Différentes stratégies pour le suivi de locuteur
Various strategies for speaker tracking

Jean-François Bonastre[1] Perrine Delacourt* [2] Corinne Fredouille [1]
Sylvain Meignier [1] Teva Merlin [1] Christian Wellekens [2]

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

[2] Institut Eurécom, 2229 route des crêtes,
BP 193, 06904 Sophia Antipolis Cedex, France
{perrine.delacourt, christian.wellekens}@eurecom.fr

Résumé

Ce travail concerne le suivi de locuteur (Speaker Tracking), une tâche proche de l'indexation selon le locuteur. Le travail à réaliser consiste à détecter les segments de document prononcés par un locuteur spécifique. Dans ce cadre, le système possède une référence correspondant au locuteur concerné (cependant, les autres locuteurs présents dans le document ne sont pas connus) et seuls les segments de document prononcés par ledit locuteur sont à prendre en compte.

Dans cet article, nous avons exploré deux stratégies différentes pour élaborer des systèmes de suivi de locuteur. La première s'appuie sur un système d'indexation selon le locuteur. Elle consiste à opérer une détection des changements de locuteurs en amont du système, sans connaissance sur les locuteurs potentiels. Une fois le signal segmenté, un système classique de vérification du locuteur est appliqué à chaque segment obtenu et détermine si ce segment a été prononcé ou non par le locuteur cible. La deuxième solution est élaborée à partir d'un système segmental de reconnaissance du locuteur, dont seule l'étape de prise de décision est adaptée à la tâche visée. Dans ce cas, la décision sur la présence ou non du locuteur cible dans le document est réalisée globalement sur l'ensemble du document. La détection des segments correspondant à ce même locuteur est menée conjointement. Enfin, une amélioration de la dernière technique est discutée, particulièrement dans le cas d'un document contenant de multiples locuteurs.

*Ce travail a bénéficié du soutien du Centre National d'Etudes des Télécommunications (CNET) au titre du contrat n° 98 1B

Mots Clés

suivi de locuteur - reconnaissance du locuteur - indexation de documents

Abstract

This work addresses speaker tracking, which is closely related to speaker indexing. The task consists in detecting the recorded segments uttered by a given speaker. In this approach, only the model of the target speaker is available and only the documents uttered by this given speaker are taken into account.
In this paper, two different stategies are explored to set up systems for speaker tracking. The first one relies on a speaker indexing tool. Speaker turns are detected in the front-end of the system without any knowledge on possible speakers. Once the signal has been segmented, a classical speaker verification process is applied to each segment and checks if this segment corresponds to the target speaker. The second solution is worked out from a segmental speaker recognition system from which only the decision step is adapted to the task at the hand. In this case, decision on the presence of the target speaker in the record is based on the whole recorded document. Segments corresponding to the target speaker are simultaneously detected. Eventually, an improvement of this last technique is discussed, more specifically for documents containing multiple speaker utterances.

Keywords

speaker tracking - speaker recognition - document indexing

1 Introduction

Dans le cadre de l'indexation par le contenu de documents multimédia, la recherche du nombre de locu-

teurs présents dans ledit document ainsi que l'affectation des différents segments de parole au locuteur correspondant constitue une tâche essentielle. Ce processus, appelé "indexation selon le locuteur", montre une grande complexité car, en plus des difficultés classiques rencontrées en traitement automatique de la parole, aucune information sur le document à traiter n'est disponible *a priori*. En particulier, le nombre d'interventions différentes, la durée moyenne de ces interventions, le nombre de locuteurs et les caractéristiques des locuteurs potentiellement présents dans le document ne sont pas connus à l'avance par le système.
Ce travail concerne le suivi de locuteur (Speaker Tracking), une tâche proche de la précédente. Le travail à réaliser consiste à détecter les segments de document prononcés par un locuteur spécifique. Deux différences principales entre l'indexation selon le locuteur et le suivi de locuteur sont à noter :

- Le locuteur est ici connu au préalable ; le système possède une référence correspondant au locuteur concerné. Cependant, les autres locuteurs présents dans le document ne sont pas connus.

- Dans le cadre du suivi de locuteur, on ne s'intéresse qu'aux segments de document prononcés par ledit locuteur. Le nombre de locuteurs et les segments correspondant à des locuteurs différents du locuteur cible ne sont pas pris en considération.

Enfin, la tâche peut être étendue au suivi simultané de plusieurs locuteurs.
Dans cet article, nous avons exploré deux stratégies différentes pour élaborer des systèmes de suivi de locuteur. La première s'appuie sur un système d'indexation selon le locuteur. Elle consiste à opérer une détection des changements de locuteurs en amont du système. Une fois le signal segmenté, un système classique de vérification du locuteur est appliqué indépendamment sur chacun des segments obtenus pour déterminer s'il a été prononcé ou non par le locuteur cible. Cette approche privilégie la qualité de la détection des segments par rapport à la vérification d'identité (réalisée sur des segments potentiellement courts). La deuxième stratégie proposée est élaborée à partir d'un système segmental de reconnaissance du locuteur, dont seule l'étape de prise de décision est adaptée à la tâche visée. Dans ce cas, la décision sur la présence ou non du locuteur cible dans le document est réalisée globalement sur l'ensemble du document. La détection des segments correspondant à ce même locuteur est menée conjointement à la phase précédente. Cette seconde solution privilégie la vérification d'identité (réalisée sur le document entier) par rapport à la fiabilité de la détection des segments.
Enfin, dans la dernière partie de l'article, une amélioration de la seconde approche, modélisant les changements de locuteur par un modèle de Markov caché (HMM), est proposée.

2 AMIRAL : un système général pour la reconnaissance du locuteur

Les deux approches proposées s'appuient sur le système de reconnaissance du locuteur AMIRAL. AMIRAL est un système multi-reconnaisseurs segmental, dédié aux tâches de reconnaissance du locuteur telles que l'Identification Automatique du Locuteur (IAL), la Vérification Automatique du Locuteur (VAL), et plus récemment, au Suivi de Locuteurs (SL). AMIRAL est composé de différents modules détaillés dans les paragraphes suivants.

2.1 Pré-traitement

Pour la phase de pré-traitement du signal de parole, le système AMIRAL utilise le module de paramétrisation standard du consortium ELISA[1]. Le signal de parole est représenté toutes les 10 ms, par 16 coefficients cepstraux, dérivés d'une analyse en bancs de filtres. Une normalisation, basée sur le retrait de la moyenne cepstrale (Cepstral Mean Subtraction) permet de minimiser les perturbations dues aux différents canaux de transmission de la voix.

2.2 Modélisation du locuteur

Le système AMIRAL exploite différentes techniques statistiques de modélisation de la voix. Dans cette étude, les locuteurs sont chacun modélisés par un modèle mono état. Les modèles de locuteurs utilisés sont des mixtures de gaussiennes (Gaussian Mixture Models [6]), entraînées à l'aide de l'algorithme EM (Expectation-Maximization [5]) basé sur le principe du maximum de vraisemblance. Les modèles sont constitués de 16 composantes, caractérisées par des matrices de covariance pleines. Enfin, la mesure de similarité utilisée entre un vecteur de paramètres décrivant une trame de signal et un modèle consiste à calculer la vraisemblance pour que ladite trame ait été émise par le modèle considéré.

2.3 Approche bloc-segmentale et normalisation

Un des aspects spécifiques du système AMIRAL est de considérer le signal de parole à un niveau segmental. Comme les segments considérés sont de taille fixe et de très courte durée (0,3 seconde), cette approche est nommée "bloc-segmentale". Durant la phase de test,

1. Le consortium ELISA est composé de laboratoires de recherche européens travaillant sur une plate-forme de référence pour l'évaluation des systèmes de reconnaissance du locuteur. Ces laboratoires sont: ENST (France), EPFL (Suisse), IDIAP (Suisse), IRISA (France), LIA (France), VUTBR (République Tchèque), RMA (Belgique), RIMO (Rice (Etats Unis) et Mons (Belgique)).

le signal de parole est découpé en blocs temporels, sur chacun desquels une mesure de similarité normalisée est calculée. Cette architecture permet de renforcer la robustesse du système en supprimant les zones non informatives lors de la décision. Elle a permis également une adaptation aisée d'AMIRAL aux tâches d'indexation ou de suivi de locuteur.
La normalisation appliquée au niveau de chaque bloc a pour but de pallier les problèmes de variabilité classiques en reconnaissance du locuteur et surtout d'homogénéiser les mesures de similarité.
La méthode utilisée combine deux techniques. Une première étape consiste à calculer un rapport de vraisemblance, pour chaque bloc, en divisant la vraisemblance obtenue par rapport au modèle du locuteur considéré (le signal a été prononcé par ledit locuteur) par la vraisemblance de l'hypothèse inverse (le signal provient d'un autre locuteur), modélisée par un modèle général de locuteur (souvent nommé "modèle du monde"). Une seconde normalisation de type MAP (Maximum A Posteriori) est, ensuite, appliquée sur chaque rapport de vraisemblance. Cette normalisation a pour objectif de prendre en compte le comportement du reconnaisseur, en remplaçant le rapport de vraisemblance par la probabilité *a posteriori* que le locuteur cible ait prononcé le segment de signal correspondant. Cette normalisation nécessite d'apprendre le comportement du système sur un ensemble séparé de données de développement et de choisir différentes probabilités *a priori* décrivant les conditions de test. L'avantage majeur de cette normalisation consiste à proposer des scores bornés ayant un sens dans le domaine probabiliste [7][8].

2.4 Architecture multi-reconnaisseurs

La deuxième particularité du système AMIRAL est de reposer sur une architecture multi-reconnaisseurs. Deux grands types de reconnaisseurs sont à distinguer selon la nature des informations prises en compte. Le premier type s'intéresse principalement au domaine spectral qui est divisé en sous-bandes fréquentielles traitées indépendamment. Le deuxième type exploite les informations dynamiques du signal de parole en concaténant des trames successives sur une fenêtre temporelle de taille constante et en sélectionnant un sous-ensemble de paramètres jugés optimaux pour la caractérisation du locuteur. Les mesures de similarité normalisées, issues de chaque reconnaisseur sont alors fusionnées au niveau de chaque bloc temporel. L'étape de fusion met en œuvre les avantages du processus de normalisation explicité dans la section 2.3.

2.5 Stratégies de décision

Le dernier module du système AMIRAL est constitué d'une étape de fusion et d'une étape de décision. La première phase consiste à fusionner les différentes mesures de similarités obtenues à raison d'une par bloc temporel. Plusieurs stratégies sont employées, de la simple moyenne arithmétique (notée AM) à des méthodes spécifiques aux tâches de suivi de locuteur en passant par des stratégies "d'élagage" ("pruning"), éliminant les blocs sur lesquels un niveau de bruit important est détecté.

3 Différentes stratégies pour le suivi de locuteur

Le but du suivi de locuteur est de rechercher dans un enregistrement audio les paroles prononcées par le locuteur cible, en d'autres termes de détecter le début et la fin des segments de parole attribuables à celui-ci. Dans cette section, plusieurs approches sont proposées.
La première approche consiste à réaliser une phase de détection des changements de locuteurs proposant une segmentation du message en segments homogènes (prononcés par un seul locuteur). Un processus de vérification du locuteur est alors appliqué indépendamment sur chacun des segments obtenus. L'étape de présegmentation n'utilise aucune connaissance *a priori* des locuteurs engagés dans la conversation. Elle est décrite au paragraphe 3.1.
La deuxième solution proposée dans cet article exploite directement les possibilités segmentales du système AMIRAL. Une phase de décision spécifique au suivi de locuteur suit l'étape de calcul des mesures de similarité entre les différents blocs temporels du message et le modèle du locuteur cible. Les processus de segmentation et de vérification du locuteur cible sont ici unifiés au sein du procédé de reconnaissance du locuteur. Le système dérivé d'AMIRAL est détaillé au paragraphe 3.2.
Enfin, un raffinement de la deuxième technique est présenté au paragraphe 3.3. Cette solution est particulièrement dédiée aux documents comprenant un nombre important de locuteurs (conférences, films, etc.) et pour lesquels un suivi simultané de plusieurs locuteurs est nécessaire.

3.1 Segmentation préliminaire en locuteurs

L'utilisation d'une segmentation préliminaire en locuteurs avant le processus de vérification repose sur l'idée suivante : le score de vérification est plus fiable si les segments considérés ne comportent que des trames (vecteurs acoustiques) provenant d'un unique locuteur. De même, la longueur des segments influe très fortement sur la qualité des résultats (cf [4]). Ainsi, le but de la segmentation en locuteurs est de découper le signal de parole en plusieurs segments homogènes : chaque segment ne doit contenir des paroles prononcées par un seul locuteur. Cette segmentation consiste à détecter les changements de locuteurs et est réalisée sans tenir compte de la connaissance du locuteur cible

(le modèle de ce locuteur n'est pas exploité).

Détection d'un changement de locuteur. Etant données deux portions de signal paramétrisées (deux séquences de vecteurs acoustiques) $\mathcal{X}_1 = \{x_1, ..., x_i\}$ et $\mathcal{X}_2 = \{x_{i+1}, ..., x_{n_\mathcal{X}}\}$, nous considérons le test d'hypothèses suivant pour un changement de locuteur à l'instant i:

- H_0 : les deux portions sont relatives au même locuteur. Leur réunion est modélisée par un unique processus gaussien : $\mathcal{X} = \mathcal{X}_1 \cup \mathcal{X}_2 \sim \mathcal{N}(\mu_\mathcal{X}, \Sigma_\mathcal{X})$
- H_1 : chaque portion a été prononcée par un locuteur différent et est modélisée par un processus gaussien différent : $\mathcal{X}_1 \sim \mathcal{N}(\mu_{\mathcal{X}_1}, \Sigma_{\mathcal{X}_1})$ et $\mathcal{X}_2 \sim \mathcal{N}(\mu_{\mathcal{X}_2}, \Sigma_{\mathcal{X}_2})$

Le rapport de vraisemblance généralisé, R, entre les hypothèses H_0 et H_1 est défini par :

$$R = \frac{L(\mathcal{X}, \mathcal{N}(\mu_\mathcal{X}, \Sigma_\mathcal{X})}{L(\mathcal{X}_1, \mathcal{N}(\mu_{\mathcal{X}_1}, \Sigma_{\mathcal{X}_1}).L(\mathcal{X}_2, \mathcal{N}(\mu_{\mathcal{X}_2}, \Sigma_{\mathcal{X}_2})}$$

Ce rapport de vraisemblance généralisé a été utilisé dans [2][3] en identification du locuteur et a prouvé son efficacité. La distance d_R est obtenue en prenant le logarithme de l'expression précédente : $d_\mathrm{R} = -\log R$ ("distance" est ici un abus de langage car d_R ne vérifie pas les propriétés d'une distance).
Une valeur élevée de R (i.e. une faible valeur de d_R) signifie que la modélisation avec une seule gaussienne (hypothèse H_0) s'accorde mieux aux données. A l'opposé, une faible valeur de R (i.e. une forte valeur de d_R) indique que l'hypothèse H_1, i.e. la modélisation par deux gaussiennes, correspond mieux aux données. Dans ce cas, un changement de locuteur est détecté à l'instant i.

Détection de tous les changements de locuteurs. La distance d_R est calculée pour chaque couple de fenêtres de signal de même durée (environ 2 secondes). Ces fenêtres doivent être suffisamment longues pour estimer de manière fiable les paramètres des gaussiennes et suffisamment courtes pour faire l'hypothèse qu'elles ne contiennent les paroles que d'un seul locuteur. Ces fenêtres sont glissantes et sont déplacées à chaque itération d'un laps de temps fixe (environ 0,1 seconde) le long du signal paramétrisé, comme le montre la figure 1.
Les distances calculées pour chaque couple de fenêtres sont stockées pour former à la fin du processus une courbe de distances. Les pics les plus significatifs (en terme d'amplitude) de cette courbe sont alors détectés : ces pics correspondent aux changements de locuteur recherchés. Un maximum local de la courbe des distances est considéré comme significatif si les différences entre son amplitude et celle des minima situés de part et d'autre sont supérieures à un certain seuil (dépendant de la variance de la distribution des distances). Nous imposons également un intervalle de temps minimal entre deux changements de locuteurs consécutifs (cf figure 2). La détection des changements de locuteurs ne se fait donc pas en considérant l'amplitude absolue des pics mais plutôt en considérant leur facteur de forme, comme détaillé dans [1].

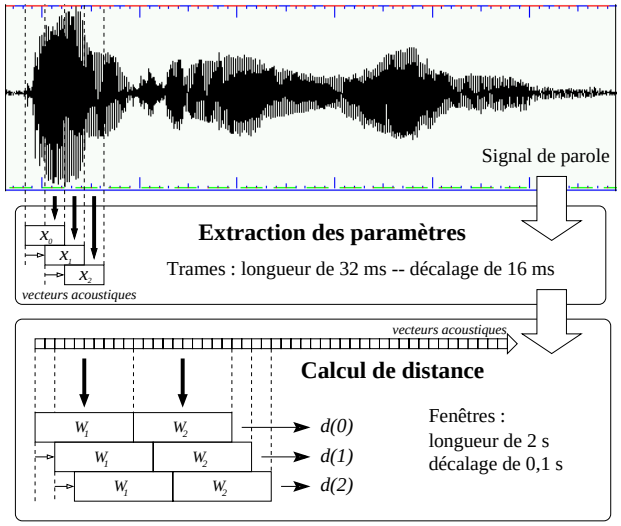

FIG. 1 – *Calcul de distance par fenêtres glissantes*

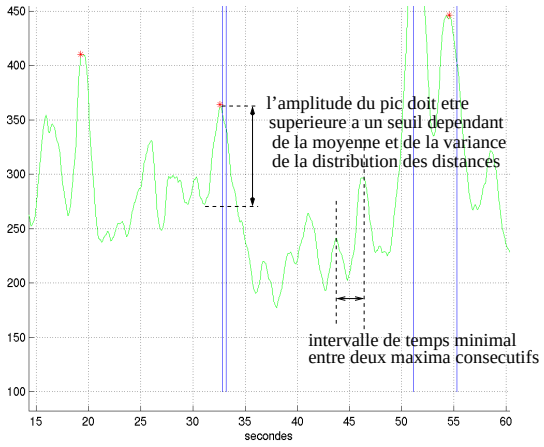

FIG. 2 – *Détection des points de changement de locuteur à l'aide du graphe de distances*

Une détection manquée (i.e. un changement de locuteur qui n'est pas détecté) est plus préjudiciable pour le processus de vérification qu'une fausse alarme (un changement est détecté alors qu'il n'existe pas). En effet, un segment contenant les paroles de plusieurs locuteurs (résultant d'une détection manquée) ne pourra être correctement identifié. Aussi, les paramètres impliqués dans la détection des changements de locu-

teurs ont été ajustés de manière à éviter les détections manquées au détriment du nombre de fausses alarmes. Le signal est probablement sur-segmenté : les paroles consécutives d'un même locuteur sont réparties sur plusieurs segments. Cependant, la durée des segments de locuteurs obtenus est suffisamment grande pour avoir une décision de vérification fiable.
Pour cette méthode, nous avons considéré que les segments temporels étaient mono-locuteur (et mono condition d'enregistrement) et nous avons configuré le système AMIRAL en conséquence : une simple moyenne arithmétique a été choisie comme procédé de fusion temporelle (pour obtenir la mesure de similarité finale, à partir des mesures provenant de chaque bloc temporel). Un procédé de décision classique, basé sur un seuil de décision, permet d'attribuer un segment (ou non) au locuteur cible.

3.2 Suivi de locuteur basé sur un système de reconnaissance du locuteur

Cette stratégie repose à la fois sur l'approche segmentale temporelle du système AMIRAL (section 2.3) et sur un algorithme de décision spécifique qui permet simultanément de choisir les zones de signal correspondant le mieux au locuteur cible et de décider si les informations retenues permettent d'identifier la présence dudit locuteur, sur l'ensemble du document. La technique proposée se décompose donc en deux phases :

1. Le système de reconnaissance du locuteur AMIRAL est mis en œuvre pour obtenir une mesure de similarité entre chaque bloc du signal et le modèle du locuteur cible;

2. Une stratégie de détection simultanée de la zone de décision appropriée et de prise de décision est alors mise en œuvre, à partir des données précédentes. Cette méthode est nommée SWGM (Sorted Weighted Geometric Mean). Elle consiste en quatre phases :

   - Une phase préliminaire, de tri des blocs dans l'ordre décroissant des mesures de similarité associées (i.e. : le plus probable en premier)
   - Une seconde phase permet la détection du sous-ensemble de blocs temporels optimal pour prendre la décision globale de présence du locuteur cible dans le document. Plus précisément, cette phase consiste à rechercher le sous-ensemble de blocs Eb tel que le score Sc attribué à l'ensemble du document soit maximal:

     $$Sc = f(Mg(Eb), Card(Eb))$$

     Avec : $Mg(Eb)$ correspondant à la moyenne géométrique des mesures de similarité associées aux blocs composant l'ensemble Eb. $Card(Eb)$ correspondant au cardinal de l'ensemble Eb. Et $f(x, y)$ une fonction pondérant la moyenne x en fonction du nombre y d'éléments à partir desquels x a été calculée. $f()$ correspond donc à une estimation de la confiance attribuée à la moyenne. Dans le cadre de cet article, $f(x, y) = x\sqrt[y]{0,1}$.
   - Une phase de décision est alors mise en œuvre. Cette phase compare le score Sc obtenu à l'étape précédente à un seuil ($Sdec$) prédéterminé. Si le score est inférieur au seuil, le document entier est rejeté (le locuteur cible n'est pas du tout présent dans ce document).
   - Enfin, dans le cas contraire, une étape d'extension du sous ensemble Eb est réalisée. Elle consiste à ajouter un à un les différents blocs dans Eb, toujours selon l'ordre décroissant des mesures de similarité associées aux blocs. Ce processus est arrêté dès que le score attribué à l'ensemble des blocs sélectionnés (par $f()$) est inférieur au seuil de décision $Sdec$.
     Finalement, ce nouvel ensemble de segments temporels est attribué dans sa totalité au locuteur cible.

NB:
Dans le cadre des évaluations NIST99 ([9]), une étape complémentaire de segmentation, facultative, a été utilisée pour prédécouper l'enregistrement en plusieurs zones. Cette étape, basée sur un seuil de rejet fixe permettant de détecter les zones très peu probables, était nécessaire vue la durée importante (plusieurs minutes) des enregistrements considérés. L'algorithme décrit dans le paragraphe précédent a ensuite été appliqué sur chacune des zones retenues.
L'avantage majeur de cette stratégie est de réaliser presque conjointement la décision dite "de vérification d'identité" et la segmentation. Cette étape de décision est plus robuste car réalisée sur l'ensemble des données présentes dans le document.

3.3 Détection simultanée de plusieurs locuteurs

La méthode présentée dans la section 3.2 offre l'avantage d'utiliser la connaissance *a priori* du locuteur cible durant les étapes de segmentation et de décision (réalisées simultanément). L'amélioration proposée ici concerne les documents contenant de multiples locuteurs. Dans ce cas, la tâche de suivi de locuteur est souvent réalisée pour différents locuteurs cibles.
Le principe novateur proposé ici consiste d'une part à réaliser simultanément l'ensemble des détections correspondant aux locuteurs cibles recherchés, en exploitant l'ensemble des modèles disponibles, et d'autre part, à modéliser les changements de locuteurs à l'aide d'un HMM.

Cette méthode réutilise la première étape de l'approche décrite dans la section 3.2. Les scores normalisés sont également calculés pour chaque bloc temporel mais maintenant pour l'ensemble des locuteurs connus du système (locuteurs cibles). De même, le système calcule des scores normalisés à partir d'un modèle générique de non parole (bruit, silence...) et un modèle générique de parole (appris sur un ensemble de données séparé). Un modèle HMM est alors construit, en associant un état à chaque locuteur cible. Deux états, correspondant respectivement au modèle de parole et au modèle de non parole sont ajoutés au modèle HMM. Un algorithme de type Viterbi attribue alors de manière optimale chaque bloc temporel à un des modèles. L'ensemble des règles utilisées pour définir la valeur des probabilités de transition du modèle HMM est exprimé sous forme d'une matrice contenant les poids de passage d'un état à un autre. Les poids choisis sont déterminés par l'opérateur en fonction des objectifs de la tâche d'indexation. En particulier, l'opérateur choisit le coût d'une erreur d'indexation d'un bloc (bloc attribué à un mauvais locuteur) par rapport au coût d'une non détection.
Les probabilités de transition vérifient trois conditions :

- La probabilité de transition entre les états doit être plus faible que la probabilité de rester dans le même état, car les interventions sont majoritairement composées de plusieurs blocs consécutifs.
- Les probabilités de rester dans le même état sont égales.
- Les locuteurs étant équiprobables, les probabilités de passer d'un état à un autre (différent) sont alors identiques.

Exemple : Transformation d'une matrice de poids en matrice de transition.
Soit la matrice de poids exprimée par le tableau 1.

Modèles	P.	Non P.	Loc. I	Loc. J ($\neq$ du Loc. I)
P.	5	1	5	5
Non P.	5	1	5	5
Loc. I	5	1	12	1

TAB. 1 – **Matrice des poids** *du modèle X (en ligne) vers le modèle Y (en colonne). P.: modèle de parole, Non P.: modèle de non parole, Loc. I, Loc. J: modèles des locuteurs I et J*

Pour un modèle de Markov à cinq états (Parole, Non Parole, Locuteur 1, Locuteur 2, Locuteur 3), le système construit la matrice de transition donnée par le tableau 2, où chaque poids est :

- reporté dans la matrice de transition;
- divisé par la somme marginale de la ligne, de sorte que la somme des probabilités des arcs sortant d'un état soit égale à 1 (propriété des modèles de Markov).

Modèles	P.	Non P.	Loc. 1	Loc. 2	Loc. 3
P.	5/21	1/21	5/21	5/21	5/21
Non P.	5/21	1/21	5/21	5/21	5/21
Loc. 1	5/20	1/20	12/20	1/20	1/20
Loc. 2	5/20	1/20	1/20	12/20	1/20
Loc. 3	5/20	1/20	1/20	1/20	12/20

TAB. 2 – **Matrice de transition** du modèle X (en ligne) vers le modèle Y (en colonne). P.: modèle de parole, Non P.: modèle de non parole, Loc. i: modèle du locuteur i

4 Expériences

4.1 Bases de données et protocoles d'évaluation

Les deux stratégies proposées ont été testées durant la campagne NIST/NSA99 d'évaluation des systèmes de reconnaissance du locuteur, qui proposait pour la première année des tests de suivi de locuteur.
Ces deux stratégies ont été élaborées de manière à respecter les conditions de cette évaluation : pour chaque test, il existe un seul locuteur cible et seule la connaissance sur ce locuteur est utilisée.
Dans ce contexte, les corpus utilisés sont composés d'enregistrements de conversations téléphoniques spontanées, issus d'un sous-ensemble du corpus Switchboard II, fourni dans le cadre de la campagne d'évaluation NIST/NSA99. Ce sous-corpus est composé de 230 hommes et 309 femmes. Les données d'entraînement pour chaque locuteur sont constituées de deux minutes de parole enregistrées sur deux sessions différentes.
L'apprentissage des modèles du monde, de bruit, et de parole, comme celui de la fonction de normalisation, sont réalisés en utilisant un corpus complètement séparé (cf section 2.3).
Les deux approches ont été évaluées durant la campagne NIST99, à l'aide de 4000 fichiers d'essai d'une minute de parole chacun.
La dernière méthode a été testée sur un corpus artificiel de 5000 messages, constitué en mélangeant des enregistrements mono-locuteur issus de la même base que précédemment (NIST/Switchboard).

4.2 Résultats et commentaires

La figure 3 montre les résultats obtenus par chacun des participants à la campagne NIST/NSA99, et en particulier par les deux systèmes présentés ici. Ces résultats sont fournis sous forme d'une courbe montrant

les fausses acceptations (blocs attribués à tort au locuteur cible) en fonction des faux rejets (blocs prononcés par le locuteur cible et rejetés à tort par le système). Les résultats sont calculés à raison d'une décision tous les centièmes de seconde.

Ces résultats montrent peu d'écart entre les différents compétiteurs de la campagne NIST/NSA99. La différence entre les sytèmes est masquée d'une part par la difficulté intrinsèque de la tâche (les écarts entre les différents compétiteurs lors de la campagne NIST99 étaient très faibles, pour le suivi de locuteur) ainsi que par le mode de calcul des performances, qui pénalise de la même manière toute erreur. En particulier, une erreur de positionnement d'une frontière de segment, de 1/100 de seconde, est comptabilisée au même niveau qu'une fausse détection de segment.

Au regard de ces différents points, la comparaison des résultats des deux systèmes ne permet pas de mettre en avant l'une ou l'autre des stratégies proposées.

Les résultats correspondant à la variante dédiée à la détection simultanée de plusieurs locuteurs sont nettement plus encourageants. Ce système a été capable d'attribuer 77 % des blocs (de 0,3 s) au locuteur cible correspondant, avec simultanément 8 % d'erreur d'affectation de blocs (blocs attribués à un mauvais locuteur), sur un total de 810047 blocs à affecter. Ces résultats doivent être nuancés par les conditions de l'expérience : tous les modèles de locuteurs étaient connus du système et le corpus de test, bien qu'issu du même ensemble de données que pour les expériences précedentes, était construit artificiellement.

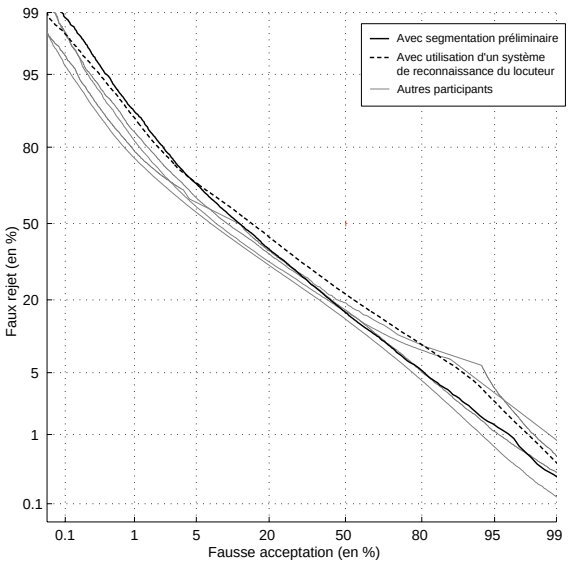

FIG. 3 – *Résultats obtenus lors de la campagne d'évaluation NIST99 – Taux de faux rejet en fonction du taux de fausse acceptation (taux calculés trame par trame)*

5 Conclusion

Nous avons présenté dans cet article deux approches très différentes dans le cadre d'un système de suivi de locuteur. Si les conditions expérimentales n'ont pas permis de conclure définitivement sur les avantages respectifs de ces méthodes, il apparaît clairement que développer une technique de suivi de locuteur à partir d'un système de reconnaissance du locuteur, sans segmentation au préalable, est une voie d'investigation intéressante.

Enfin, l'association d'un tel système et d'un HMM a montré un très bon niveau de performances. Il reste cependant à tester cette option dans le cas où tous les locuteurs ne sont pas connus du système.

Références

[1] P. Delacourt, D. Kryze, C.J. Wellekens, Speaker-based segmentation for audio data indexing, *ESCA workshop: accessing information in audio data*, 1999.

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

[3] H. Gish, N. Schmidt, Text-independent speaker identification, *IEEE Signal Processing magazine*, pp 18-32, Oct. 1991.

[4] I. Magrin-Chagnolleau, A.E. Rosenberg, S. Parthasarathy, Detection of target speakers in audio databases, *ICASSP*, 1999.

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

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

[7] C. Fredouille, J.-F. Bonastre, T. Merlin,

Segmental normalization for robust speaker verification, *Workshop on robust methods for speech recognition in adverse conditions*, 1999.

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

[9] M.A. Przybocki, A.F. Martin, NIST Speaker Recognition Evaluation 1997, *RLA2C*, pp 120-123, Apr. 1998.