Segmentation en locuteur des documents sonores par approches hybrides

Hachem KADRI[1] *Zied LACHIRI*[1, 2] *Noureddine ELLOUZE*[1]

[1]Unité Signal, Image et Reconnaissance de formes, ENIT, BP 37, 1002 le Belvédère, Tunis, Tunisie
Emails : kadri_hachem@yahoo.fr, N.ellouze@enit.rnu.tn
[2] Département Physique et Instrumentation, INSAT, BP 676, Centre Urbain Cedex, 1080, Tunis, Tunisie
Email : zied.lachiri@enit.rnu.tn

ABSTRACT

This paper deals with a new technique, DIS_T^2_BIC, for audio speaker segmentation when no prior knowledge of speakers is assumed. This technique is based on a hybrid concept which is organized in two steps: the detection of the most probable speaker turns and the validation of turns already detected. Our new technique uses a distance measure algorithm based on the Hotelling's T^2-Statistic criterion. The validation is obtained by applying the Bayesian Information Criterion (BIC) segmentation algorithm to the detected speaker turns. For measuring the performance we compare the segmentation results of the proposed method versus recent hybrid techniques. Results show that DIS_T^2_BIC method has the advantage of high accuracy speaker change detection with a low computation cost.

1. INTRODUCTION

La segmentation audio selon l'identité des locuteurs est une tâche qui devient de plus en plus nécessaire dans des domaines variées allant de l'adaptation en locuteurs pour les systèmes de reconnaissance de parole jusqu'aux tâches d'indexation dans le traitement des données multimédia [4,5,7,10]. Le but de la segmentation en locuteurs est de détecter les points de changement de locuteurs dans un document sonore sachant qu'aucune connaissance a priori n'est supposée disponible sur le nombre et l'identité des locuteurs potentiellement présents dans le document. Sous ces conditions, deux approches de segmentation sont communément utilisées : la segmentation par détection de silence et la segmentation par détection de changement de locuteur [4,5].

La segmentation par détection des silences suppose l'existence des silences entre les interventions des locuteurs et utilise un calcul. Par ailleurs, la segmentation par détection de changement de locuteurs repose sur la détection des et des changements de caractéristiques acoustiques. Cette procédure ne détermine pas les caractéristiques présentes dans le signal mais seulement leurs changements. Les techniques utilisées reposent essentiellement sur des mesures de distance [9] ou sur des critères de sélection de modèles [4] en particulier le critère d'information bayésien BIC [4]. Des études ont montré que ce critère offre la possibilité de détection des changements avec un seuil systématique tout en préservant un taux de détection élevée. Toutefois, ce critère peut introduire des erreurs d'estimations dues à l'insuffisance des données dans le cas où les changements de locuteurs sont proches les uns des autres. Pour s'affranchir de ces conditions et éliminer les problèmes d'estimation, une solution consiste à associer une technique par mesure de distance avec une autre basée sur un critère de sélection de modèle. Delacourt [7] propose l'association du rapport de vraisemblance généralisé et du critère BIC afin d'améliorer la détection des changements essentiellement ceux proches les uns des autres. Toutefois la technique qu'elle a réalisée, nommé DISTBIC [7], nécessite un seuil de détection ajustable à partir des expériences expérimentales et requiert un temps de traitement élevé. Zhou [11] a suggéré d'associer le critère hotelling's T^2 et le critère BIC pour développer la technique T^2-BIC [11]. Cette technique a permis la détection des changements de locuteurs à fur et à mesure tout en réduisant la complexité des traitements. Néanmoins, T^2-BIC n'arrive pas à détecter les changements de courte durée et dépend de quelques paramètres empiriques difficiles à estimer. Dans ce papier, nous proposons une technique de segmentation basée sur une approche hybride utilisant les critères Hotelling's T^2 et BIC permettant d'améliorer les taux de détection et de réduire la dépendance des paramètres empiriques et le temps de traitement.

Ce papier est organisé comme suit : la section 2 présente la segmentation en locuteurs par approche hybride. La section 3 décrit la technique de segmentation réalisée. Nous présentons les résultats expérimentaux obtenus dans la section 4, et enfin notre conclusion et nos perspectives sont détaillées dans la section 5.

2. SEGMENTATION AUDIO PAR APPROCHE HYBRIDE

Généralement, les techniques de segmentation par mesure de distance sont non stable et nécessitent des seuils de décisions déterminés à partir des résultats expérimentaux, quant aux techniques de segmentations basées sur des critères de sélection de modèle peuvent introduire des erreurs d'estimations à cause de l'insuffisance de données dans le cas où les changements de locuteurs sont proches les uns des autres. Ainsi, l'association de ces deux techniques nous offre la possibilité de détecter plus de changement tout en minimisant les effets du seuil de décision. La segmentation par approche hybride consiste à

associer d'une manière complémentaire deux techniques de segmentation, l'une basée sur une mesure de distance et l'autre basée sur un critère de sélection de modèle (figure 1).

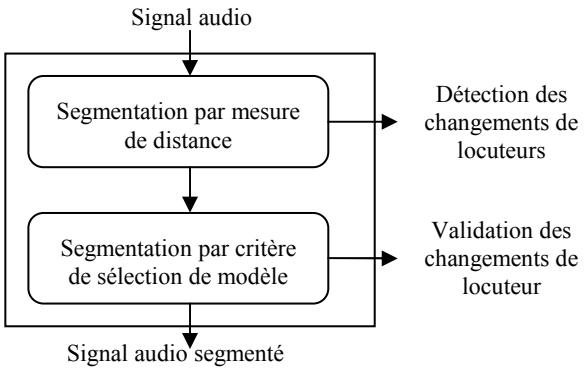

Figure 1 : Principe de la segmentation par approche hybride

L'algorithme procède en deux étapes. Premièrement une segmentation par mesure de distance est appliquée sur le signal audio et qui consiste à détecter les points de non similarité. Elle est basée sur le calcul d'une distance, à un instant t, entre deux blocs consécutifs du signal audio. La présence d'un changement de locuteurs à cet instant est d'autant plus probable que la valeur de la mesure de distance est élevée. La décision de l'existence d'une rupture est prise en comparant la valeur de la mesure à un seuil donné. Au cours de cette étape, les paramètres sont ajustés d'une manière à favoriser la détection des changements de locuteurs. La seconde étape permet d'éliminer les points de changement erronés par application d'un critère de sélection de modèle basée sur une décision statistique entre deux fenêtres du signal audio. Le critère d'information bayésien est un critère de sélection de modèle très utilisé pour la segmentation audio. Il permet, à partir des mêmes données, de choisir un modèle parmi plusieurs. En supposant que les données sont générées par un processus Gaussien, les changements de locuteurs sont détectés en comparant deux hypothèses :

- Les deux fenêtres contiennent des données générées par la même distribution
- Les deux fenêtres contiennent des données générées par deux distributions différentes

Soit $X = \{x_1,...,x_n\} \subset R^d$ une séquence de vecteurs cepstraux représentant un signal qui contient au plus n changement de locuteur. Si on suppose que X est généré par un processus Gaussien, un changement de locuteur est détecté au point $i \in \{1,...,n\}$ en calculant la valeur de ΔBIC à cet instant.

$$\Delta BIC(i) = \frac{n}{2}\log|\Sigma_X| - \frac{i}{2}\log|\Sigma_{X_1}| - \frac{(n-i)}{2}\log|\Sigma_{X_2}| - \lambda P. \quad (1)$$

Σ_X, Σ_{X_1} et Σ_{X_2} sont respectivement les matrices de covariances des vecteurs X, $X_1 = \{x_1,...,x_i\}$, $X_2 = \{x_i,...,x_n\}$ et $P = \frac{1}{2}\left(d + \frac{1}{2}d(d+1)\right)\log(n)$. λ est un facteur de pénalité et d est la dimension des vecteurs cepstraux. La valeur de i qui maximise $\Delta BIC(i)$ est le point de changement le plus probable et si $\Delta BIC(i) > 0$ alors i est l'instant de changement de locuteur.

3. SEGMENTATION DIS_T²_BIC

La technique de segmentation proposée se déroule en deux étapes. La première étape présente un nouvel algorithme de détection des changements de locuteurs qui associe les caractéristiques de la segmentation T^2 [2,10] et la segmentation basée sur le calcul d'une distance. La deuxième étape est une validation des changements de locuteurs déjà détectés lors de la première étape, fondée sur le critère BIC. L'algorithme de la technique DIS_T^2_BIC est représenté dans la figure 2.

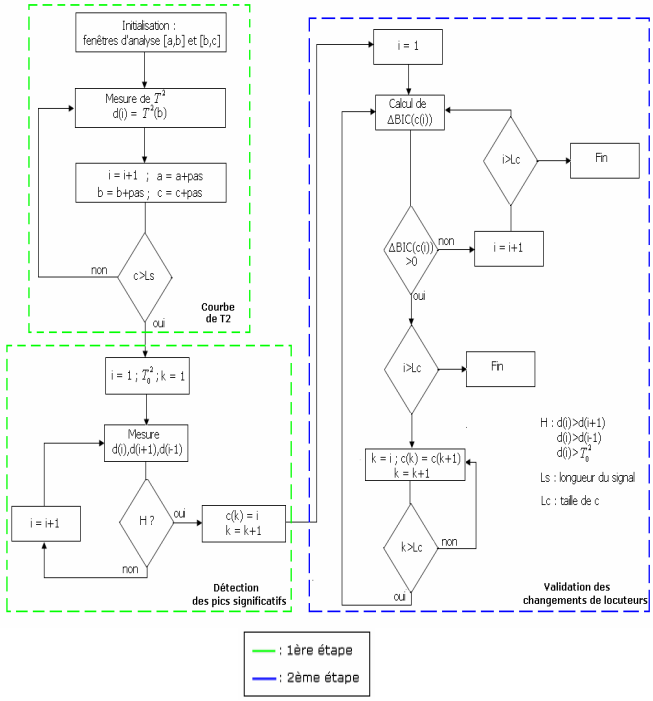

Figure 2 :Organigramme de la segmentation DIS_T^2_BIC

3.1. Application à la détection des changements de locuteurs

Hotelling's T^2 est une représentation multivariable de la distribution t de Student [2]. Une des applications de T^2 est de tester l'hypothèse que la moyenne μ_1 d'une population normale est égale à la moyenne μ_2 d'une autre dont le cas où les matrices de covariance Σ_{X_1} et Σ_{X_2} sont supposées égales mais inconnus. En terme de segmentation, le problème peut être résumé comme suit : soit une séquence de vecteurs acoustiques de longueur N, $X = x_i \in R^d, i = 1,2,...,N,$ contenant deux portions. L'une contient les b premières vecteurs cepstraux [1,b] de

moyenne μ_1 et l'autre contient les vecteurs cepstraux [b+1,N] de moyenne μ_2. Décider que ces deux portions sont homogènes, revient à tester l'hypothèse $H_0 : \mu_1 = \mu_2$ contre l'hypothèse $H_1 : \mu_1 \neq \mu_2$. Le rapport de vraisemblance associé à ce test d'hypothèse est déterminé à partir de la distribution statistique T^2 suivante :

$$T^2 = \frac{b(N-b)}{N}(\mu_1 - \mu_2)' \Sigma^{-1}(\mu_1 - \mu_2) \quad (2)$$

La méthode proposée détecte les changements de locuteur à partir de la mesure d'une valeur de T^2 entre deux portions du signal paramétré. Un couple de fenêtres adjacentes est déplacé tout au long du signal audio. Pour chaque couple une mesure de T^2 est calculée afin d'obtenir la courbe représentant la variation de T^2 au cours du temps. L'analyse de cette courbe montre l'existence de plusieurs pics d'amplitudes variées. Un changement de locuteurs se traduit par la présence d'un pic prononcé de la courbe de T^2. Ainsi, un changement de locuteurs peut être facilement détecté en recherchant les maxima de la courbe de T^2.

3.2. Choix du seuil de décision

La courbe de T^2 présente des pics de faibles amplitudes et d'autres d'amplitudes élevées. Seulement les pics prononcés correspondent à la présence d'un changement de locuteurs, il est donc nécessaire de fixer un seuil de décision afin de valider uniquement les pics significatifs. Les propriétés de la distribution T^2 nous permettent de définir un seuil automatique pour séparer un pic prononcé d'un pic de faible amplitude.

Soit $X = x_i \in R^d, i = 1,2,..., N$, tel que X suit une loi normale $N(\mu, \Sigma)$. Le rapport de maximum de vraisemblance associé au test $H : \mu = \mu_0$, à partir du critère statistique Hotelling's T^2 permet de définir une région critique dans laquelle l'hypothèse H est rejetée. Cette région est limité par T_0^2 :

$$\begin{aligned} T^2 \geq T_0^2 &= \frac{(N-1)p}{N-p} F_{p,N-p}(\alpha) \\ &= T_{p,N-1}^2(\alpha) \end{aligned} \quad (3)$$

α étant le niveau de confiance, la distribution $F_{p,N-p}$ est le rapport de deux distributions χ^2 de degré de liberté respectifs p et $N-p-1$. Une valeur de T^2 supérieur à T_0^2 implique que l'hypothèse H est rejetée. De ce fait, deux portions du signal audio, qui contiennent respectivement N_1 et N_2 échantillons, sont homogènes s'ils présentent une valeur de T^2, déterminée à partir de l'équation 2, inférieur à la valeur limite T_0^2. Ainsi, la détection de changements de locuteurs se ramène donc à la recherche des maxima de la courbe T^2 vérifiant le critère 3.

La détection des changements de locuteur à partir de la courbe T^2, nous offre la possibilité de détecter des changements proches les uns des autres tout en réduisant le temps de traitement. De plus, l'utilisation du critère Hotelling's T^2 permet de définir un seuil automatique indépendant de la nature du signal audio.

3.3. Amélioration à l'aide du critère BIC

Afin de réduire le nombre de fausse alarme lors de la première étape, nous utilisons le critère BIC pour valider les changements déjà détectés [7]. Si $\{s_1, \ldots, s_N\}$ est l'ensemble des points de changements potentiels résultant de la première étape, une valeur de ΔBIC est calculée pour chaque couple de fenêtres $[s_{i-1}, s_i]$ $[s_i, s_{i+1}]$. Si la valeur est positive, un changement de locuteur est détecté à l'instant i. Sinon, le point s_i est retiré de l'ensemble des points de changements potentiels, de telle sorte que la prochaine valeur de ΔBIC est calculée sur le nouveau couple de fenêtres $[s_{i-1}, s_{i+1}]$ $[s_{i+1}, s_{i+2}]$.

4. VALIDATION EXPÉRIMENTALE

Pour évaluer l'algorithme de segmentation proposé, nous avons utilisé deux types de document sonores.

- Une conversation créée artificiellement à partir de la base de données TIMIT (parole propre avec 48 changements de locuteurs),

- Une conversation créée en concaténant des phrases de la base ARABE [3] contenant des sons arabes phonétiquement équilibrés (parole propre avec segments courts et 240 changements de locuteurs),

- Deux émissions télévisées enregistrées de la chaîne Aljazeera [1] (segments de toutes les longueurs, parole spontanée et préparée avec 50 changements de locuteurs),

- Une conversation réalisée en groupant des documents sonores effectuées dans l'institut IDIAP [6] (segments de toutes les longueurs, parole spontanée avec 85 changements de locuteurs).

Les paramètres utilisés sont les coefficients MFCC, sans les Δ-coefficients (dérivées premières), calculés avec des fenêtres d'analyse de 20 ms et un recouvrement de 10 ms. L'évaluation des techniques de segmentation nécessite la quantification de deux types d'erreurs : le taux de fausses alarmes (TFA) et le taux de détections manquées. Une fausse alarme (FA) a lieu lorsqu'un changement de locuteur est détecté alors qu'il n'existe pas. Une détection manquée (DM) a lieu quand un changement de locuteur existant n'est pas détecté. Les taux de fausses alarmes et de détection manquée sont définis comme suit :

$$TFA = 100 \times \frac{nombre\ de\ FA}{nbre\ de\ changements\ réels + nombre\ de\ FA}\ \% \quad (5)$$

$$TDM = 100 \times \frac{nombre\ de\ DM}{nbre\ de\ changements\ réels}\ \% \quad (6)$$

La technique de segmentation DIS_T^2_BIC présente des taux de fausses alarmes et des taux détections manquées très proche de ceux de la segmentation DISTBIC. Plus

précisément, les résultats obtenus à partir de la conversation ARABE sont quasiment égaux. Ainsi, la présence de parole de courte durée et de changements de locuteurs proches les uns des autres ne pose pas un problème pour ces deux techniques. Nous déduisons que le principe de détection de changement de locuteurs à partir de la courbe de distance ou de la courbe de T^2 est efficace pour segmenter des documents audio qui présentent ces caractéristiques (par exemple conversation téléphonique). Contrairement, le TDM de la segmentation T^2-BIC pour ARABE est élevé et atteint les 40%.

La segmentation de la conversation TIMIT montre la difficulté de ces techniques d'éviter la détection d'un silence situé dans la parole d'un même locuteur. Le TFA pour TIMIT est environ de 30% pour les trois techniques et T^2-BIC présente le taux de TFA le plus faible. En effet, DISTBIC et DIS_T^2_BIC sont conçu d'une manière à favoriser le plus de détection possible afin de réduire au minimum le TDM au détriment du TFA. La présence des silences inter-locuteurs se traduit par un pic plus prononcé dans la courbe de T^2 que dans la courbe de distance ce qui justifie le fait que le TFA de la segmentation DIS_T^2_BIC plus élevé que celui de la segmentation DISTBIC.

Les résultats de la segmentation des documents réels, à savoir les documents d'Aljazeera et d'IDIAP se ressemblent. La technique de segmentation DIS_T^2_BIC présente le plus faible taux de détections manquées (38% pour le document d'Aljazeera et 41.17% pour le document d'IDIAP). Les TDM de la segmentation DISTBIC sont proches de ceux de DIS_T^2_BIC (44% pour le document d'Aljazeera et 45.17% pour le document d'IDIAP) contrairement aux TDM de la segmentation T^2-BIC qui dépassent les 50%. Par ailleurs le TFA de la segmentation T^2-BIC du document d'Aljazeera est plus faible que le TFA de la segmentation du document IDIAP (33% pour Aljazeera et 47% pour IDIAP). Cette variation est moins importante avec la segmentation DIS_T^2_BIC (40% pour Aljazeera et 42% pour IDIAP).

A partir des tests réalisés, nous remarquons que la segmentation DIS_T^2_BIC est plus appropriée que DISTBIC et T^2-BIC. Elle présente, essentiellement, les TDM les plus faibles. De plus, les paramètres y intervenant sont plus réduits et plus robuste puisqu'ils ne dépendent pas de la nature du document sonore et peuvent être déterminés d'une manière plus au moins systématique. Par ailleurs, l'utilisation du critère T^2 nous permet de définir un seuil fixe pour la détection des changements de locuteurs et d'accélérer le temps du traitement.

5. CONCLUSION

Dans cet article, nous proposons une méthode de segmentation basée sur l'association de la détection de changement de locuteurs par mesure de distance et par critère de sélection de modèle. Cette méthode utilise, en premier lieu, le critère Hotelling's T^2 pour élaborer la courbe de T^2. La localisation des pics élevés de cette courbe permet la détection des changements de locuteur les plus probables situés dans la conversation. La deuxième étape valide les changements déjà détectés à l'aide du critère BIC. La méthode proposée permet de détecter les changements de locuteurs proches les uns des autres en utilisant un seuil fixe et donc indépendant de la nature des documents audio.

Dans des travaux avenir, nous envisageons d'améliorer le seuil de détection. Le seuil utilisé est déterminé à partir d'un test d'hypothèses qui permet de définir deux régions une dans laquelle le test est vérifié et l'autre dans laquelle le test rejeté.

Tableau 1 : *TFA et TDM des techniques DISBIC, T^2-BIC et DIS-T^2-BIC*

	DISTBIC		***T^2-BIC***		***DIS_T^2_BIC***	
	FAR (%)	*MDR (%)*	*FAR (%)*	*MDR (%)*	*FAR (%)*	*MDR (%)*
TIMIT	30.43	**6.25**	26.15	15.58	34.24	**8.33**
ARABE	7.69	8.75	6.61	40.83	9.09	**6.66**
ALJAZ	36.70	44	33.33	54	40.47	**38**
IDIAP	45.16	45.88	47.85	50.58	42.17	**41.17**

BIBLIOGRAPHIE

[1] Aljazera broadcasting channel http://www.aljazeera.net

[2] T.Anderson, An introduction to multivariate statistical analysis, John Wiley & Sons, Inc., NY, 1985.

[3] M.Boudraa, B.Boudraa et B.Guerin, "Mise en place de phrases arabes phonétiquement équilibrées",XIXème JEP, Bruxelles, 1992.

[4] M.Cettolo et M.Federico, "Model selection criteria for acoustic segmentation", in Proc. ISCA Tutorial and Research Workshop ASR 2000, Paris, Sept 2000.

[5] S.Chen et, P.Gopalakrishnan, "Speaker, environment and channel change detection and clustering via the Bayesian Information Criterion", in Proc. DARPA Broadcast News Transcription, 1998.

[6] Dalle Molle Institute for Perceptual Artificial Intelligence (IDIAP), http://www.idiap.ch.

[7] P.Delacourt et C.J.Wellekens, "DISTBIC: a speaker based segmentation for audio data indexing", Speech Communication, vol. 32, pp. 111-126, Sept. 2000.

[8] R. Huang et J.H.L. Hansen, "Advances in unsupervised audio segmentation for the broadcast news and ngsw corpora" in Proc. ICASSP'04, Montreal, May 2004.

[9] M.Siegler, U.Jain, B.Raj, et R.M.Stern "Automatic segmentation, classification and clustering of broadcast news audio", in Proc. DARPA Speech Recognition Workshop, pp. 97-98, Chantilly, Virginia, USA, 1997.

[10] S. Wegmann, P. Zhan et L. Gillick, "Progress in Broadcast News Transcription at Dragon Systems", in ICASSP 99, Phoenix, Arizona, March 1999.

[11] B.W.Zhou et J.H.L.Hansen, "Unsupervised audio stream segmentation and clustering via the Bayesian Information Criterion," in Proc. ICSLP'2000, Vol. 1, pp.714-717, Beijing, China, Oct. 2000.