TWO-STEP SEMI-SUPERVISED APPROACH FOR MUSIC STRUCTURAL CLASSIFICATION

Prateek Verma, Yang-Kai Lin, Li-Fan Yu

Stanford University

ABSTRACT

Structural segmentation involves finding hoogeneous sections appearing in a song. The name of these sections depends on the genre of interest. The task is particularly challenging as each of the misclassifed labels gives spurious change points. We have proposed a two step method by finding the boudaries present in the song followed by segment labelling. Novel idea of transforming the features into posteriori space using unsupervised model fitting is elucidated along with interesting behaviour of increasing the mixtures in GMM. Advanced techniques such as hidden markov model is also shown to give reasonable results for the task of boundaries detection. Finally the task of labelling the segments within the detected boundaries are carried out unlike the existing methods which give labels such as A, B, and C. The results are presented for boundary and segment level evaluation.

1. INTRODUCTION

Given a song the task of structual segmentaiton is to identify the structure appearing within a song. This is an interesting problem as this information is not given either in the CD metadata or can be derived from artist or song name. Usually the genre, style information can be extracted to a large extent by the artist name or meta data information.

Apart from applicability to music recommendation systems, this will help in automatic music summary generation too. For some of the music selling companies, they generally give a small manually generated preview of the song for a user to buy. However it is difficult to do this for millions of the song. Our system will help in summary or preview generation of the song by learning the structure present in the song along with the type of sections.

Due to the challenges involved and wide spread applicability, this task is part of MIREX contest for the past 6 years. Various approaches have been tried and a standardized dataset of beatles song is available for comparison of results of various researchers. The approaches used in the MIREX contest are mainly carried out using distance matrix-novelty score approach in order to get the boundary. Due to the difficulty of the task, most of the work has been confined to labelling the structure in the song as A, B, and C instead of the actual labels of the song.

In this work, we want to implement an approach to classify the actual labels of music sections.We have proposed a two step approach by first trying to find the boundaries present in the song. After finding the boundaries, the segment between the boundaries are classified. For boundary detection, we have taken advantage of unsupervised clustering methods such as GMM and HMM which can use a lower dimensional feature space to accurately find the structure in the song. Posteriori transformed features proposed are robustness to the noisy feature space. We have also achieved fairly decent accuracy in terms of segment level performance.

The contributions of our work is as follows: First, novel usage of transforming the features to posteriori probability features and application to segmentation task of western music. Second, behaviour of posterior features when a high number of mixtures are used to bring out repetitive segments. Third, labelling the segments into the corresponding labels instead of giving the labels like A, B, and C and achieving comparable results.

2. DATASET

The dataset chosen for this task was 174 songs released by Beatles. The annotations were downloaded from the MIREX website, and since due to copyright issues the audio files were not available, we manually downloaded all the songs from YouTube. The dataset consists of a total of 27,395 sec of audio files. The sections are labelled as intro, chorus, verse, outro, refrain, and others as six classes.

3. SYSTEM OVERVIEW

Fig. 1 shows the simple diagram of the proposed system. In our task, we must first know where the different music structures are, and then we can treat each section as individual event for our classification problem. Thus, our system is a two-step system including, first, the boundary detection (unsupervised) module and, second, the structural classification (supervised) module.

In detail, first EM algorithm is used to compute the parameters followed by posteriori features computation and boundary detection by novelty score computation.

Another approach to predict the boundary is by assuming that the temporal features are the outputs from a discrete

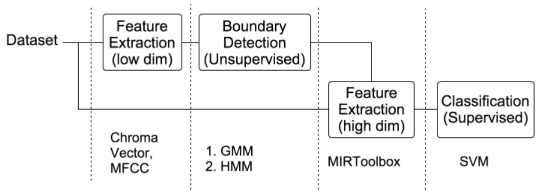
Dataset
Feature Extraction (low dim)
Boundary Detection (Unsupervised)
Feature Extraction (high dim)
Classification (Supervised)
Chroma Vector, MFCC
1. GMM
2. HMM
MIRToolbox
SVM

Fig. 1. System diagram of the proposed system.

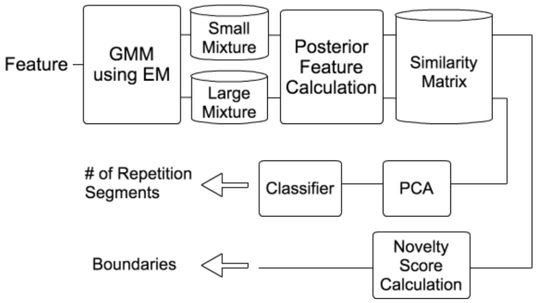

Fig. 2. Flow diagram of the boundary detection module using GMM.

Markov process, which forms a hidden Markov model. The parameters can be trained by the features, and the resulting state transition sequence can be used to determine the boundaries.

With the boundary detection results, we can move on to extract high-level feature for each sections segmented by the boundaries. Finally, with the section-level feature, the LIBLINEAR library and LIBSVM library [1] is employed for training svm models.

4. FEATURE

4.1. Low-Level Feature

For our task as it is two step approach, we describe two sets of features. For unsupervised clustering used in the first step, we do unsupervised maximum likelihood estimation on the features for clustering to identifying distinct changes occurring in the song. For this, the harmony based and timbral based features are sufficient enough to capture musically similar sections. The features we used for timbral description is the Mel-frequency cepstral coefficients. Given an audio waveform we obtain spectrogram by taking STFT.

The parameters used in our analysis are 0.01s hop with 0.03s frame for the computation of the above features. We then apply a texture window sampled at 0.1s with texture frame of 3s. This is mainly done to obtain a better homogeneous representation of the audio and remove the noisiness present in the features. Fig. 3 shows the decomposition of audio into the features desired by us which represent the harmony and the timbre of the sound.

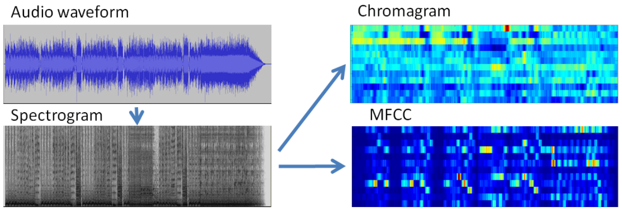

Fig. 3. Harmony (chroma) and Timbre (MFCC) features for unsupervised boundary detection.

However once we have generated the boundaries, for the task of labelling the segments, these features are not sufficient. The small set of features are helpful to cluster within the song but not across all the songs for labelling the segments. Hence we extract a very high number of features which can characterize and represent the segment labels.

4.2. High-Level Feature

For supervised classification used in the first step, we consider a number of baseline approaches for comparison. First, we use the MIRtoolbox (version 1.3.4) [2] to compute a 41 features covering the temporal, spectral, cepstral and harmonic aspects of music signals (denoted as TIMB). Second, the conventional MFCC, (delta)MFCC and (delta)(delta)MFCC are also used for their popularity (denoted as MFCC). Third, we augment MFCC and TIMB by calculating the mean, variance, and standard deviation over the local frames. Finally, we perform early fusion on MFCC and TIMB (i.e. by concatenating the corresponding clip-level representations to form a longer feature vector) to get our final feature. By doing so, we get a 317-dimensional high-level feature vector which represents various properties. Similar feature is used in [3] and considered to be very competitive for capturing every nuance of music signal. In our task, the structures of music is very complicated due to the variability from one song to another, hence we consider this high-level feature vector to be a good choice for our task.

5. FIRST STEP I: BOUNDARY DETECTION USING GMM

In literature, methods largely follow novelty score- self distance matrix (SDM) computation for a particular song [4]. On a set of features, a self distance matrix is computed whose (ith, jth) entry corresponds to the distance between the ith and the jth feature vector. For correct feature set block structure appears in the SDM corresponding to the homogeneous segments in the song. A kernel is convolved along the diagonal elements of the SDM to obtain the change points and the value obtained is called as novelty score.

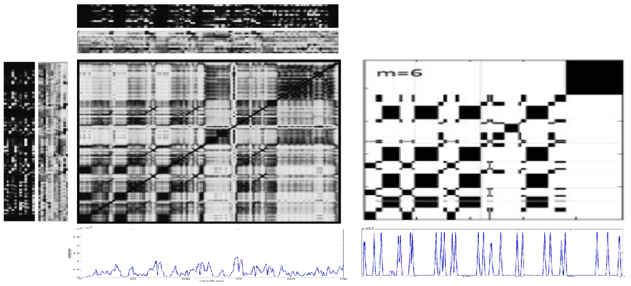

Fig. 4. Comparison of feature(left) and posteriori feature SDM(right) with the computed novelty score shown below.

It was seen that the computation of novelty score in the feature space was very noisy and we see a lot of spurious peaks in the novelty score. As shown in Fig. 4, this can be accounted for by the fact that some of the feature values are noisy which result in high value in the distance calculation in SDM. This motivated us to explore the idea of mapping the features into posteriori features as done previously in some speech recognition based applications. We present a methodology derived from speech processing applications to address this problem by mapping the feature vector space to a posteriori vector space. The difference between the approach in our case is that we do an unsupervised gaussian parameter estimation.

Let $\mathbf{X} = (x^{(1)}, x^{(2)}, \cdots, x^{(m)})$ be the features, where $x^{(i)} \in \mathbb{R}^n$ and $i = 1, 2, \cdots, m$. Assume that the distribution of $x^{(i)}$ is Gaussian mixture, i.e., $p(x^{(i)}) = \sum_{j=1}^{N} w_i f(x_i; \mu_j, \Sigma_j)$, where N is the number of mixture, μ_j, Σ_j are the mean and covariance matrix with respect to mixture j. We assume that Σ_j is a diagonal matrix for simplicity. $f(\cdot)$ is the probability density function (PDF) of the multivariate Gaussian distribution with dimension n. The EM-algorithm is then used to derive the parameters $\Theta = (\mathbf{w}, \mu, \mathbf{\Sigma})$ that maximizes the likelihood function $l(\Theta) = \prod_{i=1}^{m} p(x^{(i)}; \Theta)$, where $\mathbf{w} = [w_1 w_2 \cdots w_N]^T, \mu = [\mu_1 \mu_2 \cdots \mu_N]^T, \mathbf{\Sigma} = [\Sigma_1 \Sigma_2 \cdots \Sigma_N]^T$. After deriving Θ, it can be considered as mapping the features $\mathbf{X}$ to the posterior probability space $p_f = (p^{(1)}, p^{(2)}, \cdots, p^{(m)})$, where $p^{(i)} \in \mathbb{R}^N$. Furthermore, we have $p^{(i)} = [p(C_1 \mid x^{(1)}) p(C_1 \mid x^{(2)}) \cdots p(C_1 \mid x^{(N)})]^T$, where C_i are the i-th mixture of GMM.

Now after computation of posteriori feature SDM we observe that the novelty score is much better in terms of discrimination of section. It is seen that the posteriori features are more robust to noisy fluctuations. The outliers present in the feature space get mapped to uniformly low conditional class probabilities. In case of feature space these get mapped onto non-uniform distances depending on the feature value. Thus distances in posteriori probabilities vector space can be expected to be much more uniformly dependent on the underlying musical segments thus forming honogenous blocks in SDM.

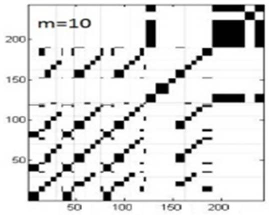

Fig. 5. Emergence of diagonals in the Posteriori SDM for high number of mixtures.

Diagonals were starting to emerge in the posteriori SDM even though they were not evident in the feature SDM when we increased the number of mixtures. As shown in Fig. 5, the appearance of diagonals in the SDM tells us about the repeating sections present in the song. For smaller mixtures these would appear as blocks if the number of mixtures chosen is larger than the number of sections present in the song. When we increase the number of mixtures, the block sizes appearing in the main diagonal reduces and will give much better temporal resolution. There will exist an optimum number of mixture as very few mixture will model two or more section together whereas higher mixtures will result in splitting up the section into one or more. Thus reducing the precision of the boundary detection task. The intra segment splitting for higher mixtures was validated as there was little change in the recall rates on increasing the number of mixtures. The number of mixtures were finally chosen to be 7 empirically giving the best performance as recall: 0.81, precision: 0.3132 and F-score 0.4706 for the boundary detection task with a tolerance of 3s. The performance reported by [5] is given as recall: 0.541, precision 0.56 and F-score 0.54 , which is only clustering of segments instead of classification in our case..

As we increase m smaller and smaller clusters starts forming. Now if we have the same smaller section present elsewhere in the song too, the cluster will model it perfectly giving low value of distance at that instance giving rise to a diagonal. Posteriori features also help us in improving the contrast of the repetitions. If the features are differing slightly, it will lead to them belonging to different clusters when the number of mixtures is large, and thus the distance between them will get amplified in the posteriori space as opposed to feature space.

Further the number of repetitions can be estimated accurately by looking at the SDM matrix. We treated this similar to MNIST challenge (Images- numbered label). We were able to predict with an error of 26% the number of repeated sections occurring in a song using a quadratic gaussian discriminative classifier. To do this, each of the SDM was converted into a matrix of 75×75. The dimension of such a large features were reduced using principal component analysis as the data was actually residing in a lower dimensional space. ($\sim$10 features accounted for $>$ 95% variance of the data)

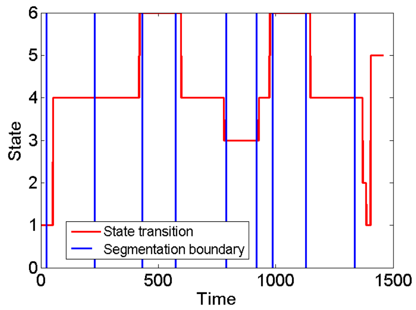

Fig. 6. Example for the state transition sequence with 6 states of a given song. The bule lines are the ground truth of the boundary location.

6. FIRST STEP II: BOUNDARY DETECTION USING HMM

6.1. Methodology

An alternative unsupervised machine learning approach we used in boundary detection is the Hidden Markov Model (HMM). HMM assumes that the system is a Markov process with unobserved states. Markov process assumes that the current state only depends on the previous state and independent of any other state, which approximates the music temporal characteristics well. The state is not directly observable, thus HMM parameters (transition probabilities and emission probabilities with respect to each state) can only be derived from the observable output, which depends on the underlying state. In the boundary detection task, we assume that the output is a single Gaussian distribution. The observable outputs are the feature of each time frame, and the state transition corresponds to the segmentation boundary. The state itself only presents the homogeneity within the segment. EM algorithm is used to determine the HMM parameters, and the Viterbi algorithm is then used to derive the state transition sequence. If the current state is the same as the previous state, the current and previous time frame are in the same segment; if state transition occurs, the boundary is predicted to occur between the time frames. An example for the state transition diagram is shown in Fig. 6.

6.2. Experiment Setup

For the HMM, only MFCC in the GMM 25-dimension features is used since it has higher homogeneity within each segment. The parameters through EM algorithm are derived by using the pmtk3 library. After deriving the boundary by the proposed algorithm, any boundary occurs after the previous one in less than 3 seconds is omitted in order to prevent overfitting. The time interval is chosen to be 3 due to the 3-second tolerance we use to evaluate the performance.

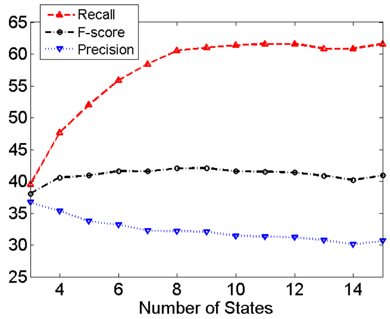

Fig. 7. The performance of the boundary detection with respect to the number of states. F-score reaches its maximum at 9 states.

	GMM		HMM		[5]
	317 raw	160 PCA	317 raw	160 PCA	
Linear	0.33	0.39	0.30	0.34	0.54
Rbf	0.40	**0.42**	0.36	0.35	

Table 1. The F-score of the classification result. Note that [6]'s task is only clustering of segments instead of classification in our case.

6.3. Experiment Result

As mentioned in the GMM part, for the boundary detection task, higher recall is preferable since the result is to be used in classification task, which is less sensitive to the overfitting. From Fig. 7, it can be seen that the number of states is the parameter actually affects the performance. By choosing larger number of states, recall will be higher. It is due to the fact that larger number of states implies that the number of segment with different characteristics are larger, which will result in larger number of predicted boundaries. The precision decreases with the increase of recall, which is a typical trade-off. F-score reaches its highest value when the number of states equals 9.

7. SECOND STEP: STRUCTURAL CLASSIFICATION USING SVM

7.1. Methodology

Support vector machine (SVM) is a well-known supervised machine learning algorithm for classification and regression analysis. In particular, SVM makes use of the kernel trick which implicitly maps the input feature to a high-dimensional feature spaces efficiently. In our case, the physical meanings behind different music structure are usually implicit in music signals, and the difference between different structures is difficult to annotate in various situation. Thus, we consider SVM to be a suitable algorithm for our task because it takes account of the implicit pattern lying behind the input feature, which may help us with this complicated task.

	Non	Man	Euc	Zsc
Linear	0.34	0.29	0.30	0.45
Rbf	0.29	0.42	0.45	**0.55**

Table 2. Results(F-score) of different normalizations.

	No PCA	122 (90%)	160 (95%)	197 (98%)
Linear	0.45	0.39	0.40	0.45
Rbf	0.55	0.62	**0.64**	0.55

Table 3. Result of different PCA components.

7.2. Experiment Setup

The features are extracted based on the ground truth annotations for training examples and boundary detection results for testing examples. The length of each feature is from 6s to 14s. To prevent overfitting, we adopt 5-fold cross-validation (CV) and leave-one-out cross-validation. For all the fold partitions in 5-fold CV, we make sure the distribution of sections over different structures is balanced. And Support Vector Machine (SVM) with linear or RBF kernel is used for classification. Note that we dont have to perform feature pooling because there is only one 317-dimensional feature vector for each section as described in section 4.

7.3. Experiment Result

To ensure we use our feature properly, we have to find a good normalization method on the feature for our task. This part is implemented off-line with all the features extracted based on the ground truth annotations. Table 1 shows the results in F-score. We can see that among z-score, Euclidean distance, Manhattan distance, and no normalization, z-score normalization gives us the best results, where the result can be improved from 0.29 to 0.55. This conforms to our expectation because in our high-level feature vector, every dimension represents different physical meaning and the unit of measurement also differs a lot, so some of them may be ignored because of the huge difference of magnitude. Thus, it is useful to apply z-score normalization so that each feature dimension has zero mean and unit variance. We then apply z-score normalization to all the remaining experiments.

We also tried the use of PCA in off-line experiment. The 317-dimensional feature we use is a large combination of different raw features, so it may be useful to apply PCA to refine it. Table.2 shows the results of using PCA with different number of components in terms of f-score. We can find that using 160 components gives us the best result improved from 0.55 to 0.64. Thus, we would also try PCA with 160 components in the remaining experiment.

Finally, we perform structural classification using the ground truth annotations for training and boundary detection results for testing. Shown in Table 3, our best result (0.42) comes from using GMM for boundary detection, PCA with 160 components, and SVM with RBF kernel, which are all the best parameters found in previous experiments.

8. CONCLUSION

We have successfully demonstrated a novel technique based on a two step appraoch by combining supervised and unsupervised algorithms for finding structure present in a song. Posterior feature mapping is an interesting idea which has been shown to be more robust to feature noise. Such an transformation can be applied to other machine learning tasks that deal with noisness in data. Further in order to increase precision of the results further, other change detection techniques such as Bayesian Information criteria can be combined to give better results. Hidden Markov model based technique which takes into account the transitional probabilities as well during the training was demonstrated to be successful in the task of boundary detection. Boundaries derived from Unsupervised posterior feature based approach was found to give better results as compared to HMM based approach both in terms of boundary F-score performance as well as segment level performance. In future this framework can be applied to PCA reduced features as well as after application of feature selection techniques. For the task of assignment of labels to a particular segment, sparse coding and deep learning based techniques can be tried in future.

9. REFERENCES

[1] C.-C. Chang and C.-J. Lin, *LIBSVM: A library for support vector machines*, 2001.

[2] O. Lartillot and P. Toiviainen, “A Matlab toolbox for musical feature extraction from audio,” in *DAFx*, 2007.

[3] L. Su L.-F. Yu and Y. H. Yang, “Sparse cepstral and phase codes for guitar playing technique classification,” in *Proc. IEEE. Int. Conf. Acoustics, Speech & Signal Processing*, 2014.

[4] Ewald Peiszer, Thomas Lidy, and Andreas Rauber, “Automatic audio segmentation: Segment boundary and structure detection in popular music,” *Proc. of LSAS*, 2008.

[5] Mark Levy and Mark Sandler, “Structural segmentation of musical audio by constrained clustering,” *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 16, no. 2, pp. 318–326, 2008.