Automatic Segmentation for Music Classification using Competitive Hidden Markov Models

Abstract

Music information retrieval has become a major topic in the last few years and we can find a wide range of applications that use it. For this reason, audio databases start growing in size as more and more digital audio resources have become available. However, the usefulness of an audio database relies not only on its size but also on its organization and structure. Therefore, much effort must be spent in the labeling process whose complexity grows with database size and diversity.

In this paper we introduce a new audio classification tool and we use its properties to develop an automatic system to segment audio material in a fully unsupervised way. The audio segments obtained with this process are automatically labeled in a way that two segments with similar psychoacoustics properties get the same label. By doing so, the audio signal is automatically segmented into a sequence of abstract acoustic events. This is specially useful to classify huge multimedia databases where a human driven segmentation is not practicable. This automatic classification allow a fast indexing and retrieval of audio fragments. This audio segmentation is done using competitive hidden Markov models as the main classification engine and, thus, no previous classified or hand-labeled data is needed. This powerful classification tool also has a great flexibility and offers the possibility to customize the matching criterion as well as the average segment length according to the application needs.

The first stage in a classification system is the parameterization part where some features are obtained from the raw audio signal. The aim of this parameterization stage is to calculate a set of values that represent the main characteristics of the audio samples. The nature of the parameterization is strongly dependent on the application since it will determine the set of abstract acoustic units and, therefore, the underlying structure of a certain sound.
The main classification engine in our system is based on Hidden Markov Models (HMM). HMMs have proven to be a very powerful tool to statistically model a process that varies in time. The idea behind them is very simple. Consider a stochastic process from an unknown source and we only have access to its output in time. HMMs are well suited to model this kind of events. From this point of view, HMMs can be seen as a doubly embedded stochastic process with a process that is not observable (hidden process) and can only be observed through another stochastic process (observable process) that produces the time set of observations. However, after examining the audio segmentation problem, we can see that usual HMM training is completely useless because we do not have a priori information about the segmentation. In traditional HMM the training procedure uses a labeled observation sequence and parameter estimation is based on learning from examples. For example, in speech recognition, the training database is labeled with the phonemes of the speech sequences. By doing so, HMMs learn from examples of each phoneme and they use this knowledge to identify similar patterns in the recognition stage. But the problem we are dealing with in automatic music segmentation is very different (and by far more difficult). Our main goal is to blindly identify similar (in some sense) audio segments and therefore we do not have any previous knowledge to train traditional HMMs. Thus, this is clearly an example of HMMs unsupervised training. Competitive Hidden Markov Models have proved to be specially well suited for this kind of situations. CoHMM are different to HMM mainly in the training stage since the recognition stage (or classification) shares the common algorithm.
In this paper we also show some results that prove that coHMM converge to a realistic segmentation architecture.

The main architecture for all experimental systems used in this paper is the same, even though other parameterizations are under study. We use mel-cepstrum analysis to obtain the feature vectors. Each feature vector is calculated from a 25 ms frame with an overlap of 10 ms. In order to take into account some temporal structure beyond two frames, the feature vector is extended with the mel-cepstrum first and second derivatives. Each derivative is calculated using two frames before and two frames after each frame. Since mel-cepstrum coefficients are invariant to energy changes (coefficient 0 is dropped out) energy information is added to the feature vector with its first and second derivatives. Raw energy is not included to avoid dependence on the music level. The coHMM topology can be different from experiment to experiment since the classification criteria can be also different. However, the topology that better matches the average experiment is a 3 three states left-to-right model and one model for each event.

As it is shown with experimental results, this system opens the possibility to tackle new classification problems and, even though the experiments are related to audio, some other segmentation, clustering and classification problems can be solved.

Even though only audio experiments were presented, competitive Hidden Markov Models are a statistical tool that can be used in a wide range of applications, not only in the classification field but also in the recognition area because the segmentation labels can be used to identify parts of the audio stream.

Author Information

Eloi Batlle
Audiovisual Institute. Universitat Pompeu Fabra
Rambla 31, 08002 Barcelona
Catalunya - Spain
eloi@iua.upf.es
http://www.iua.upf.es

Pedro Cano
Audiovisual Institute. Universitat Pompeu Fabra
Rambla 31, 08002 Barcelona
Catalunya - Spain
pcano@iua.upf.es
http://www.iua.upf.es

Suggested Readings

Batlle, E. et al. 1998. Feature Decorrelation Methods in Speech Recognition. A Comparative Study. *International Conference on Spoken Language Processing*. Vol. 3, pp. 951-954.

Baum, L.E. et al. 1970. A Maximization Technique Occurring in The Statistical Analysis of Probabilistic Functions of Markov Chains. *The Annals of Mathematical Statistics*. Vol. 41, n. 1, pp. 164-171.

Dempster, A.P. et al. 1977. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society*. Vol. 39, n. 1, pp. 1-38.

Peeters, G. et al. 2000. Instrument Sound Description in the Context of MPEG-7. *International Computer Music Conference*.

Rabiner, L.R. 1989. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceedings of the IEEE*. Vol. 77, no. 2, pp. 257-286.

Raphael, C. 1999. Automatic Segmentation of Acoustic Musical Signals Using Hidden Markov Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. Vol. 21, no. 4.

Ruggero, M.A. 1992. The Mammalian Auditory Pathway: Neurophysiology. *chap. Physiology and Coding of Sound in the Auditory Nerve, Springer-Verlag*.

Serra, X. 1997. Musical Sound Modeling with Sinusoids plus Noise. *in G.D. Poli, A. Picialli, S.T. Pope and C. Roads, editors, Musical Signal Processing. Swets \& Zeitlinger Publishers*.

Torres, L. et al. 1990. Signal Processing V: Theories and Applications. *Elsevier Science Publishers B.V.*