DETECTION AND CLUSTERING OF MUSICAL AUDIO PARTS USING FISHER LINEAR SEMI-DISCRIMINANT ANALYSIS

Theodoros Giannakopoulos and Sergios Petridis

NCSR "Demokritos"
Inst. of Informatics and Telecommunications, Computational Intelligence Laboratory
Patriarchou Grigoriou and Neapoleos St., 15310, Aghia Paraskevi, Greece

ABSTRACT

We present a method aiming at facilitating musical audio summarization by organizing the signal into a set of possibly recurring parts, such that inclusion of an expert from each part would be adequate to compactly summarize the whole audio signal. Crucial to the success of the grouping segments into parts is the underlying distance metric, which depends on the feature space and should provide distances that are low for segments of the same audio part and high for segments of different audio parts. Starting with a general purpose audio feature space, we use the information from the sequential structure of audio signals, in order to estimate in a completely unsupervised way a Fischer subspace with discriminant characteristics for the particular audio signal. The derived feature space is used in a segmentation-clustering system based on fuzzy clustering, HMM and k-NN probability estimation. The experimental results show an almost 10% performance gain when adopting the Fisher subspace with respect to using the original feature space.

Index Terms— music summarisation, Fischer discriminant analysis, clustering, audio analysis

1. INTRODUCTION

Creating a concise audio summary, also referred as audio thumbnail, that best represents an original musical audio signal is a challenging research topic in music content analysis [1, 2]. Appealing audio summaries should allow listeners to get both a good feeling and a good idea of the original music. This focus of this paper is on the latter issue, namely on facilitating audio summarization by organizing the audio signal into a set of (possibly recurring) parts, such that inclusion of an expert from each part would be adequate to represent the whole audio signal.

Our objective is akin to but not identical to music structure analysis in music of sectional form, which comes down to recovering the sectional structure of a musical piece, such as intro, chorus and verse. Our proposed method does not make any particular assumptions about the existence of musical structure, neither does it assigns sectional labels to extracted parts. Instead, it results to a partitioning of the audio stream into contrasting clusters of (possibly not contiguous) audio segments. However, these clusters often correspond to musical sections, which fact has enabled us to base our experiments in corpora of musical pieces with sectional annotation.

Several audio features and pattern analysis methods have been adopted in order to achieve segmentation of music signals. As far as the audio features are concerned, Mel-frequency cepstral coefficients (MFCCs) [1] are a widely used type of feature. Even though most methods [2, 3, 4] adopt chroma-based features, which seem to fit naturally in the music domain, MPEG-7 audio descriptors have also been used in [5], focusing on audio spectrum envelope features. Regarding the core of the music segmentation task (i.e., the method which is responsible for discovering homogeneous segments), many methods are based on the computation of the *self-similarity matrix* [1, 2], where the similarity measure can be obtained in several ways, such as the Euclidean distance or the scalar product between the audio feature vectors. Self-similarity analysis is also adopted in [3], along with dynamic time warping, while a Hidden Markov Model (HMM) model has been used towards musical key estimation. In [5], a HMM with a large number of states is used in order to extract low-level labels, based on the adopted audio features for each frame and then histograms of neighbouring frames are clustered into segment-types.

In this paper, we present our research towards estimating a feature subspace which can be used to discriminate between musical parts. Instead of applying some clustering method on the initial feature space (whatever the selection of the audio features is), we propose leveraging information from the sequential structure of the audio signal, in order to find a Fisher linear discriminant subspace, where the discrimination between the different musical parts is more accurate. Section 2 discusses the way the initial feature space is obtained, while the way the sequential structure of the audio signal is taken into account to obtain the discriminant subspace is the subject of Section 3. Finally, Sections 4 and 5 present the experimental results and the conclusions respectively.

2. FEATURE EXTRACTION

Our methodology relies first on representing the audio signal as a sequence of N_x-dimensional feature vectors corresponding to overlapping fixed-size segments of the input signal. To derive these N_x-dimensional feature vectors, we have adopted a two-step methodology, similar to the one in [6] and [7].

At a first stage, a short-time analysis is conducted, resulting in ${}^{N_x}/_2$ audio features for every w_s of audio signal:

$$\left\{\mathbf{o}[n] \in \mathcal{R}^{N_x/2}\right\}, n \in [1 \dots T/w_s],$$

where T is the duration of the audio signal and T/w_s is the number of w_s-sized non-overlapping short-term windows. The values used in our experiments are $N_x = 62$ for the number of coefficients and $w_s = 50$ms for the analysis window and analysis step. The particular audio features extracted for each short-term frame are the following:

- Zero Crossing Rate (ZCR): this is the rate of sign-changes of a signal, i.e., the number of times the signal changes from positive to negative or back, per time unit.
- Entropy of energy [7]: this feature is a measure of abrupt changes in the energy level of an audio signal and it is extracted by computing the entropy of the normalized energy values of a particular number of sub-frames.
- Spectral centroid [6]: this is the centre of "gravity" of the spectrum.
- Spectral spread: this feature is extracted by taking the root-mean-square (RMS) deviation of the spectrum from its centroid (defined above).
- Spectral entropy [8]: this feature is computed by dividing the spectrum of the short-term frame into sub-bands (bins) and then computing the entropy of the individual spectral energies of the bins.
- Spectral flux [6]: this feature measures the local spectral change between successive frames.
- Spectral rolloff [6]: this feature is equal to the frequency below which certain percentage of the magnitude distribution of the spectrum is concentrated.
- 12 MFCCs (energy coefficient is ignored) [9]: MFCCs is actually a type of cepstral representation of the signal, where the frequency bands are computed using the Mel-scale.
- 12 chroma coefficients: this type of audio features (proposed by Wakefield in [10]) is a 12-element representation of the spectral energy of a signal, known as the *Chroma Vector*. Each element of the vector corresponds to one of the twelve traditional pitch classes (i.e., twelve notes) of the equal-tempered scale of the Western music.

The second stage of the feature extraction methodology is the *mid-term statistic* calculation. In particular, the means and variances over L subsequent vectors $\mathbf{o}[n]$ are extracted here, leading to N_x-dimensional vectors $\mathbf{x}[n]$. Means constitute the first half dimensions of the vectors, while variances the second half:

$$x_i[n] = \frac{1}{L}\sum_{m=n}^{n+L} \mathbf{o}[m], i = [0 \dots {}^{N_x}/_2),$$

$$x_i[n] = \frac{1}{L}\sum_{m=n}^{n+L} \left(o_{i-N_x/2}[m] - x_{i-N_x/2}[n]\right)^2, i = [{}^{N_x}/_2 \dots N_x]$$

Each $\mathbf{x}[n]$ describes a *texture window* of duration $w_l = L \cdot w_s$. Length of windows has been set to $L_1 = 50$ ($w_l = 1000$ms) for all tasks, except for FLSD (Fisher semi -discriminant Linear Analysis, which is discussed in the sequel), where a smaller value $L_2 = 20$ ($w_l = 400$ms) is used, to allow the feature vector of the the texture window to vary within a music thread (see Section 3 for the definition of music thread). These two lengths result respectfully to two sequences of features vectors, namely $\mathbf{x}_1[n]$ and $\mathbf{x}_2[n]$.

3. USING FISHER LINEAR SEMI-DISCRIMINANT ANALYSIS IN CLUSTERING OF MUSICAL SEGMENTS

3.1. Applying FLSD to find a discriminant subspace

As described in Section 2, the audio signal is represented by a sequence of N_x- dimensional feature vectors, mapped to some musical part (music cluster). This initial feature space can be considered as the sum of two orthogonal subspaces: a discriminative subspace and a classification-irrelevant subspace. In many cases the largest part of the original feature space does not contain discriminative information, which can lead to wrong estimation of the desired cluster, due to the fact that most clustering algorithms involve the Euclidean distance. In this paper we propose using a semi-supervised method to extract the discriminative subspace, called *Fisher Linear semi-discriminant Analysis* (FLSD). A complete description of this method, though, in the context of speaker diarization, can be found in [11].

The basic idea in Fisher linear discriminant analysis (FLD) is to extract linear combinations of features, where the means of classes are far from each other and the variance within each class is small. Let $\mathbf{x}$ be a N_x-dimensional feature vector, $\mathcal{C} = \{c_k\}$ be the set of classes, and $\{\mathbf{x}^i \mapsto c^i\}$ be a set

of mappings between feature vector samples to classes. The between-class scatter matrix is defined as:

$$\mathrm{S}_b = \mathop{\mathcal{E}}_{c\in\mathcal{C}}[(\mathbf{m}_c - \mathbf{m})(\mathbf{m}_c - \mathbf{m})^\top], \tag{1}$$

where $\mathbf{m} = \mathcal{E}_{\text{all } \mathbf{x}^i}[\mathbf{x}^i]$, $\mathbf{m}_c = \mathcal{E}_{\mathbf{x}^i \mapsto c}[\mathbf{x}^i], \forall c \in \mathcal{C}$, the average within-class scatter matrix is defined as

$$\mathrm{S}_w = \mathop{\mathcal{E}}_{c\in\mathcal{C}}\left[\mathop{\mathcal{E}}_{\mathbf{x}^i\mapsto c}[(\mathbf{x}^i - \mathbf{m}_c)(\mathbf{x}^i - \mathbf{m}_c)^\top]\right]. \tag{2}$$

and the total scatter matrix of samples as

$$\mathrm{S}_m = \mathop{\mathcal{E}}_{\text{all } \mathbf{x}^i}[(\mathbf{x}^i - \mathbf{m})(\mathbf{x}^i - \mathbf{m})^\top],$$

where $\mathcal{E}$ denotes the statistical average. Note that S_m does not depend on the class mappings, while one can easily verify that $\mathrm{S}_m = \mathrm{S}_b + \mathrm{S}_w$. Given a positive integer $N_y < N_x$, the aim of FLD is to find, among all possible $N_x \times N_y$ full rank matrices A, the matrix that optimizes a criterion of the following form:

$$\hat{\mathrm{A}} = \mathrm{argmax}_{\mathrm{A}\in\mathcal{R}^{N_x\times N_y}}\, r(\mathrm{A}, \mathrm{S}_1, \mathrm{S}_2). \tag{3}$$

where $(\mathrm{S}_1, \mathrm{S}_2)$ can be any of $\{(\mathrm{S}_b, \mathrm{S}_w), (\mathrm{S}_m, \mathrm{S}_w), (\mathrm{S}_b, \mathrm{S}_m)\}$. Several criteria of this form have been studied [12, Chapter 10], Sammon [13], such as $r = \frac{|\hat{\mathrm{A}}^\top \mathrm{S}_1 \hat{\mathrm{A}}|}{|\hat{\mathrm{A}}^\top \mathrm{S}_2 \hat{\mathrm{A}}|}$ As long as $\hat{\mathrm{A}}^\top$ is found, it can be used to project the original N_x-dimensional feature vectors to their N_y-dimensional FLD-optimal subspace

$$\mathbf{y} = \hat{\mathrm{A}}^\top \mathbf{x} \tag{4}$$

In order to use the FLD criterion one needs to know the class label of each sample from a training dataset. However, this type of information may be unavailable in a clustering framework. In the musical part clustering task, we do not know *all* the audio samples (i.e., signal segments) that belong to the same musical part beforehand, but one can guess that, for each sample, *all neighbouring samples, in a relatively small window, most likely belong to the same musical part*.

At this point, let us introduce the term of *class threads*. Each class can be composed out of one or more class threads, in the sense that all samples mapped to the same class thread v, are also mapped to the same class c. The surjective mapping of class threads to class, denoted by $h(v)$, provides, for each class thread, its corresponding class. Assuming that h is not known, while we do know the mapping of samples to class threads, we can estimate the average *within-class thread* S_w^h and *between-class thread* S_b^h scatter matrices and then apply the FLD criterion using these matrices. It has been shown in [11] that, under certain conditions, the subspace found using S_w^h and S_b^h in the optimization equation can well approximate the one that would had been found if the mapping with original classes were known. This optimization obtained using the within-class thead and between-class thread scatter matrices is defined as the optimal Fisher Linear Semi-Discriminant (FLSD) matrix. Note that S_m is used to refer to both mixed-class and mixed-class thread scatter matrices, which are equal, since there is no involvement of the class or the class-thread mapping in their definition.

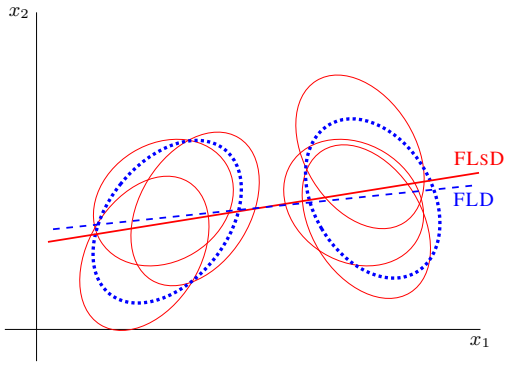

Fig. 1. A FLSD example in two dimensions with two classes and six class threads. Solid (resp. dashed) ellipses correspond to the contours of the variances of two class threads (resp. classes). The projection found by FLSD (solid line) closely approximates the one found by FLD (dashed line).

Figure 1 shows a toy example in a 2D space, with two classes composed by three class threads each. The FLD projection (dashed line in the figure) has been evaluated using the mapping to the original classes, $\{\mathbf{x}^i, c^i\}$. Evaluation of the optimal FLSD projection (solid line in the figure) uses the class threads instead, $\{\mathbf{x}^i, v^i\}$, neither the mapping to the original classes, nor h. Notice that, in this example, the FLSD projection is a close approximation to the FLD projection.

In [11] we have pointed out that we should expect a near-optimal behavior of the FLSD criterion. In particular, we can seek for an approximate discriminative subspace using class threads, instead of the classes, as long as the first-order statistics over the classes' threads do not differ much from the first-order statistics over their corresponding original ones.

Let us now describe the analytical steps of the algorithm that incrementally evaluates S_w^h through a long-term analysis of the audio signal. The algorithm proceeds by sequentially analysing fixed-size segments of duration w_l. For every music segment, a new music-part thread is created, and the feature vectors sampled within this segment are used to obtain the music-part thread mean feature vector and scatter matrix, also updating the overall within-class thread and mixed-class scatter matrices. Once all the audio signal has been processed, the scatter matrices are given as arguments to the Fisher criterion to obtain the optimal music-discriminative subspace.

3.2. Obtaining music clusters

Up to now, we have described the FLSD approach in order to find the music-discriminative subspace. In this section we

Algorithm 1: FLSD

Input: $\mathbf{x}_2[n], n \in 1 \dots ({}^{T}/_{w_s} - L_2)$ // used for the scatter matrices
Parameter: N_y // subspace dimension
Output: $\hat{\mathrm{A}}_{N_x \times N_y}$ // the optimal FLSD matrix
$n \leftarrow 1$ // initialise the analysis window sequence index
$v \leftarrow 1$ // initialise the class thread index
$\mathbf{m} \leftarrow \mathbf{0}_{N_x}, \mathrm{S}_m \leftarrow 0_{N_x \times N_x}, \mathrm{S}_w^h \leftarrow 0_{N_x \times N_x}$ // initialisation
while $n < \frac{T}{w_s} - L_1$ **do**
  $R \leftarrow [n \dots n + L_1 - L_2)$ // range of texture windows
  $\mathbf{m}_c \leftarrow \frac{1}{|R|} \sum_{k \in R} \mathbf{x}_2[k]$, // class thread mean
  $\mathrm{S}_c \leftarrow \frac{1}{|R|} \sum_{k \in R} \mathbf{x}_2[k]\mathbf{x}_2[k]'$ // class thread cov. mat.
  $\mathrm{S}_w^h \leftarrow \mathrm{S}_w^h + \frac{w_l}{T} \left(\mathrm{S}_c - \mathbf{m}_c \mathbf{m}_c^\top \right)$ // within-class
  $\mathbf{m} \leftarrow \mathbf{m} + \frac{w_l}{T} \mathbf{m}_c, \mathrm{S}_m \leftarrow \mathrm{S}_m + \frac{w_l}{T} \mathrm{S}_c$ // mixed-class
  $v \leftarrow v + 1$ // advance the musical part-thread index
  $n \leftarrow n + L_1$ // advance the analysis window
$\mathrm{S}_m \leftarrow \mathrm{S}_m - \mathbf{m}\mathbf{m}^\top$ // evaluate mixed-class scatter matrix
$\hat{\mathrm{A}} = \mathrm{argmax}_{\mathrm{A}_{N_x \times N_y}} r(\mathrm{A}, \mathrm{S}_m, \mathrm{S}_w^h)$ // apply the Fisher criterion

will show how the music clusters are obtained, after the optimal subspace has been computed. Overall, the following steps are applied:

1. For each non-overlapping window of duration w_l, a feature vector $\mathbf{x}_1[n]$ is generated (Section 2) and subsequently projected to the precomputed FLSD subspace (as discussed in Section 3.1) resulting to vector $\mathbf{y}[n] = \hat{\mathrm{A}}^\top \mathbf{x}_1[n]$.

2. The set of all projections $\mathbf{y}[n]$, independently to their order, are partitioned using a Fuzzy C-Means clustering algorithm [14]. Moreover, for each $\mathbf{y}[n]$, a cluster probability is estimated as the ratio of feature vectors attributed to the given cluster that are among the K closest to the given $\mathbf{y}[n]$. In our experiments, K has been set as the 10% of the sample set. In other words, the k-Nearest Neighbour classifier has been adopted here as a cluster probability estimator.

3. Using the previously estimated labels, the cluster transition matrix along with the prior probabilities of each cluster are evaluated. Together with the K-NN, these define an HMM model with states as many as clusters. Then, by applying the Viterbi algorithm, the most probable path is obtained.

4. Through HMM smoothing, some segments end up with having a label different from the one proposed by the clustering algorithm. It follows that the K-NN estimates of the conditional distributions are modified and hence Step 3 can be repeatedly applied to further improve the results. It has been experimentally found that this process converges within a few iterations.

5. Successive segments of the same cluster are merged, forming longer cluster-homogeneous segments.

It has to be noted that the adopted fuzzy clustering algorithm requires knowing the number of clusters beforehand. Since this information is typically not available, Steps 2 to 4 are applied for a range of number of clusters and the Silhouette Width criterion [15] is used to decide about the quality of the clustering result in each case and therefore the optimal number of clusters.

4. EXPERIMENTAL RESULTS

4.1. Dataset

As described in Section 1, our objective is not to recover the sectional structure of a musical piece, since we do not make any particular assumptions about the musical structure. Though, since an objective evaluation dataset is needed, we have adopted the Beatles annotation dataset [1], developed by the Center of Digital Music, Queen Mary. We have used the provided musical sections (e.g., intro, verse, etc) of this dataset as separate segments for our clustering task.

4.2. Performance measures

Since the purpose of the proposed algorithm is to detect clusters of musical parts, we have obviously used clustering-related performance measures. First, our approach was evaluated based on the overall accuracy rate, defined as the ratio of *correctly* clustered segments duration to the *total* signal duration. This measure is based on the optimal one-to-one mapping of the cluster labels with the true labels. This is achieved by applying the Hungarian method to the resulting confusion matrix between clusters and musical parts. In addition, we have used the Average Cluster Purity (ACP) and Average Musical Part Purity (AMP) measures, defined respectively as

$$\mathrm{ACP} = \frac{1}{N_a} \sum_{i=1}^{N_c} \max_{j=1 \dots N_s} n_{ij}$$

and

$$\mathrm{AMP} = \frac{1}{N_a} \sum_{j=1}^{N_s} \max_{i=1 \dots N_c} n_{ij}$$

where N_a is the total number of segments, N_s is the total number of musical parts, N_c is the total number of detected clusters and n_{ij} is the total number of segments classified in musical part (cluster) i and belonging to musical part j. We have also considered the Normalized Mutual Information Max (NMI_{max}) measure as suggested in [16], which is used to compare two partitions over the same data.

[1] http://isophonics.net/content/reference-annotations-beatles

Performance measure %	Feature space	
	Initial	FLSD
Accuracy	60	69
ACP	65	77
AMP	83	78
NMI_{max}	33	48

Table 1. Comparison of the clustering performance, with and without the FLSD subspace step.

4.3. Performance results

In Table 1 we present the performance measures of the proposed method when both feature spaces (original and FLSD) were adopted. The improvement when using the Fisher discriminant subspace is obvious: in terms of overall accuracy, the FLSD subspace leads to 15% relative increase, while in terms of the mutual information measure the relative increase is more than 40%. Finally, it has to be noted that the best performance was achieved for $N_y = 13$ dimensions of the FLSD subspace.

5. CONCLUSIONS

This study has shown that the estimation of a Fisher subspace in terms of music-part discrimination boosts the performance of the music clustering task. In particular, a 15% of relative increase in the overall accuracy is achieved with respect to the original feature space. It has to be noted that in this work we have taken into account mid-term audio feature statistics, generated over fix sized segments (texture windows). Therefore, no information regarding the overall musical structure of the audio signal, such as beat and rhythm information, is considered in this work. We plan to embed such music-related characteristics in the future, in order to further improve the performance of the method.

6. REFERENCES

[1] M Cooper and J. Foote, "Automatic music summarization via similarity analysis," in *Proc. of ICMIR*, 2002, pp. 81–85.

[2] M. Bartsch and G. Wakefield, "Audio thumbnailing of popular music using chroma-based representations," *IEEE Transactions on Multimedia*, vol. 7, no. 1, pp. 96–104, 2005.

[3] W. Chai, "Semantic segmentation and summarization of music," in *IEEE SP Magazine*, 2006, pp. 124–132.

[4] R Weiss and J.P. Bello, "Unsupervised discovery of temporal structure in music," *J. Sel. Topics Signal Processing*, vol. 5, no. 6, pp. 1240–1251, 2011.

[5] M. Levy and M. Sandler, "Structural segmentation of musical audio by constrained clustering," *IEEE TASLP*, vol. 16, no. 2, pp. 318–326, 2008.

[6] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE TASP*, vol. 10, pp. 293–302, 2002.

[7] T. Giannakopoulos, *Study and application of acoustic information for the detection of harmful content, and fusion with visual information*, Ph.D. thesis, University of Athens, 2009.

[8] H. Misra and et al., "Spectral entropy based feature for robust asr," in *ICASSP, Montreal, Canada*, 2004.

[9] S. Theodoridis and K. Koutroumbas, *Pattern Recognition, Fourth Edition*, Academic Press, 2009.

[10] G.H. Wakefield, "Mathematical representation of joint time-chroma distributions," in *Proceedings of the International Symposium on Optical Science, Engineering and Instrumentation (SPIE), Denver, Colorado*, 1999.

[11] T Giannakopoulos and S. Petridis, "Fisher linear semi-discriminant analysis for speaker diarization," *IEEE TASLP*, vol. 20, no. 7, pp. 1913 –1922, 2012.

[12] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, Academic Press Limited, Boston, MA, 1990.

[13] D.H. Foley and J.W. Sammon Jr, "An optimal set of discriminant vectors," *IEEE TC*, vol. 100, pp. 281–289, 1975.

[14] R. Babuka, PJ. Van der Veen, and U. Kaymak, "Improved covariance estimation for Gustafson-Kessel clustering," in *FUZZ-IEEE'02*. IEEE, 2002, vol. 2, pp. 1081–1085.

[15] L. Vendramin, R. Campello, and E.R. Hruschka, "On the comparison of relative clustering validity criteria," in *SIAM International Conference on Data Mining*, 2009, pp. 733–744.

[16] N.X. Vinh, J. Epps, and J. Bailey, "Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance," *JMLR*, pp. 2837–2854, 2010.