Segmentation en locuteurs d'un document audio

Perrine Delacourt et Christian J. Wellekens
Institut EURÉCOM, 2229 route des Crêtes,
BP 193, 06904 Sophia Antipolis Cedex, France
{perrine.delacourt,christian.wellekens}@eurecom.fr
http://www.eurecom.fr/~delacour/speech/

Résumé: Dans cet article, nous abordons le problème de la segmentation en locuteurs. Le but est d'obtenir des segments de locuteurs homogènes, c'est-à-dire ne contenant les paroles que d'un seul et même locuteur. Ces segments doivent être les plus longs possible. Dans notre étude, nous faisons les hypothèses qu'aucune connaissance a priori sur les locuteurs n'est disponible et que les personnes présentes dans la conversation ne parlent pas simultanément. Notre technique de segmentation s'effectue en deux passes : tout d'abord, les changements de locuteurs les plus probables sont détectés lors de la première passe pour être validés ou au contraire annulés lors de la seconde passe. Nous avons appliqué cette technique de segmentation à des données réelles et synthétiques. Les résultats de ces expériences démontrent l'efficacité de la technique à segmenter en locuteurs. Nous avons également comparé les performances de notre technique à une autre technique de segmentation. Pour les conversations contenant de longs segments de locuteurs, les deux techniques sont équivalentes. Par contre, notre technique est plus performante dans le cas de conversations contenant de courts segments de locuteurs.

1 Introduction

Le but de la segmentation en locuteurs est d'obtenir des segments de locuteur homogènes : chaque segment résultant ne doit contenir les paroles que d'un seul locuteur et être aussi long que possible. La segmentation en locuteurs n'a été étudiée que récemment dans la littérature [7, 1, 2, 13] en tant qu'étape préliminaire à plusieurs tâches d'indexation parmi lesquelles : la transcription automatique de journaux télévisés [14, 5], le regroupement automatique de messages [11] ou encore la poursuite de locuteur [12, 10].

L'algorithme de segmentation proposé dans cet article est conçu pour être intégré dans un système d'indexation par locuteurs. A partir d'un document audio, le système doit fournir en sortie la séquence de locuteurs présents dans la conversation en détaillant pour chacun d'eux les périodes durant lesquelles ils parlent. Le but n'est autre que de savoir qui parle et quand. Ce système d'indexation procède en deux étapes : le document audio est tout d'abord segmenté en locuteurs puis les segments apartenant à un même locuteur sont regroupés à l'aide de techniques décrites dans [11] ou [9]. Dans cet article nous présentons la première étape à savoir la segmentation en locuteurs. Nous faisons les hypothèses suivantes : aucune information sur les locuteurs n'est disponible (pas de modèle de locuteur, pas de phase d'entraînement)

et les personnes ne parlent pas simultanément.

Notre algorithme de segmentation se déroule en deux temps. Les changements de locuteurs les plus probables sont d'abord détectés à l'aide d'un algorithme de segmentation basé sur le calcul d'une distance. Ces points de changements potentiels sont ensuite validés ou au contraire annulés en utilisant le Critère d'Information Bayesien. Ce critère a été utilisé par S.Chen dans [2] pour de la segmentation en locuteurs mais il nécessite de longs segments de locuteurs (>3s).

Le paragraphe 2 détaille notre algorithme de segmentation. Les performances de cet algorithme sont évaluées au paragraphe 3 par le biais de critères présentés au paragraphe 3.2. Les résultats sont commentés au paragraphe 3.3. La comparaison de notre algorithme de segmentation avec l'algorithme proposé par S.Chen ([2]) est également faite au paragraphe 3.3. Enfin, le paragraphe 4 conclut et donne des perspectives possibles pour un travail futur.

2 Segmentation en locuteurs

Notre segmentation en locuteurs repose sur la détection des changements de locuteurs et non des silences inter-locuteurs ou des segments à l'aide de modèles de locuteurs. Elle se déroule en deux temps : une première passe se base sur le calcul d'une distance pour déterminer les changements de locuteur les plus probables et une seconde passe utilise le Critère d'Information Bayesien (CIB) pour valider ou annuler ces changements potentiels.

2.1 Segmentation basée sur le calcul d'une distance

2.1.1 Détection d'un changement de locuteur

Etant donné deux portions de signal paramétrisé (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 :

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

Ce rapport de vraisemblance généralisé a été utilisé dans [7, 6] 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 avec deux Gaussiennes correspond mieux aux données. Dans ce cas, un changement de locuteur est détecté à l'instant i.

2.1.2 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 sont suffisamment longues pour faire une estimation fiable des 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. Les distances calculées pour chaque couple de fenêtres sont stockées pour former à la fin du processus une courbe de distances. Nous nous appliquons ensuite à détecter les pics les plus significatifs (en terme d'amplitude) de cette courbe: 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. 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 [3].

Une détection manquée (i.e. un changement de locuteur existant n'est pas détecté) est plus préjudiciable pour l'étape suivante du système d'indexation par locuteurs qu'une fausse alarme (un changement est détecté alors qu'il n'existe pas): un segment corrompu (i.e. contenant plusieurs locuteurs) peut perturber l'étape de regroupement qui consiste à réunir les segments appartenant à un même locuteur. Aussi, les paramètres impliqués dans la détection des changements de locuteurs ont été ajustés de manière à éviter les détections manquées au détriment du nombre de fausses alarmes. Le signal est probablement sur-segmenté à l'issue de la première passe: les paroles consécutives d'un même locuteur sont réparties sur plusieurs segments. Une seconde passe utilisant le Critère d'Information Bayesien (CIB) est alors nécessaire pour réduire le nombre de fausses alarmes. Ce critère appliqué à la segmentation a été utilisé par S. Chen dans [2].

2.2 Raffinement avec le Critère d'Information Bayesien

Le CIB est un critère de vraisemblance pénalisé par la compléxité du modèle. Avec les mêmes notations que précédemment, le CIB est déterminé par: $\mathrm{CIB}(M) = \log L(\mathcal{X}, M) - \lambda \frac{m}{2} \log N_{\mathcal{X}}$, où $L(\mathcal{X}, M)$ est la vraisemblance de la séquence de vecteurs acoustiques $\mathcal{X}$ pour le modèle M, m est le nombre de paramètres du modèle M et λ le facteur de pénalité. Le premier terme reflète l'ajustement du modèle aux données et le deuxième terme correspond à la complexité du modèle. Ainsi, la modélisation qui maximise ce critère est conservée. Les variations de CIB entre les deux

modélisations (un processus Gaussien opposé à deux processus Gaussiens) est alors donnée par: $\Delta\text{CIB}(i) = -R(i) + \lambda P$ où $R(i)$ désigne le rapport de maximum de vraisemblance entre l'hypothèse H_0 (pas de changement de locuteur) et l'hypothèse H_1 (un changement de locuteur à l'instant i) et le terme de pénalité est donné par: $P = \frac{1}{2}(d + \frac{1}{2}d(d+1))\log N_{\mathcal{X}}$, d étant la dimension de l'espace acoustique, et λ le facteur de pénalité. Une valeur négative de Δ-CIB(i) indique que la modélisation avec les deux Gaussiennes correspond mieux aux données $\mathcal{X}$, ce qui signifie qu'un changement de locuteur a lieu à l'instant i.

Pour chaque paire de segments délimités par les points de changements trouvés lors de la première passe, une valeur de Δ-CIB est calculée. Si cette valeur est négative alors le changement de locuteur situé à la frontière commune aux deux segments est validé. Sinon, ce changement de locuteur est annulé et les deux segments sont réunis pour ne former qu'un à la prochaine itération.

3 Expériences

3.1 Données

Différents types de données de parole ont été utilisés pour comparer notre algorithme de segmentation avec l'algorithme proposé par S.Chen, appelé procédure CIB (cf [2, 3]):

- 2 conversations qui ont été créées artificiellement en concaténant des phrases de 2 secondes en moyennes extraites de la base de données TIMIT (parole propre, segments courts, anglais).
- 2 conversations créées en concaténant des phrases de 1 à 3 secondes extraites d'une bases de données fournie par le Centre National d'Etudes des Télécommunications (CNET) (parole propre, segments courts, français).
- 3 journaux télévisés extraits de la base de données de l'Institut National de l'Audiovisuel (INA) (segments de toute longueur, français).
- 3 conversations téléphoniques extraites de la base de données SWITCHBOARD ([8]) (segments de toute longueur, parole spontanée, anglais).

Pour les conversations synthétiques, les silences entre les différents locuteurs ont été réduits de manière à ressembler à des silences inter-locuteurs d'une conversation réelle. Nous avons également utilisé 4 journaux télévisés français (référencés *jt*) enregistrés dans notre laboratoire pour tester plus précisément notre approche.

Le signal de parole est paramétrisé avec 12 coefficients Mel-cepstraux. L'ajout des Δ-coefficients (dérivées premières) n'améliore pas les résultats et augmente le temps de calcul. Aussi, les Δ-coefficients ne sont pas utilisés (cf [4]).

3.2 Méthodes d'évaluation

Une bonne segmentation fournit les changements de locuteurs corrects et des segments ne contenant qu'un seul locuteur. Nous distinguons deux types d'erreur pour

la détection de changements de locuteurs. 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. Dans notre contexte, ce deuxième type d'erreur est plus grave que le premier type. En effet, nous avons vu au paragraphe 2.1.2 qu'un segment corrompu (i.e. contenant plusieurs locuteurs) pouvait détériorer l'étape de regroupement du système d'indexation par locuteurs. A l'inverse, une fausse alarme donc une sur-segmentation peut être résolue lors de cette même étape de regroupement des segments. Nous définissons le taux de fausses alarmes (TFA) comme suit :

$$\text{TFA} = \frac{\text{nombre de FA}}{\text{nbre de changements réels} + \text{nombre de FA}}$$

et le taux de détections manquées (TDM) par :

$$\text{TDM} = \frac{\text{nombre de DM}}{\text{nbre de changements réels}}$$

Une bonne segmentation est caractérisée par de faibles valeurs de TFA et de TDM.

3.3 Résultats et commentaires

Pour évaluer les performances de notre technique de segmentation, nous la comparons à la procédure CIB décrite dans [3]. Pour les deux techniques, nous indiquons le taux de fausses alarmes (TFA) et le taux de détections manquées (TDM). En ce qui concerne notre technique de segmentation, nous distinguons la segmentation basée sur la distance d_{R} (première passe) et le raffinement à l'aide du critère CIB (seconde passe). Le tableau 1 présente les résultats obtenus pour la procédure CIB appliquée à différents types de données décrits au paragraphe 3.1. Le tableau 2 présente les performances des deux passes de notre technique de segmentation appliquées aux mêmes données.

Les paramètres des deux techniques de segmentation ont été fixés pour chaque base de données. Leur valeur dépend essentiellement de la longueur des segments réels de locuteurs. Par exemple, plus ces segments sont longs, plus le paramètre λ (du critère CIB) doit être élevé.

Le TDM et le TFA de la procédure CIB (cf tableau 1) et de la deuxième passe de notre algorithme (cf tableau 2) respectivement, appliquées aux journaux télévisés de l'INA sont quasiment égaux. Cela signifie que les deux techniques de segmentation sont équivalentes avec des conversations contenant de longs segemnts de locuteurs. Nous pouvons également remarquer la baisse sensible du TFA entre la première et la seconde passe de notre algorithme. La segmentation basée sur la distance d_{R} est en fait sensible aux changements d'environnement sonore ou d'intonation du locuteur.

Les conversations téléphoniques (SWITCHBOARD dans les tableaux 1 et 2) contiennent également de longs segments mais de parole spontanée. En particulier, les conversations téléphoniques sont "clairsemées" de petits mots comme "Yeah" ou "Hum-hum". Quand ces mots sont prononcés alors que l'autre personne parle, notre hypothèse que les personnes ne parlent pas simultanément n'est pas respectée. Le

	CIB	
	TFA	TDM
TIMIT	31.5	30.5
CNET	14.3	50.0
INA	18.3	15.7
SWITCHBOARD	20.3	30.6

TAB. 1 – TFA *et* TDM *avec la procédure* CIB

	1^{rst} pass		2^{nd} pass	
	TFA	TDM	TFA	TDM
TIMIT	40.3	14.3	28.2	15.6
CNET	18.2	16.7	16.9	21.4
INA	37.4	9.03	18.5	13.5
SWITCHBOARD	39.0	29.1	25.9	29.1

TAB. 2 – TFA *et* TDM *respectivement avec la première et la seconde passes de notre technique de segmentation*

processus de segmentation se trouve détérioré par ces petits mots : ils ne sont en effet pas correctement détectés. De plus, ces petits mots ne sont pas pertinents dans le cadre de l'indexation par locuteurs : une intervention pour dire "Hum-hum" n'a pas de sens dans ce contexte. C'est pourquoi nous n'en tenons pas compte pour l'évalutation des deux techniques de segmentation. Cependant, la segmentation basée sur la distance (première passe) étant sensible aux changements d'environnements sonores, elle détecte dans la plupart des cas une des bornes de ces petits mots. C'est ce qui explique la valeur élevée du TFA de la première passe (cf tableau 2). Le TFA reste également plus élevé avec la seconde passe de notre algorithme qu'avec la procédure CIB (cf tableau 1). Par ailleurs, les TDM des deux techniques sont comparables.

Quant aux conversations contenant de courts segments (TIMIT et CNET dans les tableaux 1 et 2), notre technique de segmentation fournit de meilleurs résultats que la procédure CIB : pour ces conversations, le TDM est deux fois plus faible avec notre technique qu'avec la procédure CIB pour des valeurs de TFA comparables. Les conversations CNET sont faites de segments plus courts que les conversations TIMIT : cela explique le taux élevé de détections manquées. Nous pouvons aussi remarquer que les paramètres ne semblent pas dépendre de la langue. Ces paramètres sont quasiment identiques pour les conversations en anglais ou en français (CNET et TIMIT). Les faibles différences sont probablement dues aux conditions d'enregistrement.

Nos expériences montrent que notre technique de segmentation est plus précise que la procédure CIB en présence de segments courts, bien que les deux techniques aient les mêmes performances en présence de segments longs.

Nous avons mené d'autres expériences sur des journaux télévisés enregistrés dans notre laboratoire afin d'étudier les occurrences d'erreur. Les résultats sont présentés

	1^{rst} pass			2^{nd} pass		
	TFA	TDM	TD	TFA	TDM	TD
jt	59.0	8.9	8.4	23.7	9.4	8. 4

TAB. 3 – *Journaux télévisés :* TFA, TDM *et* TD *respectivement avec la première et la seconde passes*

dans le tableau 3. Pour évaluer plus finement notre technique de segmentation, nous définissons le taux de décalages (TD) : $\text{TD} = \frac{\text{nombre de décalages}}{\text{nbre de changements réels}}$

Un décalage est un changement de locuteur qui a été détecté à un instant décalé par rapport à sa position temporelle réelle. Un décalage correspond en fait à une fausse alarme et une détection manquée proches l'un de l'autre et qui ne devrait pas affecter le processus de regroupement. A la suite d'un décalage d'un changement de locuteur, l'un des segments contient les paroles de deux locuteurs. Cependant, la proportion de données d'un des locuteurs (de l'ordre de quelques dixièmes de seconde) est négligeable comparé au volume de données de l'autre locuteur (quelques secondes).

La plupart des détections manquées sont dues à de très courtes phrases, surtout durant les interviews : les questions des journalistes sont en général très brèves et elles ne sont pas détectées ou alors partiellement. En fait, les paramètres ont été ajustés pour détecter de longs segments de locuteurs, aussi les segments très courts ne sont pas toujours correctement détectés. Quant au TFA, sa valeur élevée s'explique par deux raisons prinicpales. Tout d'abord, quand une personne de langue étrangère est interviewée et que ses paroles sont traduites simultanèment ou plus exactement avec un léger décalage, cela créée des FA. (Remarque, l'une de nos hypothèses est dans ce cas non respectée). La deuxième raison est liée à la façon dont sont construits les reportages de journaux télévisés : les évènements sont commentés mais la bande son correspondant à ces évènements reste en fond sonore. Aussi, quand un changement d'environnement sonore intervient dans cette bande son, cela provoque bien souvent un FA.

Enfin, les taux que nous utilisons pour évaluer la segmentation permettent de quantifier les résultats mais ne reflètent pas la qualité de la segmentation. Bien que le TDM est loin d'être négligeable, les segments les plus significatifs (en termes de durée) sont détectés et leur écoute est tout à fait acceptable.

4 Conclusion et perspectives

Dans cet article, nous proposons une technique de segmentation composée d'une segmentation par calcul de distance suivie d'un raffinement à l'aide du critère CIB. Cette technique de segmentation est aussi efficace que la procédure CIB dans le cas de conversations contenant de longs segments de locuteurs et fournit de meilleurs résultats dans le cas de conversations contenant de courts segments. Nos expériences montrent également que les paramètres dépendent essentiellement de la longueur des

segments réels de locuteurs. Il reste cependant un problème : les paramètres peuvent être ajustés pour détecter plutôt des petits segments ou plutôt des longs segments mais pas les deux. Nos efforts vont maintenant consister à adapter les paramètres à la taille réelle des segments de locuteurs. Par ailleurs, cette technique de segmentation est destinée à être intégrée dans un système d'indexation par locuteurs. Aussi, notre travail futur va être de combiner l'étape de segmentation et l'étape de regroupement pour former le système complet d'indexation.

Références

[1] H.S.M. Beigi and Stéphane Maes. Speaker, channel and environment change detection. In *World congress of automation*, 1998.

[2] S.S. Chen and P.S. Gopalakrishnan. Speaker, environment and channel change detection and clustering via the Bayesian information criterion. In *DARPA speech recognition workshop*, 1998.

[3] Perrine Delacourt, David Kryze, and Christian J. Wellekens. Speaker-based segmentation for audio data indexing. In *ESCA workshop: accessing information in audio data*, 1999.

[4] Perrine Delacourt and Christian J. Wellekens. Audio data indexing: use of second-order statistics for speaker-based segmentation. In *ICMCS*, 1999.

[5] Jean-Luc Gauvain, Lori Lamel, and Gilles Adda. Partitioning and transcription of broadcast news data. In *ICSLP*, 1998.

[6] Herbert Gish and N. Schmidt. Text-independent speaker identification. *IEEE signal processing magazine*, oct. 1994.

[7] Herbert Gish, Man-Hung Siu, and Robin Rohlicek. Segregation of speakers for speech recognition and speaker identification. In *ICASSP*, pages 873–876, 1991.

[8] J.J Godfrey and al. SWITCHBOARD: telephone speech corpus for research and development. In *ICASSP*, 1992.

[9] S.E. Johnson and P.C. Woodland. Speaker clustering using direct maximisation of the MLLR-adapted likelihood. In *ICSLP*, 1998.

[10] Ivan Magrin-Chagnolleau and al. Detection of target speakers in audio databases. In *ICASSP*, 1999.

[11] D.A. Reynolds and al. Blind clustering of speech utterances based on speaker and language characteristics. In *ICSLP*, 1998.

[12] Aaron E. Rosenberg and al. Speaker detection in broadcast speech databases. In *ICSLP*, 1998.

[13] Matthew A. Siegler and al. Automatic segmentation, classification, and clustering of broadcast news audio. In *DARPA speech recognition workshop*, 1997.

[14] P.C. Woodland and al. The development of the 1996 HTK broadcast news transcription system. In *DARPA speech recognition workshop*, 1997.