A HMM-EMBEDDED UNSUPERVISED LEARNING TO MUSICAL EVENT DETECTION

Sheng Gao and Yong-wei Zhu
Institute for Infocomm Research, 21 Heng Mui Keng Terrace, Singapore, 119613
{gaosheng, ywzhu}@i2r.a-star.edu.sg

ABSTRACT

In this paper, an HMM-embedded unsupervised learning approach is proposed to detect the music events by grouping the similar segments of the music signal. This approach can cluster the segments based on their similarity of the spectral as well as the temporal structures. This is not easily done for clustering with the traditional similarity measures. Together with a Bayesian information criterion, the proposed approach can obtain a suitable event set to regularize the complexity of the model structure. The natural product of the approach is a set of music events modeled by the HMMs. Our experimental analyses show that the detected musical events have more perceptual meaning and are more consistent than the KL-distance based clustering. The learned events match better with our experience in spectrogram reading. Its capacity is further evaluated on a task of music identification. The identification error rate is reduced to 1.57%, and 56.3% relative error rate reduction is observed comparing with the system trained using the KL-distance clustering method.

1. INTRODUCTION

Music is highly structured. However, this structure is not readily available for musical signal. Automatically recovering the structure and efficiently representing it is important for music information retrieval, indexing and organization of digital audio library, and music summarization [1, 6, 12]. There are rich structures in music. But here we are interesting in one of its temporal structures, i.e. repetitive pattern (or event). There are many works for automatically discovering the repeating patterns from acoustical signal [6, 8, 11, 12].

From the experience in spectrogram reading and speech analysis, these outstanding events (or landmarks) often occur at places with significant changes in spectral and temporal characteristics. It is therefore possible to find the musical structures using the data-driven approaches.

To detect the repeating patterns, the audio sequence is segmented and grouped together based on their similarities of the features. In [3, 6, 12], the feature sequence is mapped into its similarity representation by calculating the distance between any pair of frames, and then the segment boundary position is located using the heuristic methods. [2, 8] apply dynamic programming (DP) for segmentation with the pre-defined costs of insertion and deletion. These methods need to compute the similarity of all frame pairs and it is not efficient for real-time analysis of music excerpts that could last for a few minutes. To address this problem, the segment- based (fix-length e.g. [3] or variable-length based on the beat detection e.g. [11]) methods are applied. Then the unsupervised clustering methods, such as heuristic clustering [12] or *k*-means clustering [3], are used to merge similar segments to obtain a set of events.

In the proposed methods, the temporal structure of an excerpt of music is often ignored in the clustering. For example, in [2, 3, 8, 12], since the similarity is calculated from a pair of frames, the frame dependent information is not considered. For the segment-based methods [3, 11, 12], the similarity of a pair of segments is calculated from a global statistics of the segment and its temporal feature of a segment is excluded from calculating the similarity.

Considering the importance of the temporal information of the musical signal, it should be included in the similarity measure. In this paper, a hidden Markov model (HMM) embedded unsupervised learning approach is proposed to group the similar music segments together by measuring the similarity of their spectral and temporal structure. This learning algorithm is a modified K-means clustering method, where the similarity between the variable-length segments is measured by the HMM, not the Euclidean distance or KL-distance. To get a suitable set of clusters, a Bayesian information criterion (BIC) is used to regularize the complexity of the model structure. This learning algorithm can automatically generate a set of music event, each with a HMM for describing its temporal structure of the corresponding segments.

We will experimentally analyze its property of the proposed unsupervised learning approach for detecting music events and its effectiveness and efficiency for indexing music on a task of music identification. Since the learning algorithm is to group the music segments, we will first introduce the music segmentation algorithm in the next section.

2. MUSIC SEGMENTATION

Music segmentation is to partition the music signal into a sequence of segments. The similar segments are grouped into a cluster or event. Here the term, *event*, is used to refer to the cluster with the similar segments. In [1], it is defined by a set of coherent characteristics with some striking properties.

The beat and onset is the low-level perception feature that human perceives as well as it can be automatically detected from the signal. The onset position indicates a beginning of the next segment. So the onset-based segmentation should be more coherent and robust [10]. Here we will apply the *maximum a posteriori* (MAP) based adaptive learning approach for the beat and onset detection [10].

2.1 MAP-based Beat Detection

The beat of a piece of music is a sequence of equally spaced phenomenal impulses, which defines a tempo for the music [10]. Given a piece of music, a feature sequence can be extracted. Let $X = (\vec{o}_1, \cdots, \vec{o}_t, \cdots, \vec{o}_T)$ denote a sequence of *D*-dimensional feature vectors, and *T* be its length. A temporal window (or block) is applied to analyze the beat. Assume that its size is *L*, and there are *M* blocks in the feature sequence, then *X* can be re-denoted as $X = (O_1, \cdots, O_t, \cdots, O_M)$, where $O_t = (o_1^t, o_2^t, \cdots, o_L^t)$. If only the tempos in a range of $[\tau_a, \tau_b]$ are considered, then tempo induction can be formulated as,

$$\tau^* = \arg\max_{\tau \in \text{All possible tempo sequence}} P(\tau | X) \quad (1)$$

where $\tau = (\tau_1, \cdots, \tau_t, \cdots, \tau_M), \tau_t \in [\tau_a, \tau_b]$ is any possible tempo sequence, and $\tau^* = (\tau_1^*, \cdots, \tau_t^*, \cdots, \tau_M^*)$ the optimal one.

To simplify the optimization in Eq. (1), it is assumed that

0-7803-9332-5/05/$20.00 ©2005 IEEE

τ_t^* is estimated only from the block O_t but with a conditional probability, i.e. $P(\tau_t|\tau_{t-1})$, which is derived from its previous block. Then, Eq. (1) can be simplified as

$$\tau_t^* = \arg\max_{\tau_t \in [\tau_a, \tau_b]} (1-\eta)\cdot \log\left(P(O_t|A^t, \tau_t, \Sigma^t)\right) + \eta \cdot \log(P(\tau_t|\tau_{t-1})), \quad t \in [1, M]. \quad (2)$$

Here η is a constant weight. The first term in the right hand is the likelihood of the sub-sequence O_t. And the second is our model about the tempo for the block, O_t, given the known previous tempo.

The first term can be estimated from the observed data if a linear regression model is assumed. Given a block of sub-sequence evidence, O_t, the linear model is defined as,

$$\vec{o}_k^t = \mathrm{A}^t \cdot \vec{o}_{k-\tau_t}^t + \Theta^t \quad (3)$$

where $\Theta^t = (\theta_1^t, \theta_2^t, \cdots, \theta_D^t)^T$ is a prediction error vector. In this paper, A^t is a diagonal transformation matrix, and Θ^t is assumed to be a multivariate Gaussian distribution with a zero mean and a diagonal covariance, Σ^t.

So the probability distribution of $\vec{o}_k^t$ is also a Gaussian distribution (mean equal to $A^t \cdot \vec{o}_{k-\tau_t}^t$ and covariance, Σ^t), i.e.,

$$P(\vec{o}_k^t | A^t, \tau_t, \Sigma^t) \sim N(A^t \cdot \vec{o}_{k-\tau_t}^t, \Sigma^t) \quad (4)$$

Then the likelihood of the evidence, O_t, in Eq. (2) can be derived from Eq. (4) as,

$$\log\left(P(O_t|A^t, \tau_t, \Sigma^t)\right) = \sum_k \log\left(P(\vec{o}_k^t|A^t, \tau_t, \Sigma^t)\right) \quad (5)$$

Because the likelihood defined in Eq. (5) is a function of a tempo, the second term in Eq.(2) can be approximated by a logistic function as,

$$P(\tau_t|\tau_{t-1}) = \frac{1}{1+\exp\left(-\lambda \cdot \left(P(O_{t-1}|A^t, \tau_{t-1}, \Sigma^{t-1}) - \beta\right)\right)} \quad (6)$$

where λ is a scale coefficient and β is a bias. The normalization is performed to make $\sum_{\tau_t} P(\tau_t|\tau_{t-1}) = 1$.

With the above definitions, the optimal tempo, $\tau^* = (\tau_1^*, \cdots, \tau_t^*, \cdots, \tau_M^*)$, can be estimated using the EM algorithm based on the MAP criterion.

2.2 Beat Onset Decision

After the beat period is determined, the onset can be determined. Assume that the detected beat period is τ_t^* for a sub-sequence $O_t = (o_1^t, o_2^t, \cdots, o_L^t)$, and its corresponding energy envelope is $En_t = (en_1^t, en_2^t, \cdots, en_L^t)$, the sub-sequence is equally divided by its beat period. Let $O_t(i) = (o_{(i-1)\cdot\tau_t^*+1}^t, o_{(i-1)\cdot\tau_t^*+2}^t, \cdots, o_{(i-1)\cdot\tau_t^*+\tau_t^*}^t)$ be the feature in the *i-th* beat period (with $i \in [1, L/\tau_t^*]$), and $En_t(i) = (en_{(i-1)\cdot\tau_t^*+1}^t, en_{(i-1)\cdot\tau_t^*+2}^t, \cdots, en_{(i-1)\cdot\tau_t^*+\tau_t^*}^t)$.

The onset is defined as the time with the maximal energy. To extract the onset in each beat period, the averaging onset, $\overline{on}^t$, is first calculated from the averaging energy envelope as,

$$\overline{on}^t = \arg\max_{j \in [1, \tau_t^*]} \frac{1}{\tau_t^*} \sum_{i=1}^{L/\tau_t^*} en_{(i-1)\cdot\tau_t^*+j}^t \quad . \quad (7)$$

With the assumption that the onset in each beat period will have a bias (here maximum bias is set to 10% of the beat period) centered at the average onset, the real onset can be determined as,

$$on^t(i) = \arg\max_{j \in [\overline{on}^t - bias\tau_t^*, \overline{on}^t + bias\tau_t^*]} en_{(i-1)\cdot\tau_t^*+j}^t \quad (8)$$

3. HMM-EMBEDDED UNSUPERVISED LEARNING APPROACH

After a musical signal is segmented with the beat onset, its beat-level structure is obtained. This segmentation has some perceptual meaning, especially for percussion music. Due to the highly structured nature of music, many repetitive segments are often observed. To group these similar segments into some meaningful musical events without using any knowledge, the unsupervised clustering with a chosen similarity metric is applied.

Assume that the detected onset positions divide the feature sequence, $X = (\vec{o}_1, \cdots, \vec{o}_t, \cdots, \vec{o}_T)$, into *N* segments, which is denoted by $S = (s_1, \cdots, s_i, \cdots, s_N)$. s_i is the i-th segment with the length T_i, which is a subsequence having the feature sequence $X_i = (\vec{o}_1^i, \cdots, \vec{o}_{T_i}^i)$. Our task is to group these *N* variable-length segments into the musical event clusters based on their observation and the chosen similarity metric. Since the temporal variation of the music segment characterizes some specific properties of a piece of music (e.g. rhythm) and it contains rich information that is not represented by the frame-level feature, it is interesting to cluster the segments using a metric calculated from the spectral as well as the temporal features. The popular metrics (e.g. Euclidean or KL distance) will not do it well.

Although the dynamic time warping (DTW) can handle the similarity measure between the variable-length sequences, the HMM-based measure is more preferred since it can model and recover the hidden structures for the music segment [4, 9]. [7] studied the sequence clustering with the HMM. Here we will study the HMM-embedded K-means clustering approach for detecting music event.

3.1 HMM-embedded K-means Clustering

Given *N* music segments, $S = (s_1, \cdots, s_i, \cdots, s_N)$, each with a sequential observation having the variable frames, the K-means clustering can be used to get *C* clusters or music events [9] by minimizing an objective function defined as,

$$L(S|\Lambda) = \sum_{i=1}^N \min_j d(s_i, \Lambda_j), \quad (9)$$

where $\Lambda = (\Lambda_1, \cdots, \Lambda_C)$ is the parameter set for *C* clusters, and $d(s_i, \Lambda_j)$ measures the similarity between the *i-th* segment, s_i, and *j-th* cluster modeled by Λ_j. Since the length of the segments is variable and we would like to measure the similarity of the segments based on their statistical spectral distribution as well as the temporal structure. HMM is a good choice for this task. Its hidden states and state sequence will model the unobserved temporal structure of the music segment while the statistical spectral distribution of a segment is characterized by the state probabilistic distribution. In this case, the parameter set for Λ_j will include the transition matrix, A_j, among the states, the means, μ_j, and the variances, v_j, if the Gaussian distribution is used for the state description.

Since the segmentation is based on the beat onset and a segment describes the music changing property in an interval

between two continuous onsets, it is reasonable to use a HMM with the left-to-right state transition. So the similarity function, $d(\cdot)$, can be defined as the log-likelihood function derived from HMM as,

$$d(s_i, \Lambda_j) = -\log\left(P(s_i|\Lambda_j)\right), \qquad (10)$$

where $P(s_i|\Lambda_j)$ is the probability generated from the *j-th* cluster or HMM for the *i-th* segment.

To estimate the parameter set for C HMMs, i.e. $\Lambda = \{\Lambda_j = (A_j, \mu_j, v_j) | 1 \le j \le N\}$, the EM algorithm can be used. The HMM-embedded K-means clustering algorithm is shown in the following:

1. Initialization:
   - Assign N segments into C clusters randomly.
   - Estimate the initial parameters for each cluster (or HMM) based on the assigned segments.
2. Iterative estimation
   - Calculate the similarity defined in Eq. (10) between a given segment and any HMM using the Viterbi algorithm.
   - According to the above similarity, re-assign the segment into one of C clusters with the minimal similarity.
   - Re-estimate the parameters for C clusters based on the re-assignments using the segmental K-means algorithm for training HMM (Baum-Welsh algorithm can also be used. But computation cost is high) [4].
3. Terminate if the stop criterion defined in the following is reached. Otherwise, goto (2).
   - The preset maximal iteration cycle is reached. Or,
   - The relative improvement of the overall similarity is smaller than a predefined threshold.

With the above algorithm, we can cluster N segments or sequence with the variable length into C clusters based on the sequence similarity measured by the HMM. The statistical property of all segments in each cluster is characterized by a HMM, whose parameters are trained in the clustering procedure at the same time. The states of HMM will describe the hidden structure contained in an event. For example, if the correct beat and onset could be obtained for a piano music, the first state would describe the acoustical property for the attack while the last would describe for the decay. This is contrast to the K-means clustering using the Euclidean or KL distance as a measure, where only the overall statistics for a segment can be described.

3.2 Model Selection

In the above we have introduced the HMM-embedded clustering approach for the known cluster number. However, it is not true for many real-world applications. The fixed-size clusters will not work well for the music event detection and representation.

Given any piece of music, it is difficult to know about how many musical events are sufficient to describe it. Even for the experts, the definition is sometimes confused because of the hieratical organization of the music structure. However, significant differences of the musical structure are often observed. Some music excerpts are simple, which maybe played by a single instrument with the repeats of a few similar events, while others are complex where many instruments play simultaneously with the diverse chords and rhythms. This implies that the number of the musical events should partially depend on the complexity of music. For music with a simple structure, only a few clusters may be sufficient, while more clusters are needed for modeling music with the complex structure. So it is necessary to automatically determine the event number.

Many model selection criterions are proposed [9]. In this paper, we adopt a *Bayesian information criterion* (BIC) to choose an optimal model from a set of models, each of which has a different event number and is trained using the HMM-embedded K-means approach discussed in the above. To generate the set of models with the variable cluster number, the top-down clustering procedure is applied.

For any piece of music, we assume that at least *minC* clusters are needed to represent it and the maximal number of the clusters is set to *maxC*. The set of models, each corresponding to a cluster set, is $\Phi = \{\Lambda^n | n \in [\min C, \max C]\}$. A candidate model set with n clusters is denoted by $\Lambda^n = \{\Lambda_j^n = (A_j^n, \mu_j^n, v_j^n) | j \in [1, n]\}$, where Λ_j^n is the parameter set (i.e. transition matrix, means and variances) for the *j-th* cluster (or HMM) among n clusters. The set of model is trained using the top-down procedure together with the HMM-embedded K-means clustering. Then the BIC criterion is used to select the optimal model set, Λ^{n^*},

$$\Lambda^{n^*} = \min_{\Lambda^n \in \Phi} BIC(n) \qquad (11)$$

where *BIC(n)* is defined as,

$$BIC(n) = L(S|\Lambda^n) + \frac{1}{2}\kappa \cdot Q(n), \qquad (12)$$

The first term in the right hand side in Eq. (12) is the overall similarity between N segments and the model, Λ^n. It can be calculated using Eqs. (9-10). κ is a penalty weight. And *Q(n)* is a measure of the complexity of a model set. It is defined as,

$$Q(n) = \left(|A_j^n| + |\mu_j^n| + |v_j^n|\right) \cdot \log(N) \qquad (13)$$

For a HMM with R states, left-to-right state transition without any skip, and single Gaussian with the diagonal covariance for each state, $Q(n) = (2R + 2R \cdot D) \cdot \log(N)$ (D: feature dimension).

4. EXPERIMENTAL ANALYSIS

To analyze our proposed learning approach, a database with 807 pieces of music (average length is ~240 seconds), is first built. The diverse genres (e.g. western popular music, Chinese classical music, songs by various singers, etc) are covered. All pieces of music are converted to the standard wav format with a 16-bit resolution and 8-kHz sampling rate. Then the 36-dimensional feature vector is extracted, including 12-dimensional MFCC plus their first- and second –order difference [4].

In the beat and onset detection algorithm, η is set to 0.5, the interested tempo is between 60bpm and 250bpm, and the length of the block to analyze the tempo is 5 seconds. The desired maximal number of the clusters is set equal to 20 in the music event detection. The penalty weight in Eq. (12) is equal to 1.0e-4. Each HMM has 3 states with the left-to-right state transition and without any skip and single Gaussian is used for describing the state distribution.

4.1 Musical Event Detection Analysis

Now we will compare the HMM-embedded event detection approach with the KL-distance based method in [11], and then analyze its properties. Two excerpts of music are chosen from the dataset for illustration. One (*M1*) is a Chinese classical piece played by the flute and another (*M2*) is a piano music. The number of the event clusters is set equal to 10 for comparison. The onset position and the event labels for the segments are superimposed in its spectrogram, and are shown in Figure 1 and 2, respectively. Only a 5-sec piece is selected for display. The total length for each is ~240-sec. The labels (*"A","B",...*) are

only used to distinguish the different event clusters for each piece of music and there is not any other meaning.

These figures clearly show the effects on the clustering when the temporal structure and variation is embedded into the similarity measure. For *M1*, the first two segments (labeled as "*C*" and "*H*" in *M1-a*) have different notes as well as the temporal variation. The proposed HMM-embedded method can detect the difference while the KL-distance based method [11] assigns them into one same cluster (labeled as "*B*" in *M1-b*). For *M2*, the first two segments (both labeled as "*B*" in *M2-a*) have the same notes and very similar temporal structure. The proposed method detects it well while the KL-based method cannot.

The above analyses show that the HMM-embedded unsupervised clustering approach gives more reasonable results. The detected events are more consistent with our perception and experience of the spectral reading.

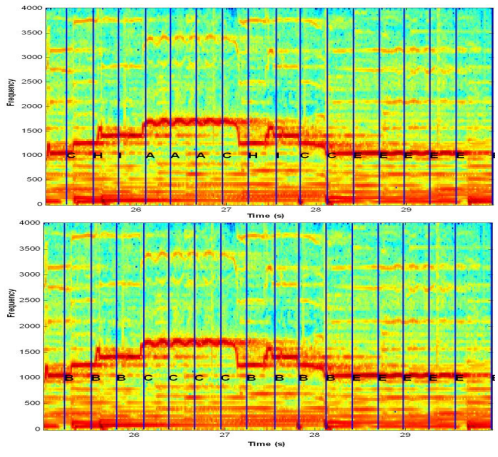

Figure 1 Musical events grouping (Chinese music, Vertical line: onset position, "A" ,"B", …,: events)

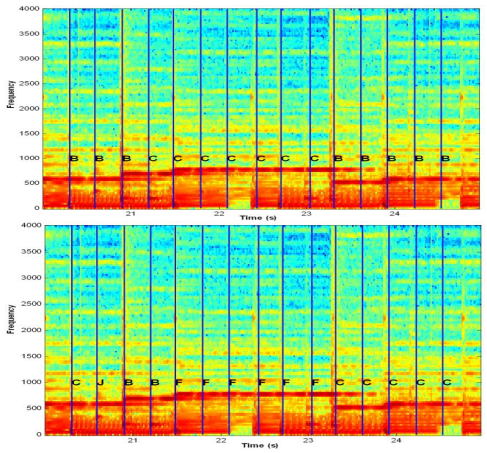

Figure 2 Musical events grouping (Piano music. Vertical line: onset position, "A" ,"B", …,: event)

4.2 Application for Music Identification

Now we will use the learned events for modeling the piece of music and apply them for the task of music identification. A test dataset with 2,421 query excerpts is built by randomly selecting three 20-second excerpts from 807 pieces of music in the database. The HMM-embedded method generates total 7,140 events for 807 pieces of music. So there are 7,140*3 single Gaussians. The baseline system is trained using the KL-distance as a similarity measure [11]. To make two systems have the approximate size of the model parameters, 27 music events for each piece of music are learned using the method introduced in [11]. Table 1 lists the error rate for the identification. Comparing with the baseline system, the system using the HMM-embedded clustering approach reduces the identification error rate from 3.59% to 1.57%. A relative error reduction, 56.3%, is seen.

Table 1 Comparison of the identification error rates

	KL-based	HMM-based	Rel (%)
Err (%)	3.59	1.57	56.3

5. CONCLUSION

We propose a HMM-embedded unsupervised learning approach to detect the music events by grouping the similar segments of the music signal. This approach can cluster the segments based on the similarity of the spectral as well as the temporal structures. This is not easily done for clustering with the traditional similarity measures such as the Euclidean or KL distances. Together with a Bayesian information criterion, a suitable event set is obtained to regularize the complexity of the model structure. The natural product of the proposed approach is a set of music events modeled by the HMMs. Our experimental analyses show that the detected musical events have more perceptual meaning and are more consistent than the KL-distance based clustering. The learned events match better with our experience in spectrogram reading. Its capacity is further evaluated on a task of music identification. The identification error rate is reduced to 1.57% and a relative error rate reduction, 56.3%, is observed comparing with the system trained using the KL-distance clustering. In the future, more applications will be studied such as music information retrieval, indexing and summarization.

6. REFERENCES

[1] E. D. Scheirer, "Bregman's chimerae: music perception as auditory scene analysis," *Prof. of ICMPC'96*.

[2] J. J. Aucouturier and M. Sandler, "Segmentation of musical signals using hidden Markov models," *Proc. of 110th AES Convention*, 2001.

[3] J. T. Foote and M.L. Cooper, "Media segmentation using self-similarity decomposition," *Proc. of SPIE Storage and Retrieval for Multimedia Databases*, 2003.

[4] L. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*, Englewood Cliffs, NJ: Prentice-Hall, 1993.

[5] M. Goto, "A chorus-selection detecting method for musical audio signals", *Proc. of ICASSP'03*.

[6] M. Wang, et al., "Repeating Pattern Discovery from Acoustic Musical Signals". *Proc. of ICME'04*.

[7] P. Smyth, "Clustering sequences with Hidden Markov Models", *Proc. of NIPS'97*.

[8] R. B. Dannenberg and N. Hu, "Pattern Discovery Techniques for Music Audio," *Proc. of ISMIR'02*.

[9] R. O. Duda, P. E. Hart and D. G. Stork, *Pattern Classification*, Second edition, Wiley, 2001.

[10] S. Gao and C.-H. Lee, "An adaptive learning approach to music tempo and beat analysis," *Proc. ICASSP 2004*.

[11] S.Gao, C.-H. Lee and Y.-W. Zhu, "An unsupervised learning approach to music event detection", *Proc. of ICME'04*.

[12] W. Chai and B. Vercoe, "Structural analysis of musical signals for indexing and thumbnailing," *Proc. of ACM/ IEEE Joint Conference on Digital Libraries*, 2003.