10)$$

où $|A|$ est le cardinal de A, $\delta_A(\omega) = 1$ si $\omega \in A$ et $\delta_A(\omega) = 0$ si $\omega \notin A$.

3.2 Définition des masses de croyance

Dans cet article, nous proposons d'améliorer le système proposé dans [6, 7] en considérant la *cohérence* des informations au sein de tours de parole *contigus*. En effet dans un *même* tour, comme on l'a vu, plusieurs occurrences correspondant à des noms complets différents peuvent être détectées. Plusieurs noms complets sont donc en compétition ou en conflit en tant que locuteur courant potentiel (ou précédent, ou suivant).
On s'intéresse à un tour t_k appartenant au locuteur c_l et ayant n_k occurrences. Soient n_{k+r}, le nombre d'occurrences tours précédent ($r = -1$) et suivant ($r = 1$) de ce tour. Soient $\{o^k_{j,r}\}$, avec $r = -1, 0, 1$ et $j = 1, \ldots, n_{k+r}$, les occurrences de noms complets détectés dans ces trois tours. Chaque occurrence $o^k_{j,r}$, correspondant à une étiquette e_i, représente une certaine information sur le locuteur du tour t_k qui peut être décrite par une masse de croyance $m^{jr}_{t_k}$ sur $\mathcal{E}$ à support simple, focalisée sur e_i et $\mathcal{E}$:

$$\begin{cases} m^{jr}_{t_k}(\{e_i\}) = \alpha_{il} P(o^k_{j,r}, t_{k-r}) \text{ si } o^k_{j,r} = e_i \\ m^{jr}_{t_k}(\mathcal{E}) = 1 - \alpha_{ij} P(o^k_{j,r}, t_{k-r}), \end{cases} \qquad (11)$$

où $\alpha_{il} \in [0, 1]$ est une mesure de compatibilité de genre entre e_i et c_l. Si les genres sont connus avec certitude, $\alpha_{il} = 0$ si $g(e_i) \neq g(c_l)$ et $\alpha_{il} = 1$ si $g(e_i) = g(c_l)$. Si les prénoms sont ambigus (comme Dominique) ou non spécifiés, ou si le genre du locuteur est incertain, $\alpha_{il} \in]0, 1[$ est estimé à partir d'une base de prénoms et du corpus d'apprentissage. Le tableau 4 présente la masse de croyance concernant le locuteur du tour t_{k+1} dans l'exemple vu en section 2.5. La masse de croyance du nom complet Jean-Claude Pajak se maintient tandis que celle des autres candidats est considérablement réduite.

Table 4: Calcul de score dans un tour de parole: méthode proposée.

Eléments focaux	$m^*_{t_{k+1}}(\{e_i\})$
Oscar Temaru (e_1)	$0,011$
Hamid Karzaï (e_2)	$0,011$
Jacques Chirac (e_3)	$0,047$
Jean-Claude Pajak	**0,904**
$\mathcal{E}$	$0,027$

3.3 Combinaison par tour de parole et par locuteur

La première étape de la combinaison consiste à agréger les informations au sein d'un tour de parole donné. La combinaison des $n_{k-1} + n_k + n_{k+1}$ masses ciblées sur le tour t_k et obtenues par l'équation 11 peut se faire par la règle conjonctive de Dempster *non normalisée* (cf. équations 7-8), afin d'assurer l'associativité et la commutativité de la combinaison: on obtient une masse de croyance m_{t_k} sur l'identité du locuteur du tour t_k, définie par

$$m_{t_k} = \bigcap_{r=-1}^{1} \bigcap_{j=1}^{n_{k+r}} m^{jr}_{t_k} \qquad (12)$$

et la masse de croyance normalisée $m^*_{t_k}$ correspondante (cf. équation 9).
La deuxième étape de la combinaison consiste à agréger les résultats obtenus par tour de parole pour l'ensemble de l'émission. Plusieurs possibilités sont envisageables *a priori* pour combiner ces informations. Il est possible de procéder de façon analogue à celle de l'équation 1. Pour l'assignation d'un nom complet e_i à un locuteur c_l donné, nous pourrions calculer un "score" pour chaque e_i, dénoté

$sm_l(e_i)$, comme la somme des masses de croyance concernant les tours de parole du locuteur c_l :

$$sm_l(e_i) = \sum_{t_k \in c_l} m^*_{t_k}(\{e_i\}). \quad (13)$$

Les scores obtenus ne sont donc pas nécessairement normalisés (i.e. $\leq$ 1), ce qui rendra là encore difficile l'interprétation des résultats lors du processus de décision. Nous préférons rester dans le cadre de la théorie des croyances: une manière assez naturelle consiste en effet à continuer de combiner toutes les masses de croyance relatives au même locuteur c_l par la même règle conjonctive de Dempster et donc à fusionner toutes les masses de croyance relatives aux différents tours de parole t_k de ce locuteur. On obtient alors une masse de croyance globale M_l sur l'identité du locuteur c_l de l'émission, définie par

$$M_l = \bigcap_{t_k \in c_l} m_{t_k} \quad (14)$$

et la masse de croyance normalisée M^*_l correspondante.

3.4 Règle de décision

Pour la décision d'affectation, nous nous appuyons sur la méthode exposée en section 2.4, mais celle-ci se trouve simplifiée et unifiée grâce à l'utilisation des masses de croyance. Nous transformons les masses de croyance M_l en une probabilité pignistique P_{M_l} (cf. équation 10) et nous utilisons la règle $\mathbf{R}$ suivante :

$$\begin{gathered}\forall c_l \in \mathcal{C} \setminus \mathcal{D}, e^*_i = \arg\max_{e_i \in \mathcal{E}} \mathrm{P}_{M_l}(e_i) \Rightarrow f(c_l) = e^*_i \\ \forall c_l \in \mathcal{D}, f(c_l) = \text{Anonyme}\end{gathered} \quad (15)$$

Ensuite, comme certains noms complets sont attribués lors de cette première étape à plusieurs locuteurs différents, il faut là encore réorganiser le partage des noms complets entre les locuteurs anonymes. On applique le même processus de décision que celui proposé en 2.4, en remplaçant les scores SC_l par les probabilités pignistiques P_{M_l}. En reprenant l'exemple proposé en 2.4, le nom complet "Jacques Derrida" est là encore assigné aux départ aux trois locuteurs c_{13}, c_{14} et c_{15} (cf. tableau 5). Mais finalement,"Jacques Derrida" est également assigné à c_{15}, car c'est pour ce locuteur qu'il obtient la plus grande probabilité pignistique.

Table 5: Décision avec deux itérations (décision en gras, probabilités pignistiques $P_{M_l}(e^*_i)$ entre parenthèses).

Locuteur	e^*_i (1ère itération)	2ème itération
c_{13}	J. Derrida $(0,89)$	**N. Demorand** $(0,11)$
c_{14}	J. Derrida $(0,71)$	**A. Adler** $(0,25)$
c_{15}	**J. Derrida** $(0,99)$	-
c_{16}	**O. Duhamel** $(0,88)$	-

4 Évaluation du système proposé

4.1 Description des corpus

L'évaluation du système proposé est réalisée à partir d'émissions radiophoniques en français de la campagne ESTER 1 phase II [4]. La majorité de ces émissions contient essentiellement de la parole lue ou préparée, et peu de parole spontanée: 15% du corpus correspond à des interventions de personnes parlant au téléphone. Les émissions proviennent de 5 radios françaises et de *Radio Télévision Marocaine* et durent de 10 à 60 min. Elles sont réparties en 3 corpus utilisés pour l'apprentissage de l'arbre de classification, le développement et l'évaluation du système. Le corpus d'apprentissage contient 76h de données (7416 tours de parole) dans lesquels 11292 noms complets sont détectés. Le corpus d'évaluation contient 10h (1082 tours de parole, 1541 noms complets détectés). Ce découpage correspond à celui de la campagne d'évaluation officielle ESTER 1 phase II.

4.2 Métriques utilisées

Le système d'identification nommée est évalué en comparant l'hypothèse générée par celui-ci à la référence distribuée avec le corpus. Cette comparaison met en évidence 5 cas d'erreur ou de succès possibles relatifs aux situations suivantes :

- l'identité proposée est correcte (C_1): le système propose une identité correspondant à celle indiquée dans la référence ;
- erreur de substitution (S): le système propose une identité différente de l'identité présente dans la référence ;
- erreur de suppression (D): le système ne propose pas d'identité alors que le locuteur est identifié dans la référence ;
- erreur d'insertion (I): le système propose une identité alors que le locuteur n'est pas identifié dans la référence ;
- pas d'identité (C_2): le système ne propose pas d'identité et la référence n'en contient pas non plus.

Parmi les différentes mesures définies dans [11, 2, 6, 7], celle qui nous semble le mieux synthétiser les résultats est le taux d'erreur Err global calculé à partir de ces 5 quantités.

$$Err = \frac{S + I + D}{S + I + D + C_2 + C_1}. \quad (16)$$

Il a l'avantage de mesurer la qualité des résultats du système d'identification nommée en une seule valeur , facilitant les comparaisons entre les systèmes. Les erreurs peuvent être calculées en terme de durée ou en terme de nombre de locuteurs.

Système	ErrDur	ErrLoc
Base	26,6%	37,4%
Référence (règle $\mathbf{R_1}$)	20,6%	20,2%
Référence (règle $\mathbf{R_2}$)	16,6%	19,5%
Proposé (règle $\mathbf{R}$)	$\mathbf{13,7}\%$	$\mathbf{14,9}\%$

Table 6: Comparaison système proposé et système de référence selon la règle de décision sur le corpus d'évaluation ESTER 1 phase II; ErrDur: taux d'erreur en durée; ErrLoc: taux d'erreur en nombre de locuteurs.

4.3 Évaluation du système

Dans les expériences, on suppose que le système connaît tous les noms complets susceptibles d'être des locuteurs. Cette liste est constituée de 1008 noms complets. La comparaison entre le système de base décrit dans [8], le système de référence le plus récent (cf. [7] et section 2.3), et le système proposé est effectuée sur des transcriptions et segmentations manuelles: il n'y a pas d'erreur de segmentation et de classification en locuteurs. De même, tous les noms complets de locuteurs sont correctement transcrits. En revanche, la détection des entités nommées comporte des erreurs. Le modèle de base n'utilisait pas la réaffectation de noms complets en cas d'étiquetage initial multiple de locuteurs, le modèle de référence correspond au système décrit en section 2.3 avec les deux règles utilisant les scores $s_l(e_i)$ et $SC_l(e_i)$ (cf. équations 2 et 4) et notre modèle est basé sur la maximisation des masses de croyance (équation 15).
Comme le montre le tableau 6, dans le système actuel, le taux d'erreur en durée ($ErrDur$) est de 3 points de moins que le système de référence avec la règle $\mathbf{R}_2$ et de 7 points avec la règle $\mathbf{R}_1$. Outre le fait qu'elles sont plus facilement interprétables, l'utilisation des masses de croyance élimine donc beaucoup d'erreurs. En ce qui concerne le nombre de locuteurs identifiés, le résultat est encore plus flagrant :le nouveau système étiquette correctement bien plus de locuteurs que le système de base, et améliore également davantage le système de référence. En conclusion, la prise en compte d'informations globales sur les locuteurs au sein des différents tours de parole, avant la décision, a permis d'améliorer notablement les résultats aussi bien en termes de durée que de nombres de locuteurs.

5 Conclusion

La méthode d'identification des locuteurs proposée dans cet article permet d'extraire les identités des locuteurs des transcriptions. L'identification est réalisée à l'aide d'un arbre de classification sémantique qui attribue les noms complets détectés dans la transcription aux locuteurs s'exprimant dans l'enregistrement. Dans cet article, nous proposons un nouveau système qui combine de manière cohérente les différentes informations sur les locuteurs potentiels à partir des fonctions de croyance. En particulier, le système tient compte du conflit éventuel d'informations concernant les locuteurs au sein même d'un tour de parole et la prise en compte d'incertitude sur leur genre. Le choix des identités des locuteurs est reporté en fin de processus où tous les noms complets candidats sont mis en concurrence. Les expériences ont été réalisées sur des émissions radiophoniques en français issues de la campagne d'évaluation ESTER 1 et le système obtient de très bonnes performances. Les travaux futurs se focaliseront sur le traitement des transcriptions automatiques. Différents types d'incertitude, dues aux erreurs de segmentation et de classification en locuteurs ou à la mauvaise transcription de noms complets de locuteurs devront alors être pris en compte. Nous nous intéresserons également au cas d'un système ouvert où la liste des locuteurs possibles n'est pas connue.

References

[1] L. Canseco-Rodriguez, L. Lamel, J. L. Gauvain. A comparative study using manual and automatic transcriptions for diarization. *Automatic Speech Recognition and Understanding (ASRU)*, 2005.

[2] M. Chengyuan, P. Nguyen, M. Mahajan. Finding speaker identities with a conditional maximum entropy model. *ICASSP'07*, 2007.

[3] P. Deléglise, Y. Estève, S. Meignier, T. Merlin. The LIUM speech transcription system: a CMU Sphinx III-based system for French broadcast news. *Eurospeech'05*, 2005.

[4] S. Galliano, E. Geffroy, D. Mostefa, K. Choukri, J.-F. Bonastre, G. Gravier. The ESTER phase II evaluation campaign for the rich transcription of French broadcast news. *Eurospeech'05*, 2005.

[5] R. Kuhn, R. De Mori. The application of semantic classification trees to natural language understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*,17(5):449-460, 1995.

[6] V. Jousse, S. Petitrenaud, S. Meignier, Y. Estève, C. Jacquin. Automatic named identification of speakers using diarization and asr systems. *ICASSP'09, 2009*.

[7] V. Jousse, S. Meignier, C. Jacquin, S. Petitrenaud, Y. Estève, B. Daille. Analyse conjointe du signal sonore et de sa transcription pour l'identification nommée de locuteur. *Traitement automatique des langues*, 50(1), 2009.

[8] Mauclair, J. and Meignier, S. and Estève, Y. Speaker diarization: about whom the speaker is talking? *IEEE Odyssey, 2006*.

[9] G. Shafer. *A Mathematical Theory of Evidence*. Princeton University Press, Princeton, 1976.

[10] P. Smets, R. Kennes. The transferable belief model. *Artificial Intelligence*, 66 : 191-234, 1994.

[11] S. E. Tranter. Who really spoke when? Finding speaker turns and identities in broadcast news audio. *ICASSP'06, 2006*, pp. 1013-1016.