

CAMPAGNE D'ÉVALUATION ETAPE 2012 : SEGMENTATION ET REGROUPEMENT EN LOCUTEURS

H. Khemiri^{1,2}

Marwa Thlithi^{1,2}

D. Petrovska-Delacrétaz¹

G. Chollet²

¹ Département Electronique et Physique

TELECOM SudParis, CNRS-SAMOVAR

² Département Traitement du Signal et des Images

TELECOM ParisTech, CNRS-LTCI

{khemiri, thlithi, chollet}@telecom-paristech.fr {dijana.petrovska}@it-sudparis.eu

Introduction

L'institut Mines Télécom a participé à la tâche de segmentation et regroupement en locuteurs (SRL). Le processus se déroule en trois étapes :

- identification audio
- détection d'activité vocale
- segmentation et regroupement en locuteurs

1 Architecture générale du système Proposé

Comme le montre la Figure 1, l'architecture générale du système proposé est basée sur les modules suivants :

- identification audio pour enlever les jingles et les segments de bruit,
- détection d'activité vocale (segmentation parole/non parole) pour ne garder que les segments parole,
- segmentation GLR-BIC,
- regroupement BIC,
- décodage Viterbi,
- regroupement CLR.

2 Identification audio basée sur ALISP

L'identification audio par le contenu consiste à retrouver des métadonnées (artiste, nom de l'album, nom de la chanson, nom de la publicité, nom de l'émission, etc.) à partir d'un extrait audio inconnu. Pour traiter ce problème, il existe deux grandes approches : le tatouage audio et l'extraction d'empreinte. Nous sommes intéressés par des méthodes basées sur l'extraction d'empreintes audio, qui sont plus appropriée pour notre tâche.

L'identification audio par extraction d'empreinte est composée de deux modules : un module d'extraction d'empreinte et un module de comparaison. La première étape dans un système d'identification audio basé sur l'extraction d'empreinte est la création d'une base d'empreintes à partir d'une base de références.

La base de références contient les documents audio (musique, publicités, jingles) que le système pourrait identifier. Dans la deuxième étape un extrait audio inconnu est identifié en comparant son empreinte avec celles de la base de références.

Notre système d'identification audio est basé sur l'approche ALISP (Automatic Language Independent Speech Processing). Cette approche a été développée initialement pour le codage de la parole à très bas débit [1] et exploitée avec succès pour d'autres tâches, telles que la reconnaissance du locuteur [2] et de la langue [3] ou encore l'identification audio [4].

Les outils ALISP ont comme avantage de fournir une segmentation automatique en unités pseudo-phonétiques, apprises à partir d'un corpus audio, et qui n'ont pas besoin de la transcription textuelle pendant la phase d'apprentissage. Rappelons que pour les systèmes de reconnaissance classiques (basées sur une segmentation en phonèmes), des bases de données, avec leur transcription phonétique sont indispensables pendant la phase d'apprentissage.

Comme l'origine des documents audio de la base de développement (TV/Radio Show) est identique à celle de la base d'évaluation, nous avons décidé d'identifier les segments audio similaires entre les deux bases.

Ces segments représentent généralement les jingles ou les segments de bruits. Notre système d'identification audio utilise les unités segmentales fournies par les outils ALISP pour identifier les segments audio récurrents. En effet, les transcriptions ALISP de la base de développement et d'apprentissage sont calculées en utilisant les modèles HMM fournis par les outils ALISP et stockés dans une base de référence, ensuite ces références sont comparées aux transcriptions de la base d'évaluation en utilisant la distance de Levenshtein [5]. La distance de Levenshtein mesure la similarité entre deux chaînes de caractères. Elle est égale au nombre minimal de caractères qu'il faut supprimer, insérer ou remplacer pour passer d'une chaîne à l'autre.

Comme expliqué dans [4], l'ensemble des unités ALISP est automatiquement acquis par la paramétrisation MFCC

(Mel Frequency Cepstral Coefficients), la décomposition temporelle [6] [7], la quantification vectorielle [8], et les modèles de Markov cachés.

Après l'acquisition des modèles ALISP, une base de référence pour chaque émission est construite à partir des transcriptions ALISP des segments audio présents dans la base d'apprentissage et de développement.

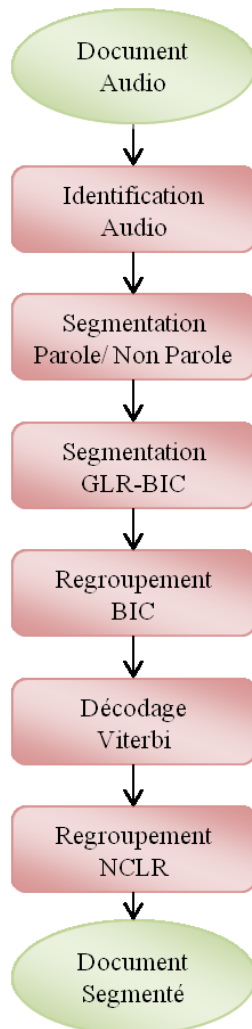


Figure 1 – Architecture générale du système proposé

Ensuite, le document radio de test est transformé en une séquence de symboles ALISP. Une fois les transcriptions ALISP des références et de données de test sont obtenues, nous pouvons passer à l'étape d'appariement.

A ce stade, la méthode de recherche utilisée dans notre système est très élémentaire. A chaque itération on avance par une unité ALISP dans le document audio de test et la distance de Levenshtein est calculée entre la transcription de référence et la transcription de l'extrait sélectionné dans le document audio. Au moment où la distance de Levenshtein est inférieure à un certain seuil, cela signifie que nous avons un chevauchement avec la référence. Puis nous continuons

la comparaison en avançant par un symbole ALISP jusqu'à ce que la distance de Levenshtein augmente par rapport à sa valeur à l'itération précédente. Ce point indique l'appariement optimal, où toute la référence a été détectée.

3 Segmentation parole/non parole

La segmentation du signal audio en segments de parole et de non parole est une étape indispensable dans la segmentation et le regroupement de locuteur. Un système de segmentation parole/non parole a été développé et appliqué sur la sortie du système d'identification audio.

3.1 Paramétrisation

La paramétrisation est fondée sur les coefficients MFCC et le taux de passage par zéro. Un vecteur spectral est calculé toutes les 10ms sur des fenêtres de 20ms. Chaque vecteur spectral est composé de 12 coefficients MFCC, les dérivées de premier et de second ordre et le taux de passage par zéro. Le vecteur résultant est de dimension 37.

3.2 Modélisation

La segmentation Parole/non parole repose sur la mise en compétition de 2 modèles constitués de mélanges de 64 gaussiennes (GMM, Gaussian Mixture Models). Pour des raisons pratiques, ces GMMs ont été codés sous la forme de modèles de Markov cachés à un état dont l'apprentissage a été réalisé à l'aide de la boîte à outils HTK [9].

3.3 Processus de décision

La mise en compétition directe trame par trame des deux modèles parole/non parole peut conduire à des segments de parole ou de non parole ayant une durée minimale de 10 ms. De ce fait une durée minimale d'une demi-seconde pour chaque segment reconnu a été imposée. En effet, comme le montre la Figure 2 chaque classe recherchée est modélisée comme la succession de 50 modèles HMM à un état.

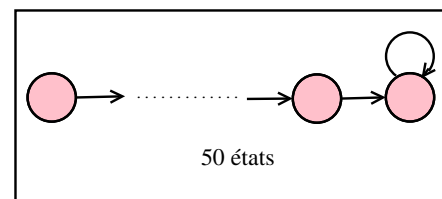


Figure 2 – La topologie de chaque modèle

4 Segmentation et Regroupement en Locuteur

Le module de segmentation et regroupement en locuteur est composé de la segmentation GLR-BIC (General Likelihood Ratio-Bayesian Information Criterion), le regroupement BIC, le décodage Viterbi et regroupement NCLR

(Normalized Cross Likelihood Ratio). Pour ce module le système *LIUM SPKDIARIZATION* a été utilisé [10]

4.1 Paramétrisation

12 MFCC et la log-énergie sont calculés dans une fenêtre glissante de 20 ms avec un décalage de 10 ms.

4.2 Segmentation GLR-BIC

La première étape consiste à segmenter le signal en segments homogènes (du même locuteur). Le principe repose sur le calcul d'une distance entre deux fenêtres adjacentes de taille 1.5s toutes les 0.16s. La distance utilisée est basée sur le critère GLR (Generalized Likelihood Ratio [11]) et obtenue par :

$$R(i) = \frac{N}{2} \log |\Sigma_{s_{i,i+1}}| - \frac{N_i}{2} \log |\Sigma_{s_i}| - \frac{N_{i+1}}{2} \log |\Sigma_{s_{i+1}}| \quad (1)$$

où s_i et s_{i+1} sont deux fenêtres adjacentes, $s_{i,i+1}$ est la fenêtre résultante de la concaténation de s_i et s_{i+1} . N , N_i et N_{i+1} sont le nombre de vecteurs acoustiques des fenêtres et Σ est la matrice de covariance.

A partir des scores GLR, les frontières des segments sont déterminées parmi les maxima locaux. Une frontière candidate est rejetée s'il existe un score supérieur dans les 1.5s suivantes du signal.

A partir de cette segmentation, Le critère BIC est appliqué pour regrouper les segments voisins qui semblent avoir été prononcés par le même locuteur. Ce critère est donné par :

$$\Delta BIC(i) = R(i) - \lambda P \quad (2)$$

avec

$$P = \frac{1}{2} \left(d + \frac{d(d+1)}{(2)} \right) \log N \quad (3)$$

où P est la complexité du modèle et λ un facteur de normalisation. Si $\Delta BIC(i) < 0$, les deux segments sont fusionnés.

4.3 Regroupement BIC

À la suite de la segmentation GLR-BIC, une classification hiérarchique est appliquée. Le critère BIC est appliqué pour regrouper les segments appartenant au même locuteur. Dans chaque itération les deux clusters les plus proches sont fusionnés jusqu'à ce que $\Delta BIC(i, j) < 0$.

4.4 Décodage de Viterbi

Un décodage par Viterbi est appliqué pour affiner les frontières des segments obtenus. Chaque cluster (locuteur) est modélisé par un GMM à 8 composantes à matrices diagonales appris sur les segments du cluster.

4.5 Regroupement NCLR

Dans les étapes précédentes de segmentation et de regroupement, les coefficients MFCCs n'étaient pas normalisés afin de préserver l'information sur l'effet du canal, ce qui contribue à différencier entre les locuteurs. À ce point,

chaque cluster contient la voix d'un seul locuteur, mais plusieurs clusters peuvent être liés à un même locuteur. Donc une étape de normalisation des coefficients MFCCs est effectuée avant une dernière étape de regroupement en utilisant *Feature Warping* [12]. Un modèle UBM (Universal Background Model) est appris à partir d'un ensemble d'apprentissage avec 512 composantes gaussiennes, ensuite ce modèle est adapté à chaque cluster pour obtenir le modèle de locuteur relatif à ce cluster. A chaque itération, les deux clusters qui maximisent le NCLR (Normalized Cross Likelihood Ratio) [13] sont fusionnés. La mesure NCLR est donné par :

$$NCLR(C_i, C_j) = \frac{1}{N_i} \log \frac{L(C_i|M_i)}{L(C_i|M_j)} + \frac{1}{N_j} \log \frac{L(C_j|M_j)}{L(C_j|M_i)} \quad (4)$$

avec M_i et M_j sont les modèles adaptés des clusters C_i et C_j et $L(.)$ est la mesure de vraisemblance. Le regroupement s'arrête lorsque la mesure du CLR dépasse un certain seuil.

5 Résultats

Les résultats sont résumés dans le tableau 1. La mesure utilisée dans cette tâche est la *Diarization Error Rate (DER)*. Nous avons calculé deux DER :

- DER1 : Diarization Error Rate pour le système proposé (résultats soumis).
- DER2 : Dizarization Error Rate sans le module de l'identification audio.

Fichier	DER1	DER2
BFMTV-BFMStory-175900	13.95	14.52
LCP-CaVousRegarde-235900	7.59	8.43
LCP-EntreLesLignes-192800-1	12.44	13.03
LCP-EntreLesLignes-192800-2	12.35	13.02
LCP-PilesEtFace-192800	10.13	10.93
LCP-TopQuestions-000400	28.66	34.78
LCP-TopQuestions-213800	1.02	6.35
TV8-LaPlaceDuVillage-201300	15.78	17.98
TV8-LaPlaceDuVillage-172800	14.95	15.79
EST2BC-FRE-FR-1000	9.91	12.45
EST2BC-FRE-FR-1750	14.33	14.89
EST2BC-FRE-FR-2152-1	16.01	16.76
EST2BC-FRE-FR-2152-2	16.3	17.09
EST2BC-FRE-FR-0910	4.82	7.39
EST2BC-FRE-FR-2004	9.52	10.69
Tous	12.52	14.43

Tableau 1 – DER pour chaque fichier audio de la base d'évaluation

On note que l'utilisation de l'identification audio basée sur ALISP a permis d'améliorer les résultats de 2%. Pour LCP-TopQuestions-000400 le fichier «uem» fournit lors de la campagne d'évaluation contient une partie qui n'est pas prise en compte par la vérité terrain ce qui explique le taux d'erreur élevé.

Lors du traitement des données de test, le temps moyen de calcul nécessaire pour traiter 60 secondes du signal audio est de 80 secondes avec une machine 3.00GHz Intel Core 2 Duo 4 Go de RAM. Notons que la recherche des segments récurrents avec le système d'identification audio est exhaustive, cette recherche coûte 30 secondes supplémentaires.

Conclusion et Perspectives

Cet article présente le système proposé par l'Institut Mines-Télécom (Télécom ParisTech & Télécom SudParis) pour la tâche de segmentation et regroupement en locuteurs dans la campagne d'évaluation ETAPE. Le système est composé des modules d'identification audio basée sur ALISP, segmentation parole/non parole basée des GMMs et segmentation et regroupement en locuteurs avec le système *LIUM SPKDIARIZATION*. avec ce système nous avons obtenu un DER global de 12.52% et nous avons montré que l'identification basée sur ALISP a amélioré les performances de système de 2%.

Les travaux futurs seront consacrés à l'intégration d'un nouveau algorithme de recherche à ce système pour accélérer l'identification audio. Cet algorithme est inspiré de BLAST (Basic Local Alignment Search Tools) [14] qui est souvent utilisé pour comparer des séquences biologiques

Références

- [1] Gérard Chollet, Jan Cernocký, Andrei Constantinescu, Sabine Deline, et Frederic Bimbot. *Towards ALISP : a proposal for Automatic Language Independent Speech Processing*, pages 375–387. 1999.
- [2] Asmaa ElHannani, Dijana Petrovska-Delacrétaz, Benoît Fauve, Aurélien Mayo, John Mason, Jean-François Bonastre, et Gérard Chollet. Text-independent speaker verification. Dans *Guide to Biometric Reference Systems and Performance Evaluation*, pages 167–211. 2009.
- [3] Gérard Chollet, Kevin McTait, et Dijana Petrovska-Delacrétaz. Data driven approaches to speech and language processing. Dans *Non-linear Speech Modeling and Applications*, volume 3445 de *Lecture Notes in Computer Science*, pages 164–198. 2005.
- [4] Houssemeddine Khemiri, Gérard Chollet, et Dijana Petrovska-Delacrétaz. Automatic detection of known advertisements in radio broadcast with data-driven alisp transcriptions. *Multimedia Tools and Applications*, pages 1–15, 2012.
- [5] Vladimir Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. Dans *Cybernetics and control theory*, pages 707–710, 1966.
- [6] Bishnu Atal. Efficient coding of lpc parameters by temporal decomposition. Dans *ICASSP*, pages 81–84, 1983.
- [7] Frederic Bimbot. An evaluation of temporal decomposition. Rapport technique, Acoustic Research Department AT&T Bell Labs, 1990.
- [8] Joseph Linde, Andres Buzo, et Robert M. Gray. An algorithm for vector quantizer design. *Communications, IEEE Transactions on*, 28(1) :84–95, Janvier 2003.
- [9] Steve Young, Dan Kershaw, Julian Odell, Dave Ollason, Valtcho Valtchev, et Phil Woodland. The htk book. Rapport technique, Entropics Cambridge Research Lab, 1996.
- [10] Sylvain Meignier et Teva Merlin. Lium spkdiarization : An open source toolkit for diarization. Dans *CMU SPUD Workshop*, 2009.
- [11] Herbert Gish, Man-Hung Siu, et Robin Rohlicek. Segregation of speakers for speech recognition and speaker identification. Dans *ICASSP*, pages 873–876, 1991.
- [12] Jason Pelecanos et Sridha Sridharan. Feature warping for robust speaker verification. Dans *ISCA Workshop on Speaker Recognition*, 2001.
- [13] Viet-Bac Le, Odile Mella, et Dominique Fohr. Speaker diarization using normalized cross-likelihood ratio. Dans *Interspeech*, 2007.
- [14] Stephen F. Altschul, Warren Gish, Webb Miller, Eugene W. Myers, et David J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215 :403–410, Mai 1990.