Deux niveaux et deux outils d'analyse pour une meilleure segmentation de données audio

Sébastien LEFÈVRE, Benjamin MAILLARD, Nicole VINCENT

Laboratoire d'Informatique
Université de Tours
64, Avenue Jean Portalis - 37200 Tours
lefevre@univ-tours.fr,vincent@univ-tours.fr

Résumé – Dans cet article, nous abordons le problème de la segmentation de données audio. Nous proposons un processus de segmentation à deux niveaux qui permet de diviser les pistes audio en courtes séquences qui sont étiquetées dans différentes classes. La segmentation est effectuée en calculant différentes caractéristiques pour chaque séquence audio. Ces caractéristiques sont calculées soit sur un segment audio complet, soit sur une trame (ensemble d'échantillons) qui est un sous-ensemble d'un segment audio. L'approche proposée pour la microsegmentation des données audio consiste en une combinaison d'un classifieur K-Means au niveau des segments et d'un système de chaînes de Markov cachées multidimensionnelles utilisant une décomposition du signal en trames. Une première classification est obtenue en utilisant le classifieur K-Means et les caractéristiques relatives aux segments. Le résultat final est alors fourni par l'utilisation des chaînes de Markov cachées multidimensionnelles et les caractéristiques relatives aux trames, en se basant sur les résultats intermédiaires fournis par la première étape. Les chaînes de Markov cachées multidimensionnelles sont une extension des chaînes de Markov cachées classiques qui permet la prise en compte de données multicomposantes. Elles sont particulièrement adaptées dans notre cas où chaque segment audio peut être représenté par plusieurs caractéristiques de différentes natures.

Abstract – We are dealing in this paper with audio segmentation. We propose a two level segmentation process that enables the audio tracks to be sampled in short sequences which are classified into several classes. The segmentation is performed by computing several features for each audio sequence. These features are computed either on a complete audio segment or on a frame (set of samples) which is a subset of the audio segment. The proposed approach for microsegmentation of audio data consists of a combination of a K-Means classifier at the segment level and of a Multidimensional Hidden Markov Model system using the frame decomposition of the signal. A first classification is obtained using the K-Means classifier and segment-based features. Then final result comes from the use of Multidimensional Hidden Markov Models and frame-based features involving intermediary results. Multidimensional Hidden Markov Models are an extension of classical Hidden Markov Model dedicated to multicomponents data. They are particularly adapted in our case where each audio segment can be characterized by several features of different natures.

1 Introduction

L'analyse et la classification de données audio est une tâche importante dans de nombreuses applications, telles que la reconnaissance de la parole ou l'indexation basée sur le contenu. Plusieurs outils mathématiques sont fréquemment utilisés dans ce domaine, comme par exemple les réseaux de neurones (RN) ou les chaînes de Markov cachées (CMC). La plupart des approches basées sur les CMC sont dédiées à la reconnaissance de la parole. Dans cet article, nous n'abordons pas ce problème mais plutôt celui de la segmentation de données audio en différentes classes définies a priori. L'application considérée ici est l'analyse de la piste audio d'une retransmission d'un événement sportif (un match de football). Trois éléments doivent être identifiés : le sifflet de l'arbitre, le bruit de la foule, et la voix du commentateur. La segmentation des extraits audio dans l'une de ces trois classes fournit suffisamment d'information pour permettre l'interprétation globale de l'événement sportif. En effet, une analyse plus contextuelle peut ensuite être effectuée pour chacun des extraits audio, selon la classe à laquelle il appartient.

La difficulté la plus importante pour résoudre un tel problème provient de la variabilité des caractéristiques au sein de chaque classe utilisée dans le processus de segmentation. Nous avons donc cherché une méthode de segmentation prenant en compte cette difficulté intrinsèque. Deux niveaux d'observation sont considérés, le niveau des segments et celui des trames. Dans une première étape, les segments complets sont regroupés afin de traiter des ensembles plus homogènes. Les ensembles obtenus ne sont que temporaires et n'ont pas de lien avec la classification finale que l'on souhaite obtenir. Leur but est de permettre d'adapter le processus à des caractéristiques partagées au sein de chaque ensemble. Il est alors possible, dans une seconde étape, d'utiliser un modèle Markovien original et défini pour chaque classe dans chaque ensemble.

Dans cet article, nous commencerons par rappeler les fondements des outils mathématiques utilisés dans notre méthode, c'est-à-dire le classifieur K-Means et les chaînes de Markov cachées. Nous décrirons ensuite la méthode proposée, en présentant les deux étapes qui la composent. Finalement, des résultats obtenus sur des pistes audio de retransmissions de matches de football seront présentés et commentés.

2 Traitement des données audio

La segmentation et la classification des données audio a été étudiée par de nombreux chercheurs. Ce traitement peut être vu comme un problème de reconnaissance de formes où deux décisions critiques doivent être prises : la sélection des caractéristiques audio et le choix du classifieur. Li et al [10] ont étudié un ensemble de 143 caractéristiques pour déterminer leur pouvoir discriminant. Wold et al [15] analysent et comparent différentes caractéristiques audio dans un but d'indexation audio basée sur le contenu. Li [11] compare expérimentalement différentes méthodes de classification et différents ensembles de caractéristiques. Bocchieri et Wilpon [3] discutent de l'influence du nombre de caractéristiques et de la nécessité d'une sélection des caractéristiques.

Nous devons déterminer si l'analyse des séquences audio sera effectuée selon une observation globale ou locale. Dans le premier cas, le but est d'étiqueter des séquences audio complètes, comme dans l'approche de Wang et al [14] pour étiqueter des pistes audio TV. Il est aussi possible d'étiqueter de courts segments audio (d'une durée inférieure à une seconde) afin de détecter des événements dans les séquences audio, comme dans les travaux de Kermit et Eide [7]. Cette détection peut même être effectuée en temps réel [16].

Plusieurs chercheurs ont proposé d'utiliser les modèles de Markov pour effectuer l'analyse de données audio. Ainsi, Kimber et Wilcox [8] construisent un modèle pour chaque orateur ou classe acoustique. L'apprentissage et la reconnaissance sont effectués respectivement en utilisant les algorithmes de Baum-Welch et Viterbi. Battle et Cano [1] proposent d'utiliser des modèles compétitifs au lieu des modèles classiques dans le cas d'un apprentissage non supervisé. Finalement, Hirsch [6] utilise une architecture de modèles adaptatifs pour traiter des signaux audio pour les télécommunications.

La méthode que nous proposons combine un classifieur K-Means et des modèles de Markov multidimensionnels. Nous rappelons ces outils dans la prochaine section.

3 Outils théoriques

3.1 Classifieur K-Means

Le classifieur K-Means est un outil largement utilisé en reconnaissance des formes [5]. En considérant le nombre K d'ensembles connu a priori, il permet de générer itérativement une partition de l'espace de données qui est définie par la position des centres des différents ensembles. Ces positions sont obtenues en minimisant les distorsions moyennes de tous les points appartenant aux K ensembles.

3.2 Modèles de Markov

Les chaînes de Markov cachées sont dédiées à la modélisation statistique de processus évoluant au cours du temps. Elles sont donc particulièrement adaptées à l'analyse de données audio [13]. Cependant, lorsqu'on considère des données d'observation multidimensionnelles, les CMC originales ne sont pas adaptées. Dans ce cas, il est possible d'utiliser une extension de ces modèles, les chaînes de Markov cachées multidimensionnelles à processus indépendants [4]. Dans cette section nous rappelons brièvement les principaux concepts des CMC avant de présenter les CMC multidimensionnelles.

3.2.1 Chaînes de Markov cachées

Une chaîne de Markov cachée est un ensemble de variables aléatoires représentant les états d'un processus stochastique discret composé d'une partie cachée et d'une partie observable. Il est caractérisé par un ensemble $\mathrm{S} = \{\mathrm{S}_1, \ldots, \mathrm{S}_\mathrm{N}\}$ d'états cachés de la CMC, un ensemble $\mathrm{V} = \{\mathrm{V}_1, \ldots, \mathrm{V}_\mathrm{M}\}$ de symboles pouvant être générés par la CMC, une matrice B de distribution de probabilités de génération des symboles dans chaque état, une matrice A de distribution de probabilités de transitions entre états, et un vecteur Π de distribution de probabilités de l'état initial. Une CMC peut donc être modélisée par le triplet $\lambda = \{\mathrm{A}, \mathrm{B}, \Pi\}$.

La méthode de segmentation proposée dans cet article utilise une architecture de CMC ergodique. Les étapes d'apprentissage et de reconnaissance sont respectivement basées sur les algorithmes de Baum-Welch [2] et Forward [13].

3.2.2 CMC Multidimensionnelles

Les CMC multidimensionnelles sont particulièrement adaptées lorsque l'on traite des données composées de plusieurs composantes indépendantes. En effet, à un instant t, la CMC ne générera pas seulement un seul symbole mais plutôt R symboles différents. Contrairement à l'utilisation de vecteurs d'observation, ici les composantes des vecteurs peuvent ici être de nature différente. L'observation décrit ici l'évolution de R processus au lieu de l'évolution d'un seul processus dans un espace de dimension R avec du bruit. On évite ainsi une étape de changement de représentation où les symboles construits font perdre de l'information.

L'architecture des CMC est donc modifiée. Le modèle ne contient toujours qu'une matrice A de transitions entre états mais R matrices B de génération des symboles, une pour chacun des processus $\mathrm{P_r}$ observables simultanément. L'architecture est caractérisée par le nombre R de processus liés à la CMC multidimensionnelle, l'ensemble $\mathrm{V^r} = \{\mathrm{V}^\mathrm{r}_1, \ldots, \mathrm{V}^\mathrm{r}_{\mathrm{M}^\mathrm{r}}\}$ de symboles liés au processus $\mathrm{P_r}$, la matrice $\mathrm{B_r}$ de distribution des probabilités de génération des symboles associés au processus $\mathrm{P_r}$, l'ensemble $\mathrm{V} = \{\mathrm{V}_1, \ldots, \mathrm{V}_\mathrm{R}\}$ des dictionnaires des symboles relatifs à chaque processus, et finalement l'ensemble $\mathrm{B} = \{\mathrm{B}_1, \ldots, \mathrm{B}_\mathrm{R}\}$ des matrices de distribution des probabilités de génération des symboles.

La méthode que nous proposons ici est basée sur une version multidimensionnelle [4] des algorithmes de Baum-Welch et Forward. Nous détaillons dans la prochaine section comment nous combinons ces modèles Markoviens multidimensionnels avec un classifieur K-Means.

4 Combinaison des deux outils

Afin d'effectuer un apprentissage et une reconnaissance des données analysées, nous devons choisir un ensemble de caractéristiques sur lesquelles baser nos conclusions. Nous pensons qu'il est important d'obtenir de l'information à différents niveaux d'observation, permettant ainsi une complémentarité des informations obtenues. Il est cependant difficile de gérer ces

différents niveaux au sein d'un seul modèle Markovien. Nous avons donc décidé de considérer deux niveaux d'observation dans la piste audio, et d'utiliser chacun des deux outils présentés dans la section précédente à un niveau particulier. La piste audio (notée PA) est divisée en segments (notés SA) qui sont eux-mêmes découpés en trames (notées TA). Les caractéristiques calculées concernent donc soit le niveau des segments, soit le niveau des trames. Au niveau des segments, le classifieur K-Means va nous permettre de regrouper les segments dans K ensembles ou classes virtuelles. Ces classes n'ont pas de lien réel avec les C étiquettes possibles que nous voulons associer à chaque segment mais permettent d'obtenir une variabilité intra-classe plus faible du signal vis-à-vis des caractéristiques utilisées. Il est ensuite possible d'utiliser un autre classifieur pour étiqueter, dans chaque classe virtuelle, les différents segments. Dans ce cadre nous utilisons les caractéristiques relatives aux trames avec un ensemble (de cardinalité C $\times$ K) de CMC. Le diagramme de l'approche proposée est donné dans la figure 1.

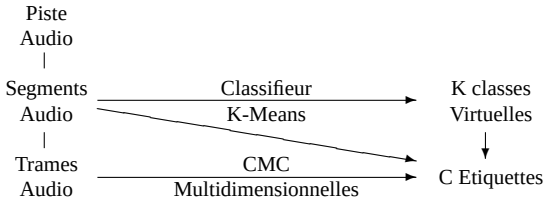

FIG. 1 – Diagramme de l'approche en deux étapes.

Nous allons maintenant préciser les caractéristiques audio que nous utilisons, ainsi que les phases d'apprentissage et de reconnaissance.

4.1 Caractéristiques audio

Nous effectuons une segmentation audio en utilisant un ensemble de 12 caractéristiques. Certaines d'entre elles sont tirées de [12] : non-silence ratio (NSR), volume standard deviation (VSTD), standard deviation of zero crossing rate (ZSTD), volume dynamic range (VDR), standard deviation of pitch period (PSTD), smooth pitch ratio (SPR), non-pitch ration (NPR), frequency centroid (FC), frequency bandwith (FB), 4 Hz modulation energy (4ME), energy ratio of subband 1-3 (ESRB1-3). Nous utilisons aussi une caractéristique relative au cepstre (CBF : cepstrum-based feature). Parmi ces caractéristiques, 5 peuvent être calculées au niveau du segment (NSR, SPR, NPR, VDR, 4ME) et seront utilisées dans la première étape de segmentation. Les autres caractéristiques (VSTD, ZSTD, PSTD, FC, FB, ESRB, CBF) peuvent être calculées au niveau des trames et seront analysées par les CMC multidimensionnelles.

4.2 Phase d'apprentissage

Nous effectuons l'apprentissage d'une manière supervisée en considérant 3 étapes successives : le calcul des caractéristiques, la classification en classes virtuelles par l'algorithme des K-Means, et la création des CMC multidimensionnelles.

Afin d'analyser un segment audio SA, nous devons au préalable calculer les caractéristiques qui lui sont associées. Dans une étape de prétraitement, nous découpons les segments audio en N trames chevauchantes et notées de TA_1 à TA_N.

Le classifieur K-Means peut alors être appliqué en utilisant seulement les caractéristiques basées sur les segments afin d'obtenir une première classification en K ensembles ou classes virtuelles (figure 2a). Le paramètre K doit être fixé a priori et a une influence sur le nombre de CMC multidimensionnelles à construire par la suite. L'utilisation du classifieur K-Means dans le processus de création des CMC permet d'améliorer la qualité de l'étiquetage.

Une fois que les segments audio ont été classés dans l'une des K classes virtuelles, ils vont être utilisés dans le processus d'apprentissage des CMC concernées. Plus précisément, nous n'utilisons que les données de l'un des K ensembles, et nous élaborons C CMC pour chacun de ces K ensembles. La figure 2b illustre ce principe où p segments audio $SA^j_{V_i}$ sont utilisés pour créer les CMC^i_1 à CMC^i_c liées à la classe virtuelle SA_{V_i}. L'algorithme d'apprentissage utilisé est celui de Baum-Welch dans sa version multidimensionnelle, introduit dans [4]. Il utilise les caractéristiques relatives aux trames qui permettent de décrire les segments par une observation dont la longueur est égale au nombre de trames. La méthode proposée nécessite K $\times$ C CMC multidimensionnelles afin d'effectuer la segmentation en C classes.

4.3 Phase de reconnaissance

Le but de la phase de reconnaissance est de classer chaque segment audio dans l'une des C classes. Les traitements nécessaires sont relativement similaires à ceux utilisés dans l'étape d'apprentissage. Les caractéristiques des segments audio sont tout d'abord calculées. Chaque segment est ensuite classé dans l'une des K classes virtuelles à l'aide du classifieur K-Means, comme le montre la figure 2a. Nous appliquons ensuite l'algorithme Forward (là encore dans sa version multidimensionnelle introduite dans [4]) sur les C CMC multidimensionnelles CMC^i_1 à CMC^i_c liées à la classe virtuelle sélectionnée SA_{V_i}. Le segment audio est finalement étiqueté comme appartenant à la classe pour laquelle l'algorithme Forward retourne le score le plus élevé P^i_{max} (figure 2c).

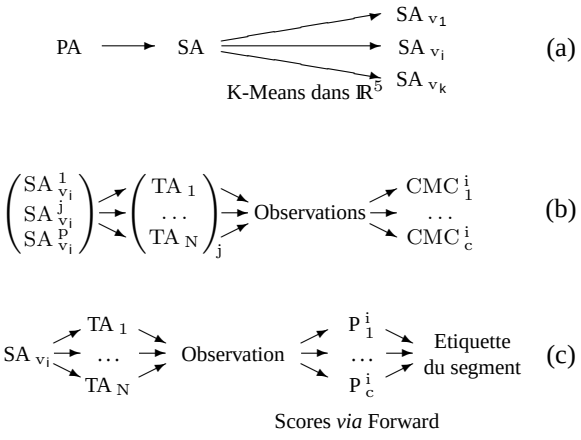

FIG. 2 – Description des étapes successives : (a) classification dans des classes virtuelles suivie par (b) l'apprentissage ou (c) la reconnaissance.

5 Résultats

La méthode proposée ici a été appliquée à des pistes audio de retransmission de matches de football [9]. Le but est de classer chaque segment audio en l'une des trois classes suivantes ($C = 3$) : sifflet de l'arbitre, bruit de la foule, et voix du commentateur. Une fois la classification d'un segment audio obtenue, il est alors possible d'analyser plus précisément celui-ci à l'aide d'une approche adaptée (par exemple la reconnaissance de la parole n'est effectuée que sur les segments de voix du commentateur).

La durée des segments audio analysés dans notre application est égale à 0,5 seconde. Les caractéristiques relatives aux trames sont calculées en divisant le segment audio en trames contenant 1024 échantillons. Deux trames successives sont décalées de 512 échantillons afin de conserver la propriété de continuité. Nous utilisons un classifieur K-Means prenant en compte 3 ensembles ($K = 3$). La méthode proposée nécessite donc 9 CMC multidimensionnelles.

Les tests ont été effectués sur un ensemble de 616 segments audio (21 pour le sifflet de l'arbitre, 148 pour le bruit de la foule, et 447 pour la voix du commentateur) extraits de différentes retransmissions. L'ensemble d'apprentissage a été défini de manière indépendante et contient 108 segments audio. Les taux de reconnaissance ont été calculés en terme de rappel et de précision et sont donnés dans le tableau 1. A titre de comparaison, les approches basées uniquement sur les CMC classiques donnent, pour un taux de précision similaire, un taux de rappel inférieur de 10 à 15 % [9].

TAB. 1 – Résults d'une segmentation en 3 classes.

Classe	Rappel	Précision
Sifflet	95 %	86 %
Foule	75 %	86 %
Commentateur	95 %	90 %

6 Conclusion

Dans cet article, nous avons proposé une méthode pour la microsegmentation de séquences audio. Nous combinons un classifieur K-Means et des chaînes de Markov cachées multidimensionnelles afin d'analyser des segments audio pour lesquels on calcule des caractéristiques à différents niveaux (segment ou trame). L'utilisation du classifieur K-Means nous permet de diminuer la variabilité intra-classe et donc d'améliorer la qualité des chaînes de Markov utilisées. Le caractère multidimensionnel du modèle Markovien utilisé permet quant à lui de traiter des données composées de plusieurs caractéristiques indépendantes. Les résultats ont montré l'intérêt de cette méthode par rapport à des approches classiques basées sur les CMC.

Parmi les perspectives envisagées, nous souhaitons évaluer d'autres caractéristiques audio et d'autres classifieurs (notamment des classifieurs non-supervisés) afin de confirmer l'amélioration des taux de reconnaissance. Afin d'obtenir un traitement temps-réel nécessaire dans l'application globale d'indexation d'événements sportifs ou dans d'autres applications à définir, la méthode doit également être implémentée sur une station multi-processeurs.

Références

[1] E. Battle and P. Cano. Automatic segmentation for music classification using competitive hidden markov models. In *International Symposium on Music Information Retrieval*, Plymouth, MA, Octobre 2000.

[2] L.E. Baum and J.A. Eagon. An inequality with applications to statistical estimation for probabilistic functions of markov processes and to a model for ecology. *Bull. American Society*, 73 :360–363, 1967.

[3] E.L. Bocchieri and J.G. Wilpon. Discriminative feature selection for speech recognition. *Computer Speech & Language*, 7(3) :229–246, Juillet 1993.

[4] T. Brouard. *Algorithmes Hybrides d'Apprentissage de Chaînes de Markov Cachées : Conception et Applications à la Reconnaissance de Formes*. Thèse de doctorat, University de Tours, France, Janvier 1999.

[5] K. Fukunaga. *Statistical Pattern Recognition*. Academic Press, 1990.

[6] H.G. Hirsch. Hmm adapatation for applications in telecommunication. *Speech Communication*, 34 :127–139, 2001.

[7] M. Kermit and A.J. Eide. Audio signal identification via pattern capture and template matching. *Pattern Recognition Letters*, 21 :269–275, 2000.

[8] D. Kimber and L. Wilcox. Acoustic segmentation for audio browsers. In *Interface Conference*, Sydney, Australia, Juillet 1996.

[9] S. Lefèvre, B. Maillard, and N. Vincent. 3 class segmentation for analysis of football audio sequences. In *IEEE International Conference on Digital Signal Processing*, volume 2, pages 975–978, Santorin, Grèce, Août 2002.

[10] D. Li, I.K. Sethi, N. Dimitrova, and T. McGee. Classification of general audio data for content-based retrieval. *Pattern Recognition Letters*, 22 :533–544, 2001.

[11] S.Z. Li. Content-based classification and retrieval of audio using the nearest feature line method. *IEEE Transactions on Speech and Audio Processing*, 8(5) :619–625, Septembre 2000.

[12] Z. Liu, Y. Wang, and T. Chen. Audio feature extraction and analysis for scene segmentation and classification. *Journal of VLSI Signal Processing for Signal, Image, and Video Technology*, 20(1/2) :61–79, Octobre 1998.

[13] L.R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2) :257–286, 1989.

[14] Y. Wang, Z. Liu, and J.C. Huang. Multimedia content analysis using both audio and visual clues. *IEEE Signal Processing Magazine*, 17(6) :12–36, Novembre 2000.

[15] E. Wold, T. Blum, D. Keislar, and J. Wheaton. Classification, search, and retrieval of audio. In *CRC Handbook of Multimedia Computing*. CRC Press, 1999.

[16] T. Zhang and C.C.J. Kuo. Heuristic approach for generic audio data segmentation and annotation. In *ACM International Conference on Multimedia*, volume 1, pages 67–76, Orlando, USA, Octobre 1999.