Modèle de Markov évolutif pour les tâches de suivi de locuteurs

Sylvain Meignier, Jean-François Bonastre, Corinne Fredouille, Teva Merlin

Modèle de Markov évolutif pour les tâches de suivi de locuteurs

Sylvain Meignier, Jean-François Bonastre, Corinne Fredouille, Teva Merlin*

LIA/CERI Université d'Avignon, Agroparc,
BP 1228, 84911 Avignon Cedex 9, France.
{sylvain.meignier, jean-francois.bonastre, corinne.fredouille, teva.merlin}@lia.univ-avignon.fr

Abstract

Seeking within a speech sequence the speaker utterances is one of the main tasks of indexing.

In this paper, the proposed speaker tracking system is defined in the case where all speaker identities are known beforehand. The conversation is modeled as an evolutive Markov Model, in which speaker models computed are added one by one. A temporary indexing is proposed after each speaker adding and then challenged at the next step. This process is iterated until all the speakers are detected.

The system has been assessed using multi-speaker messages generated by concatenation of Switchboard mono-speaker segments. The obtained results show the potentiality of the proposed solution.

1. Introduction

La recherche au sein d'une conversation des paroles prononcées par différents locuteurs constitue une tâche essentielle pour l'indexation par le contenu de documents multimédia.

Deux approches usuelles, en indexation ou en suivi de locuteurs (*speaker tracking*), sont communément envisagées. La première méthode, décrite notamment dans [1] et [2], repose dans une première phase sur une détection des ruptures provoquées par les changements de locuteurs. Une seconde phase, dite de *clustering* détermine le nombre de locuteurs et groupe les segments par locuteur. Aucune information sur les locuteurs potentiels n'est utilisée pendant ces phases. Cette caractéristique rend la méthode bien adaptée aux tâches d'indexation en aveugle. La seconde méthode, proposée en particulier dans [3], est basée sur un système de reconnaissance du locuteur. La détection des segments et l'attribution de ceux-ci aux différents locuteurs sont réalisées simultanément. Dans cette approche, les locuteurs potentiels (et leurs modèles respectifs) doivent être connus par avance. Cette dernière contrainte destine particulièrement cette technique aux tâches de suivi de locuteurs.

Dans cet article, nous proposons un système apparenté à la deuxième méthode pour les tâches de suivi de locuteurs. Toutes les identités des intervenants sont connues *a priori*. La conversation est modélisée par un modèle de Markov évolutif (proche de celui proposé dans [3]). Au cours du processus d'indexation, le modèle évolue à chaque nouvelle détection d'un locuteur.

Le système proposé a été testé sur des messages générés à partir de la concaténation de segments de parole mono-locuteur issus de la base Switchboard (NIST 1998[1]). La base est constituée de segments de parole réelle, bruitée, provenant de conversations téléphoniques.

2. Modèle de conversation

2.1. Structure du modèle de conversation

La dynamique de la conversation est représentée par un modèle de Markov ergodique, où les états caractérisent les locuteurs du message et les transitions entre les différents états modélisent les changements de locuteurs.

Le modèle est défini par :

- Un ensemble d'états représentant les modèles de locuteurs, le modèle de Parole et le modèle de Non Parole. A chaque état est associé un ensemble de probabilités d'émission, calculées par un système de vérification du locuteur (en utilisant le modèle de locuteur correspondant à cet état).
- Un ensemble de poids de transition entre les états, représentant les pénalités de changement de modèle (locuteurs, Parole, Non Parole).

L'état Parole détecte les blocs contenant des données devant être attribuées à un modèle de locuteur alors que l'état Non Parole indique les blocs contenant des silences, des perturbations importantes du canal de transmission, ou des bruits environnant les locuteurs enregistrés.

Le modèle ergodique n'intègre pas de connaissance *a priori* sur le nombre et la durée des différentes interventions, ni sur la structure de la conversation.

Les poids de transition sont établis en fonction d'un ensemble de règles. Les poids entre les états des locuteurs vérifient trois conditions :

* projet RAVOL : support financier du Conseil général de la région Provence Alpes Côte d'Azur et de DigiFrance.

[1] http://www.nist.gov/speech/spkrec98.html

- Les probabilités ($P(s_i \rightarrow s_i)$) de rester dans le même état (s_i) sont égales pour chaque état.
- Les probabilités ($P(s_i \rightarrow s_j)$) entre deux différents états (s_i, s_j) sont égales.
- Quelque soient s_i et s_j deux états différents alors $P(s_i \rightarrow s_j) < P(s_i \rightarrow s_i)$.

Les poids associés aux modèles Parole et Non Parole sont fixés par l'opérateur en fonction du coût d'une erreur d'indexation (bloc attribué à un mauvais locuteur) par rapport au coût d'une non décision (bloc non attribué à un locuteur).

2.2. Construction du modèle par un processus itératif

La construction du modèle de conversation est réalisée par un processus itératif, où les locuteurs sont détectés et ajoutés un à un au modèle de Markov.

A l'initialisation du processus (figure 1), le modèle de Markov est composé de deux états, qui représentent le modèle de Parole et le modèle de Non Parole.

En fin d'initialisation, l'algorithme de Viterbi, appliqué au modèle de Markov, donne une première segmentation, qui sera remise en cause à l'étape suivante.

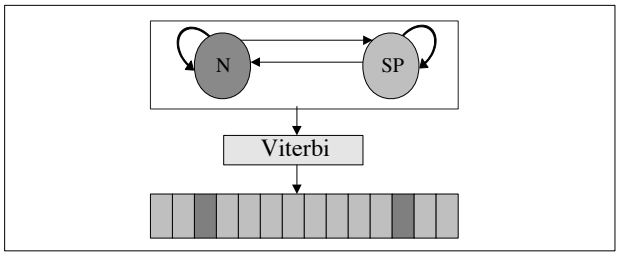

Figure 1: *Processus itératif : initialisation*

Dans la partie itérative du processus (figure 2), les différentes phases sont :

1. A partir des zones indexées Parole, le modèle le plus probable est détecté parmi les modèles de locuteurs qui n'ont pas encore été introduits dans le modèle de conversation. Cette phase est réalisée à l'aide de l'algorithme SWGM[2] [8].
2. Un nouvel état représentant le locuteur choisi en (1) est ajouté au modèle de Markov. Les probabilités d'émission de cet état sont calculées. Les poids des transitions sont adaptés pour prendre en considération le nouveau nombre d'états.
3. L'algorithme de Viterbi est appliqué pour obtenir l'alignement optimal par rapport à la topologie actuelle du modèle de Markov. On obtient une indexation temporaire, qui sera de nouveau remise en cause à l'étape suivante.
4. À la fin de l'itération, le critère d'arrêt est testé : l'indexation actuelle est-elle meilleure que l'indexation précédente ? Si un gain est constaté, une nouvelle itération commence.

[2]Sorted Weighted Geometrical Mean

NB : En pratique, le système teste un deuxième critère d'arrêt : Reste-t-il des blocs étiquetés Parole pour la sélection et l'ajout d'un nouveau modèle de locuteur ?

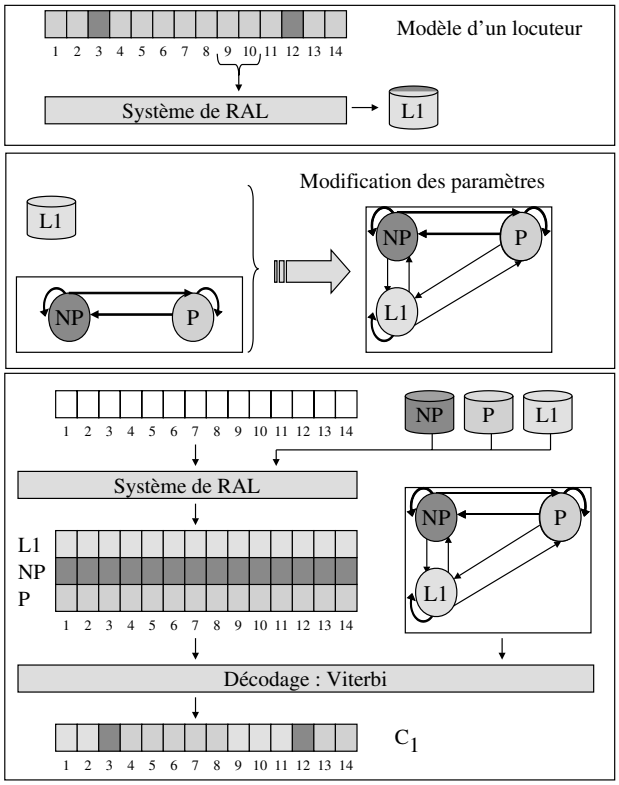

Figure 2: *Processus itératif : traitement du premier locuteur détecté*

3. Système de vérification du locuteur

Les modèles de locuteurs employés et l'ensemble des probabilités d'émission sont calculés par le système de reconnaissance du locuteur AMIRAL[3] [5], développé au LIA.

La paramétrisation acoustique (coefficients cepstraux) est effectuée grâce au module développé par le consortium ELISA[4].

Les locuteurs sont modélisés par des mixtures de gaussiennes (GMM à 16 composantes avec une matrice de covariance pleine [7]) apprises par l'algorithme EM[5] [6] (critère Maximum Likelihood).

Une spécificité d'AMIRAL est de considérer le signal de parole au niveau de blocs de 0,3 s (= 30 trames), sur lesquels des scores de vraisemblance normalisés (par modèle du monde et MAP[6] [4]) sont calculés.

[3]Architecture Multi-reconnaisseurs pour l'Indexation et la Reconnaissance Automatique du Locuteur

[4]Le consortium ELISA est composé de laboratoires de recherche européens qui travaillent sur une plate-forme commune. Les laboratoires ayant participé à NIST 1999 sont : ENST (France), EPFL (Suisse), IDIAP (Suisse), IRISA (France), LIA (France), RIMO — Rice (USA) et Mons (Belgique) —, RMA (Belgique), VUTBR (République Tchèque).

[5]Expectation-Maximization

[6]Maximum A Posteriori

La normalisation MAP employée permet d'obtenir des scores assimilables à des probabilités. AMIRAL prend également en charge la sélection des modèles à l'aide de l'algorithme SWGM, adapté à l'identification du locuteur à partir d'enregistrement pluri-locuteurs.

4. Expériences

4.1. Ensembles de données

La méthode proposée dans cet article a été expérimentée sur un sous-ensemble de données issues de la campagne d'évaluation NIST 1998. Les messages ont été simulés à partir de la concaténation de segments de parole téléphonique (conversation réelle) mono-locuteur issus de la base Switchboard.

Deux sous-ensembles indépendants (définis par le consortium ELISA) sont utilisés :

- Un premier ensemble de 25 locuteurs est consacré au développement (Dev).
- Un ensemble Eva permet la validation des paramètres mis au point sur Dev. Ce corpus a la même taille et la même structure que Dev, mais les populations de locuteurs sont disjointes.

Pour chaque sous-ensemble, nous disposons de 2 minutes de parole par locuteur pour l'apprentissage du modèle et de 30 secondes pour la génération des messages de test.

4.2. Génération des messages

Les messages sont générés par concaténation de blocs (0,3s) de parole mono-locuteur. La méthode utilisée est :

- l différents locuteurs sont sélectionnés parmi les 25 locuteurs disponibles.
- i ($avec\ i \geq l$) différents segments sont choisis, tels que chaque locuteur soit présent au moins une fois.
- La durée d de chaque intervention est déterminée.

l, i and d sont des nombres tirés aléatoirement dans des distributions gaussiennes (Table 1). La sélection des locuteurs, l'ordre d'apparition des segments et le choix des segments sont générés à partir de distributions uniformes.

5000 messages ont été générés pour chaque corpus.

NB : Il n'existe pas de situation où deux locuteurs parlent en même temps.

Parameter	Moyenne	écart type
l	5	1
i	15	2
d (# 0,3s)	10	2

Table 1: *Paramèters pour la génération des données*

4.3. Tests et résultats

Les critères sélectionnés pour mesurer la performance du système sont :
• Le pourcentage de blocs correctement indexés (BC). Le pourcentage correspond au taux de blocs pour lesquels le système propose une identité identique à l'identité réelle.
• Le pourcentage de blocs mal indexés (BM) : Il est calculé sur les blocs pour lesquels le système ne propose pas une identité identique à l'identité réelle.

Un taux d'erreur d'attribution (équ. 1) d'un bloc à un état (EA) est calculé à partir de ces deux valeurs :

$$EA = \frac{BM}{BM + BC} \quad (1)$$

Deux valeurs mesurent le taux d'indécision :
• Le pourcentage de segments étiquetés Non Parole (BNP).
• Le pourcentage de segments étiquetés Parole (BP).

Les résultats portés dans Table 2 montrent un taux sur Dev de 66% de blocs correctement indexés (respectivement 62,1% sur Eva).

	Décision			
Corpus	BC	BM	Total	EA
Dev	66,0%	28,6%	94,6%	30,2%
Eva	62,1%	26,4%	88,5%	29,8%

Table 2: *Résultats du système de suivi de locuteurs : Taux calculés sur Dev et Eva (5000 messages chacun). Pour tous les blocs : BC= % de blocs correctement indexés, BM= % de blocs mal indexés, EA= taux d'erreur d'attribution.*

Le taux d'erreur d'attribution (EA) reste élevé (environ 30% sur Dev et Eva) ; mais le taux d'erreur obtenu par une simple décision bayésienne[7] (équ. 2) atteint 54,5% d'erreur d'attribution sur Dev (et 53,8% sur Eva). Ce dernier résultat montre que le taux d'erreur d'affectation provient principalement de la difficulté intrinseque de la base Switchboard et des erreurs du système AMIRAL.

$$dec_{Bay}(i) = \underset{l \in M}{\mathrm{ArgMax}}(p(s_i|M_l)) \quad (2)$$

M est l'ensemble des modèles de locuteurs. s_i est le $i^{ième}$ bloc du signal.

Le taux d'indécision observé (blocs attribués à aucun modèle de locuteur) est de 11,5% (BP + BNP) sur EVA (Table 3, le taux de décision présenté dans Table 2 est rappelé). Le taux obtenu est faible au vu de la difficulté de la tâche.

Le taux de locuteurs détectés dans les messages est proche de 90% sur Dev et Eva. Néanmoins, le nombre de locuteurs ajoutés à tort au modèle de Markov

[7] pour un bloc donné, le locuteur le plus probable est choisi parmi les 25 locuteurs du corpus

Corpus	Indécision			Décision
	BNP	BP	Total	
Dev	0,9%	4,5%	5,4%	94,6%
Eva	1,6%	9,9%	11,5%	88,5%

Table 3: *Décision, Indécision du système de suivi de locuteurs : Taux calculés sur Dev et Eva (5000 messages chacun). Pour tous les blocs : BP= % blocs attribués à Parole, BNP= % blocs attribués à Non Parole,*

est important : environ 70% des locuteurs ajoutés sont des locuteurs qui ne font pas partie du message, que ce soit sur Dev ou Eva. Ces locuteurs représentent environ les 2/3 (représentant 20%) des erreurs d'attribution relevées durant les tests (EA $\simeq$ 30% sur Dev et Eva).

5. Conclusion

Dans cet article, le système de suivi de locuteurs utilise un modèle de Markov évolutif pour modéliser la conversation et pour déterminer automatiquement les locuteurs présents dans les messages. L'approche est basée sur un algorithme itératif qui détecte et ajoute les modèles de locuteur un à un. A chaque étape, une indexation est proposée, en fonction de l'ensemble des connaissances disponibles. Cette indexation est remise en cause à l'itération suivante jusqu'à l'optimal (détection de tous les locuteurs du message).

Les résultats obtenus sont encourageants au vu du taux d'erreur d'attribution et du taux de non décision. Cependant, trop de locuteurs absents des messages sont ajoutés au modèle de conversation.

Les travaux futurs devraient porter sur trois points :

- La méthode de sélection des locuteurs à ajouter doit être améliorée.
- Actuellement, le modèle de Markov n'utilise pas de modèle de durée explicite, l'utilisation d'un tel modèle devrait permettre de limiter les problèmes de sur-segmentation.
- Le système actuel est adapté aux tâches de suivi de locuteurs, il sera étendu à des tâches d'indexation, où les modèles de locuteurs doivent être construits, ou adaptés, à partir des données du message.

Bibliographie

[1] P. Delacourt, D. Kryze, C.J. Wellekens. Use of second order statistic for speaker-based segmentation, *EUROSPEECH*, 1999.

[2] H. Gish, H-H Siu, R. Rohlicek. Segregation of speakers for speech recognition and speaker identification, *ICASSP*, pages 873-876, 1991.

[3] K. Sönmez, L. Heck, M. Weintraub, Speaker tracking and detection with multiple speakers, *EUROSPEECH*, 1999.

[4] T. Matsui, S. Furui. Likelihood normalization for speaker verification using a phoneme and speaker-independent model, *Speech communication*, pages 109-116, August 1995.

[5] C. Fredouille, J.-F. Bonastre, T. Merlin, Similarity normalization method based on world model and a posteriori probability for speaker verification, *EUROSPEECH*, 1999.

[6] D. Dempster, N. Larid, D. Rubin, Maximum likelihood from incomplete data via EM algorithm, *J. Roy. Stat. Soc.*, Vol. 39, pp 1-38, 1977.

[7] D. A. Reynolds, Speaker identification and verification using gaussian mixture speaker models, *Speech Communication*, pp 91-108, Aug. 1995.

[8] J-F. Bonastre, P. Delacourt, C. Fredouille, S. Meignier, T. Merlin, C. Wellekens, Différentes stratégies pour le suivi de locuteur, *RFIA 2000*, Jan. 2000.