Segmentation parole/musique par Machines à Vecteurs de Support

Mathieu Ramona

RTL (Ediradio)
22 rue Bayard, 75008 Paris
mathieu.ramona@rtl.fr
TELECOM ParisTech / LTCI-CNRS

Gaël Richard

TELECOM ParisTech / LTCI-CNRS
37 rue Dareau, 75014 Paris
gael.richard@telecom-paristech.fr

ABSTRACT

We compare in this paper diverse hierarchical and multi-class approaches for the speech/music segmentation task, based on Support Vector Machines, combined with a median filter post-processing. We show the advantage of the multi-class approaches over the hierarchical schemes evaluated. Quantitative results provide a F-mesure over 96% that largely exceeds the results gathered by the ESTER evaluation campaign. We also show the relevance of the SVM with very low feature vector dimension on this task.
Keywords: Support Vector Machines, Audio segmentation, Speech detection, Hierarchical classification, ESTER

1. Introduction

La segmentation de données audio apparaît comme un besoin majeur dans la plupart des domaines de l'indexation audio. En effet, aussi bien les systèmes de reconnaissance de locuteurs, ou de transcription automatique, que les systèmes d'extraction d'information musicale (reconnaissance du genre...), ont besoin d'isoler, sur un signal inconnu, les zones pertinentes pour un type d'analyse donné. On trouve ainsi des applications dans le suivi automatique de la radio FM [10], pour le codage [2] ou encore pour la transcription de bulletins d'informations radiophoniques [11]. C'est dans ce dernier contexte que se place la campagne d'évaluation ESTER[1], sur laquelle nous avons basé l'évaluation des résultats de notre système.

La plupart des participants d'ESTER se basent sur des modèles de Markov cachés ergodiques avec modélisation des observations par mixtures de gaussiennes (GMM)[5]. On trouve également des applications simples des GMM sur des ensembles très variables de descripteurs audio [10] ; néanmoins les Machines à Vecteurs de Support (SVM), utilisées dans cet article, restent relativement peu exploitées pour cette problématique [6]. En effet, les SVM imposent une contrainte de discrimination entre deux classes, et diverses solutions ont été proposées dans la littérature pour étendre le domaine d'application des SVM aux problèmes multiclasses. Nous proposons ici une extension du système précédent [9] en comparant différentes approches hiérarchiques ou multiclasses pour la segmentation à partir de 4 classes d'apprentissage : voix, musique, voix sur musique (*mix* dans la suite), et chant. Nous évaluons en outre l'influence de la prise en compte du chant dans l'apprentissage.

L'architecture globale ainsi que les points particuliers du système sont présentés dans la section 2. Nous détaillons par la suite le protocole expérimental dans la section 3. Les résultats sont présentés et commentés dans la section 4 et nous exposons quelques discussions et perspectives globales dans la section 5.

2. Processus de classification

2.1. Architecture générale

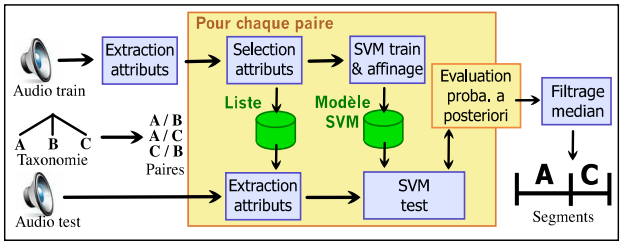

Fig. 1: Architecture du système proposé

Le système proposé combine des processus de discrimination par SVM en comparant des approches multiclasses à divers approches hiérarchiques. Nous suivons ici le principe traditionnel en apprentissage statistique, basé sur la classification de vecteurs de descripteurs acoustiques extraits sur un ensemble de trames chevauchantes couvrant le signal analysé. Chaque trame est associée à l'une des trois classes `voix`/`musique`/`mix`. La présence de chant sur certains des segments de musique du corpus peut introduire une confusion avec la classe `mix`, si tant est que le chant présent soit trop similaire à la parole. Nous avons donc introduit une quatrième classe (`chant`), utilisée ou non, selon la taxonomie adoptée, lors de la phase d'apprentissage et assimilée lors de la décision à la classe `musique`. La séquence des probabilités a posteriori est par la suite lissée par l'application d'un filtre médian, avant la prise de décision et le regroupement en segments homogènes.

2.2. Discrimination par paires

Taxonomies de classification La méthode de classification exploitée dans notre cadre est fondamenta-

[1] Évaluation des Systèmes de Transcription Enrichie d'Émissions Radiophoniques

lement discriminative. Différentes statégies (taxonomies), présentées en figure 2, sont retenues et comparées dans cet article pour étendre celle-ci à plus de 2 classes. Les arbres non-binaires exploitent le résultat des discriminations de toutes les paires des classes mises en jeu.

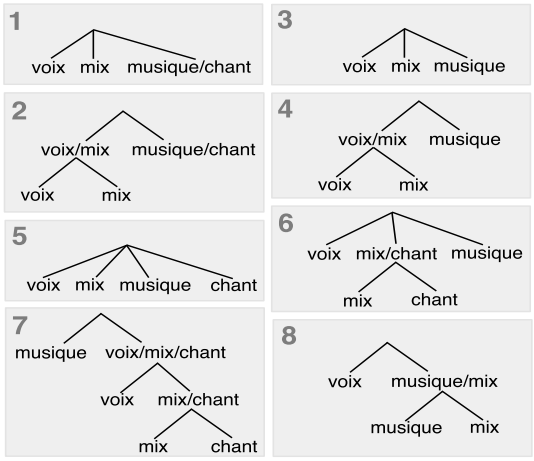

Fig. 2: Taxonomies retenues pour la classification

On constate ainsi qu'un grand nombre de paires sont communes à plusieurs arbres. On dénombre au total 14 paires (pouvant impliquer des unions de classes) ; pour chacune d'entre elles un discriminateur SVM est entrainé et affiné.

Descripteurs acoustiques Notre système exploite une large variété de plus de 600 descripteurs temporels, spectraux, cepstraux, perceptuels[2]. La majorité de ces descripteurs sont calculés sur des trames courtes de 32ms (avec un pas de 16ms), tandis que certains sont calculés sur des trames longues de 1s. Les descripteurs sur trames courtes sont remplacés par leur moyenne et variance sur chaque trame longue, permettant ainsi d'associer tous les descripteurs au sein d'un même vecteur. Chaque descripteur est centré et normalisé par sa variance évaluée sur l'ensemble d'apprentissage. Par la suite, pour chacune des paires, les descripteurs les plus pertinents sont sélectionnés à l'aide de l'algorithme IRMFSP, présenté dans [7]. Chacune des stratégies de classification a été évaluée avec un nombre de descripteurs commun à toutes les paires, variant de 4 à 70.

Discrimination L'approche de classification utilisée dans cette étude est basée sur les Machines à Vecteurs de Support (SVM). Les SVM appliquent une transformation non-linéaire (par l'application d'une fonction noyau $\boldsymbol{k}$, appelée *kernel*) sur les vecteurs de dimension d, dans un espace de dimension supérieure où les deux classes sont séparées linéairement sous contrainte de maximisation de la marge. Les noyaux exploités ici sont les *noyaux radiaux exponentiels* : $k(\boldsymbol{x}, \boldsymbol{y}) = exp(-\frac{\|\boldsymbol{x}-\boldsymbol{y}\|^2}{d\sigma^2})$.

On montre alors que la décision n'implique qu'un nombre limité de vecteurs de la base d'apprentissage ($\boldsymbol{x_i}$), proches de l'hyperplan de séparation, appelés *vecteurs supports*. La fonction de décision f pour un vecteur $\boldsymbol{x}$ a l'expression suivante :

$$f(\boldsymbol{x}) = \sum_i y_i \alpha_i \boldsymbol{k}(\boldsymbol{x_i}, \boldsymbol{x}) + b$$

Son résultat n'est malheureusement pas borné et ne représente donc pas une valeur probabiliste. La bijection par sigmoïde proposée dans [8] constitue une solution couramment utilisée pour obtenir une sortie probabiliste des SVM :

$$p(x) = \frac{1}{1 + \exp(Af(x) + B)}$$

Les constantes A et B sont évaluées à partir de la distribution des $f(\boldsymbol{x})$ sur l'ensemble d'apprentissage.

2.3. Probabilités a posteriori

On distingue deux schémas de base dans les arbres de classification présentés figure 2 :

Décision multi-classes : concerne les arbres 1, 3, 5 et 6. Dans ce cas chaque paire de classes est discriminée par la machine SVM associée. Nous utilisons l'algorithme proposé par Hastie et Tibshirani dans [4] permettant l'estimation des probabilités a posteriori de l'ensemble des classes à partir des résultats sur chaque paire.

Décision hiérarchique : concerne les arbres 2, 4, 6, 7 et 8. Les noeuds sont traités séquentiellement à la seule condition qu'un noeud père précède l'évaluation de ses noeuds fils. Chaque noeud discrimine l'une des classes en deux nouvelles classes. Les probabilités a posteriori des classes non concernées sont inchangées. Celles des deux nouvelles classes sont le résultat de la discrimination pondérée par la probabilité a posteriori de la classe évaluée. Ainsi la somme des probabilités sur toutes les classes demeure unitaire.

2.4. Lissage par filtrage médian

Afin de lisser les probabilités calculées sur la séquence des trames, on applique un filtrage médian (de longueur F_{med}) sur les probabilités a posteriori obtenues pour chacune des classes.

Par la suite, à chaque trame est attribuée la classe maximisant la probabilité a posteriori. Les trames adjacentes de même classe sont regroupées au sein de segments temporels.

3. Protocole expérimental

3.1. Corpora

Nous avons exploité durant notre expérimentation le corpus d'évaluation ESTER, dans le cadre de la tâche SES de segmentation des événements sonores. Nous avons réannoté le corpus d'apprentissage en différenciant les segments de musique seule et les segments chantés [3]. Les ressources sont réparties entre un ensemble d'apprentissage et un ensemble de développement de 12h30 respectant la distribution des diverses

[2] Voir la liste dans [9], auxquels il faut ajouter certains descripteurs liés à la fréquence fondamentale [1]

[3] les annotations actualisées sont consultables sur `http://www.telecom-paristech.fr/~ramona/dc/jep2008/`

radios du corpus. Les annotations du corpus de test n'ont pas été modifiés, afin de garder la pertinence de nos résultats au sein de la campagne d'évaluation.

Nous avons également testé l'efficacité des systèmes mis en place sur un corpus interne constitué de 12h d'une même journée (de 8h à 20h) diffusées par la radio RTL. Celui-ci est beaucoup plus diversifié que le corpus ESTER puisqu'il contient plusieurs émissions musicales ou de divertissement, dont le contenu diffère significativement des bulletins d'information fournis dans le corpus ESTER.

3.2. Affinage des paramètres

- **Dimension** d des vecteurs d'attributs. Chacun des classificateurs SVM a été entrainé pour des vecteurs de dimension d s'échelonnant entre 4 et 70.
- **Facteur de pénalité** C des SVM : Prend systématiquement la valeur C_{dat} proposée par Joachims dans l'implémentation de `SVMlight`[4].
- **Paramètre** σ des SVM : la valeur dans l'intervalle $[0.1; 1]$ (parcouru avec un pas de 0.1), minimisant le taux d'erreur moyen par trame sur l'ensemble de développement a été retenue.
- **Longueur du filtre médian** F_{med} : pour chaque arbre de classification et pour chaque dimension d, est retenue la valeur dans l'intervalle $[1; 35]$ maximisant la F-mesure sur l'ensemble de développement.

3.3. Évaluation

L'évaluation des résultats suit le protocole propre à la campagne d'évaluation ESTER. Les trois classes considérées dans la classification (`voix`, `mix` et `musique`) sont ramenées à deux classes (`voix` et `musique`) pouvant se chevaucher. Sur chacune de ces classes, et sur leur ensemble, sont évaluées les mesures de *Rappel (R)* et de *Précision (P)* définies comme le rapport de la durée cumulée où la classe est correctement détectée sur, respectivement, la durée cumulée où la classe est réellement présente et la durée cumulée où la classe est détectée. La *F-mesure (F)* est la moyenne harmonique de ces deux mesures (soit $F = \frac{2RP}{R+P}$). Sont également considérés les taux de fausse alarme (*fa*) et de faux rejet (*fr*).

Nous utilisons l'outil `trackeval` fourni dans le cadre de la campagne, pour effectuer les évaluations des différents critères.

4. Résultats

4.1. Corpus ESTER

La figure 3 montre l'évolution des F-mesures générales obtenues par chacun des 8 arbres de classification, en fonction de la dimension du vecteur de descripteurs.

On constate tout d'abord la nette distinction à haute dimension ($d \geq 20$), entre les arbres de classification 2, 4, 7 et 8 d'une part (en pointillés), et les arbres 1, 3, 5 et 6 d'autre part (en lignes entières). Ce constat montre le net avantage (un gain absolu de 1% environ sur la F-mesure) des approches multi-classes sur les

[4] `http://svmlight.joachims.org/`

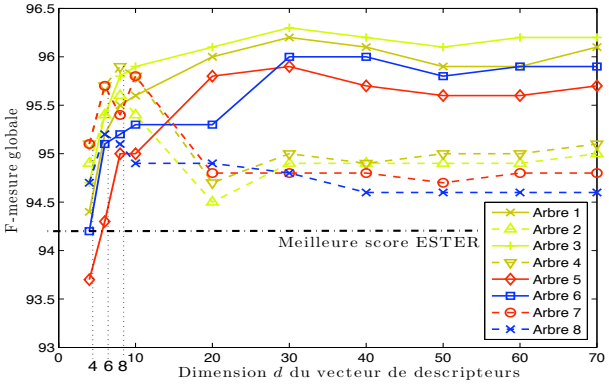

Fig. 3: Résultats de l'évaluation sur l'ensemble de test du corpus ESTER

approches hiérarchiques dans notre cadre. Nous précisons que la distinction entre ces deux groupes se retrouve sur l'ensemble de validation, ce qui aurait permis le choix d'une des approches multi-classes, sans connaissance de l'ensemble de test.

On remarque, dans le second ensemble, un léger avantage de l'arbre 3, qui exclut la classe `chant` de la phase d'apprentissage. Par contre l'arbre 5, qui introduit la classe `chant` au même niveau que les autres dans la décision multi-classes, semble être le moins avantageux. Ceci semble montrer la pertinence de l'exclusion des segments chantés pour l'apprentissage de la classe `musique`. Néanmoins, ce constat est à tempérer, en raison de la faible présence du chant dans l'ensemble de test.

Il est à noter également que l'augmentation de la dimension au delà de $d = 30$ ne semble pas apporter de gain significatif aux performances du système, quelle que soit l'arbre de classification choisi. De plus on note que les performances restent très acceptables même avec un vecteur de très faible dimension ($d = 4$) et sont supérieures à celles reportées dans le cadre de la campagne ESTER à partir d'une dimension $d = 6$. Le tableau 1 reproduit les résultats des trois meilleurs participants pour la tâche SES (à noter que la plupart se sont avant tout focalisés sur la détection de voix dans leur travail). On pourra consulter le tableau original dans l'article résumant les résultats de la campagne ESTER [3].

Le tableau 1 résume également les performances obtenues par les 4 meilleures approches, pour un vecteur de dimension $d = 30$. Notre système apporte un gain absolu d'environ 2% sur la F-mesure globale par rapport aux modèles proposés. On note une nette amélioration des performances par rapport à l'état de l'art, en particulier sur les segments de musique, avec un gain absolu de 25% environ. Nous précisons que le gain absolu apporté par le lissage par filtrage médian varie de 1 à 3% environ. Celui-ci compense dans une plus large mesure les erreurs observées à basse dimension, confirmant sa fonction de suppression des *outliers*.

Le tableau 2 représente la matrice de confusion des résultats trame à trame obtenus par l'arbre de clas-

Systèmes	général			voix			musique		
	F	%fa	%fr	F	%fa	%fr	F	%fa	%fr
Arbre 1	96.2	3.4	4.8	99.0	15.1	1.3	77.3	2.8	26.8
Arbre 3	96.3	3.4	4.5	99.1	17.7	1.0	77.7	2.7	26.8
Arbre 5	95.9	3.4	5.2	98.7	12.4	1.9	77.5	3.0	25.9
Arbre 6	96.0	4.1	4.6	99.0	13.0	1.5	76.8	3.6	24.7
ESTER 1	94.2	2.1	9.5	98.8	30.1	1.5	52.7	1.2	61.7
ESTER 2	93.1	1.3	12.1	98.9	9.7	1.9	33.7	1.0	78.5
ESTER 3	92.7	11.7	5.7	99.2	36.6	0.7	54.8	10.9	38.7

Tab. 1: Performances des meilleures approches comparées aux meilleurs résultats de la campagne ESTER (tâche SES)

sification 3, pour $d = 30$. On constate avant tout la confusion quasi-nulle entre les classes de voix et de musique. L'essentiel des erreurs demeure dans la détection erronée de voix seule pour des trames de mix. Ce problème est principalement dû à la présence dans le corpus ESTER de nombreuses zones de voix bruitée dans les segments de voix ; ainsi notre approche peine à distinguer le bruit de la musique de fond.

Classe	volume	mix	mus	voix
mix	13.7%	76.7	1.6	21.7
mus	6.4%	20.7	77.9	1.4
voix	79.9%	3.5	0	96.5

Tab. 2: Matrice de confusion pour l'arbre 3 à $d = 30$

4.2. Corpus RTL

L'application du meilleur système (arbre 3 pour $d = 30$) sur le corpus RTL produit une F-mesure globale de *90.7%* et de 74.6% et 98.6% respectivement pour les classes de musique et de voix. La matrice de confusion, table 3, nous permet de comprendre ce résultat. En effet on observe une aggravation de la tendance décrite précédemment à prendre des segments de mix pour des segments de voix seule (confusion de 62.4%). Cette confusion s'explique par la présence quasi permanente dans les émissions de divertissement d'un fond musical à peine audible. Ce défaut est très atténué par l'évaluation ESTER sur deux classes puisque la classe `voix` inclut les segments de voix et de mix, et ne tient donc pas compte de ce genre de confusions, qui pénalisent seulement la classe `musique`.

Classe	volume	mix	mus	voix
mix	4.9%	34.3	3.3	62.4
mus	1.2%	8.8	81.2	10.0
voix	93.9%	0.1	0	99.9

Tab. 3: Matrice de confusion sur le corpus RTL

5. Conclusion

Nous avons montré la pertinence de l'application des SVM pour la segmentation parole/musique, même à très faible dimension. En effet les résultats obtenus pour des vecteurs de dimension $d = 6$ sont supérieurs au meilleur résultat affiché dans le cadre de la campagne d'évaluation ESTER. Nous avons pu constater en outre la saturation du comportement des SVM au delà de $d = 30$.

Une comparaison de différentes approches de classification montre une prédominance des approches multi-classes sur les approches hiérarchiques, ainsi que l'effet bénéfique pour l'apprentissage de la suppression des zones de chant parmi les segments de musique. Nous avons enfin montré que l'essentiel des erreurs réside dans la confusion entre voix bruité et voix sur fond musical, résultat confirmé par notre évaluation sur le corpus RTL.

Références

[1] A. de Cheveigné and H. Kawahara. YIN, a fundamental frequency estimator for speech and music. *JASA*, 2002.

[2] K. El-Maleh, M. Klein, G. Petrucci, and P. Kabal. Speech/music discrimination for multimedia applications. In *Proc. ICASSP 2000*.

[3] S. Galliano, E. Geoffrois, D. Mostefa, K. Choukri, J-F. Bonastre, and G. Gravier. The ESTER Phase II Evaluation Campaign for the Rich Transcription of French Broadcast News. In *Proc. Interspeech 2005*.

[4] T. Hastie and R. Tibshirani. Classification by pairwise coupling. In *Adv. in Neural Information Proc. Systems*, 1998.

[5] D. Istrate, N. Scheffer, C. Fredouille, and J-F. Bonastre. Broadcast news speaker tracking for ESTER 2005 campaign. In *Proc. Interspeech 2005*.

[6] L.Lu, S.Z.Li, and H.J.Zhang. Content-based audio segmentation using support vector machines. In *Proc. ICME Multimedia and Expo 2001*.

[7] G. Peeters and X. Rodet. Hierarchical gaussian tree with inertia ratio maximization for the classification of large musical instrument database. In *Proc. DAFX 2003*.

[8] John C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large Margin Classifiers*. MIT Press, Cambridge, MA, 1999.

[9] G. Richard, M. Ramona, and S. Essid. Combined supervised and unsupervised approaches for automatic segmentation of radiophonic audio streams. In *Proc. ICASSP 2007*.

[10] J. Saunders. Real-time discrimination of broadcast speech music. In *Proc. ICASSP 1996*.

[11] G. Williams and D. P. W. Ellis. Speech/music discrimination based on posterior probability features. In *Proc. Eurospeech 1999*.