### music structural segmentation by hmm clustering

*Hua Xiao, Si Chen*

Center for Music Technology

Georgia Institute of Technology

#### Abstract

The abstract should appear at the top of the left-hand column of text, about 0.5 inch (12 mm) below the title area and no more than 3.125 inches (80 mm) in length. Leave a 0.5 inch (12 mm) space between the end of the abstract and the beginning of the main text. The abstract should contain about 100 to 150 words, and should be identical to the abstract text submitted electronically along with the paper cover sheet. All manuscripts must be in English, printed in black ink.

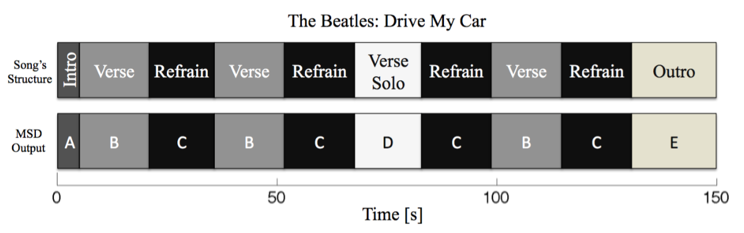
**Index Terms—**Audio, segmentation, structure, HMM, clustering

**1. Introduction**

These guidelines include complete descriptions of the fonts, spacing, and related information for producing your proceedings manuscripts. Please follow them and if you have any questions, direct them to Conference Management Services: Phone +1-979-846-6800 or email: [icassp2015@cmsworkshops.com](mailto:icassp2015@cmsworkshops.com).

**1.1. Problem definition**

Subheadings should appear in lower case (initial word capitalized) in boldface.  They should start at the left margin on a separate line.



**Figure 1**. Output example of the structural analysis of the song Drive My Car by the Beatles.

**1.2. Application**

Subheadings should appear in lower case (initial word capitalized) in boldface.  They should start at the left margin on a separate line.

**1.3. Paper organization**

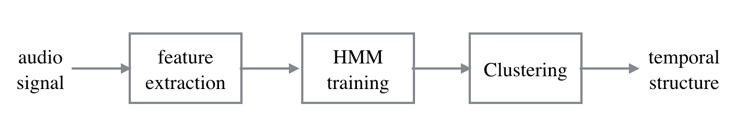
Algorithms for structural segmentation of music take an audio signal as input and give output information about its temporal structure. In our approach, a couple of subtasks can be identified. A rough overview of these subtasks is given in Figure 2.

**2. Related work**

All printed material, including text, illustrations, and charts, must be kept within a print area of 7 inches (178 mm) wide by 9 inches (229 mm) high. Do not write or print anything outside the print area. The top margin must be 1 inch (25 mm), except for the title page, and the left margin must be 0.75 inch (19 mm). All *text* must be in a two-column format. Columns are to be 3.39 inches (86 mm) wide, with a 0.24 inch (6 mm) space between them. Text must be fully justified.

**3. algorithm overview**

Algorithms for structural segmentation of music take an audio signal as input and give output information about its temporal structure. In our approach, a couple of subtasks can be identified. A rough overview of these subtasks is given in Figure 2. This section will be organized accordingly.



**Figure 2**. System overview: music structural segmentation by clustering based on HMM.

**3.1. Low-level feature extraction**

Our method of low-level feature extraction is based on the audio spectrum envelope, audio spectrum projection and sound model as described in the MPEG-7 standard [3]. This step yields a 21-dimensional audio spectrum projection feature vectors. The first 20 dimensions represent the spectral shape, by reducing dimension of the original normalized power spectrum envelope (64-dimension) using PCA method. And the final dimension is the relative power of each block window, which will be discussed in section 3.1.3.

*3.1.1. Beat tracking*

The dataset audios are mixed from stereo to mono by taking the mean of two channels. The original sampling rate is 44100Hz which is not necessary to be that high, however we downsampled to 11025Hz.

We use a hop size equal to the beat-length of the music (typically 400-600ms), and a window of three times the hop. While beat-lengths are estimated by a beat-tracking algorithm developed by labROSA [4]. For accuracy, we use the highest estimation of tempo instead of the lowest estimation (typically half of the highest tempo).

*3.1.2. Audio spectrum envelope*

The fundamental audio feature is the audio spectrum envelope. It is a standard MPEG-7 descriptor that uses logarithmic frequency scale and decibel scale to describe the power spectrum. Using logarithmic scale imitate the response of the human ear better.

We implement a constant Q transform (CQT) function to extract the audio spectrum envelope. The bin number of CQT is 8, which means there are 8 frequency bands in an octave. The lowest and highest frequency for CQT are 62.5Hz and 16kHz respectively. The output spectrum of the CQT is multiplied by its conjugate which yields a power spectrum.

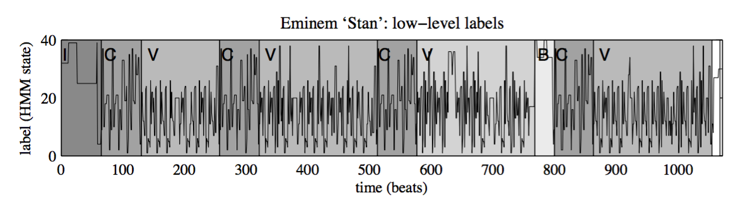
*3.1.3. Audio spectrum projection*

We applied the principal component analysis on the whole feature vector sequence to reduce the dimensionality of feature vector. Only the first 20 principal components are used. We also append the L2-norm, which represents the over all power, of each feature vector, to each feature vector. The norm is normalized by the max value, making all its values between [0, 1]. These 21-dimensional feature vectors, or audio spectrum projection (since PCA is a linear transform that project the feature vector to the principal components), are our low level features.

**3.2. HMM**

We use Baum–Welch algorithm to train an 80-state hidden Markov model on the entire sequence of feature vectors, with a single Gaussian output distribution for each state, and a single covariance matrix tied across all states. We then use Viterbi algorithm to decode the most likely state sequence, which is equivalent to assign a state to each beat.

Fig. 3 shows the resulting state sequence labels for the sample track.

**Figure 3**. Sequence of low-level labels for the sample track.

**3.3. Clustering**

We use a sliding window (16-beat long) to estimate the local state distribution at each beat of the sequence by counting 15 neighbouring states. This yields a 80-bin histogram of local states.

We then use kmeans with K=6 to cluster the histograms (state distributions). The value of K, 6, represents the number of common segmentation type of a song.

**4. Evaluation**

For the segmentation results to be evaluated and compared with a groundtruth, music pieces must be previously manually annotated. Example annotations can be found in the description of our databases

We use Beatle’s data set…

**4.1. Conditional Entropies**

One type of evaluation method that is being commonly used for the evaluation of music structure segmentation relies on information theoretic background and uses the notion of conditional. “These measures were proposed to provide a segmentation evaluation independent of the number of labels given by the estimated segmentation and allow for a better comparison of systems” [2].

The performance is shown in Table 1.

**4.2. Pair-wise F-measure**

The evaluation method of using the pairwise precision, recall and F-measure was introduced in [1]. The temporal segmentation step produces a set of boundaries between sections. These sections are evaluated by their comparison with the annotated groundtruth.

Considering Pm the set of similarly labelled frames in the reference annotation, and Ph the set of similarly labelled frames in the estimated structure, the pairwise precision, recall and F-measure are respectively defined as:

The performance is shown in Table 2.

**5. discussion**

Why our result is not good?

What may be the problem?

**6. conclusion**

Print your properly formatted text on high-quality, 8.5 x 11-inch white printer paper. A4 paper is also acceptable, but please leave the extra 0.5 inch (12 mm) empty at the BOTTOM of the page and follow the top and left margins as specified. If the last page of your paper is only partially filled, arrange the columns so that they are evenly balanced if possible, rather than having one long column.

**7. References**

[1] M. Levy and M. Sandler. “Structural Segmentation of Musical Audio by Constrained Clustering,” *IEEE Transactions on Audio, Speech & Language Processing*, vol. 16, no. 2, pages 318–326, 2008.

[2] F. Kaiser, “Music Structure Segmentation,” Diss. Universitätsbibliothek der Technischen Universität Berlin, 2012.

[3] M. Casey, “General sound classification and similarity in mpeg-7,”  *Organised Sound*, vol. 6, no. 2, p. 153..164, 2001.

[4] D. Ellis., “Beat Tracking by Dynamic Programming,” *J. New Music Research*, Special Issue on Beat and Tempo Extraction, vol. 36, no. 1, pp. 51-60, 2007.