Towards Music Structural Segmentation Across Genres: Features, Structural Hypotheses and Annotation Principles

TIAN, M; Sandler, MARKB

Towards Music Structural Segmentation Across Genres: Features, Structural Hypotheses and Annotation Principles

MI TIAN, Centre for Digital Music, Queen Mary University of London
MARK B. SANDLER, Centre for Digital Music, Queen Mary University of London

This paper faces the problem of how different audio features and segmentation methods work with different music genres. A new annotated corpus of Chinese traditional Jingju music is presented. We incorporate this dataset with two existing music datasets from the literature in an integrated retrieval system to evaluate existing features, structural hypotheses and segmentation algorithms outside a Western bias. A harmonic-percussive source separation technique is introduced to the feature extraction process and brings significant improvement to the segmentation. Results show that different features capture the structural patterns of different music genres in different ways. Novelty- or homogeneity-based segmentation algorithms and timbre features can surpass the investigated alternatives for the structure analysis of Jingju due to their lack of harmonic repetition patterns. Findings indicate that the design of audio features and segmentation paradigms and the associated signal processing techniques, the consideration of annotation principles and contextual information related to the music genre should be considered together in an effective segmentation system.

CCS Concepts: •**Information systems** → **Information retrieval; Evaluation of retrieval results; Presentation of retrieval results;** *Retrieval effectiveness;*

Additional Key Words and Phrases: Music information retrieval, music structural segmentation, data collection, harmonic-percussive source separation, evaluation, non-Western music

ACM Reference Format:
Mi Tian and Mark B. Sandler, 2015. Towards Music Structural Segmentation Across Genres: Features, Structural Hypotheses and Annotation Principles. *ACM Trans. Intell. Syst. Technol.* 9, 4, Article 39 (March 2010), 20 pages.
DOI: 0000001.0000001

1. INTRODUCTION

Music information retrieval (MIR) is a research field concerning the extraction of meaningful information from music content, usually using computational methods and with broad applications [Schedl et al. 2014]. Music is primarily an event-based phenomenon comprising a series of musical elements such as melody, harmony or rhythm that unfold in time. Both human listening and analysis activities suggest music boundaries to facilitate portraying content with specific within-piece sectional characteristics. *Music structural segmentation* (MSS) deals with the structural analysis of an entire piece. It involves dividing a music signal into its structural parts by giving it boundaries. What we target with an MSS task then depends on the subjective understanding we have of what defines the music structure. Smith studied several segmentation algorithms and suggested that algorithms designed originally for the structural

This work is supported by China Scholarship Council (CSC) and EPSRC project (EP/L019981/1) Fusing Semantic and Audio Technologies for Intelligent Music Production and Consumption (FAST-IMPACt). Sandler acknowledges the support of the Royal Society as a recipient of a Wolfson Research Merit Award.
Author's addresses: M. Tian and M. B. Sandler, Centre for Digital Music, School of Electronic Engineering and Computer Science, Queen Mary University of London, E1 4NS, London, U.K.

DOI: 0000001.0000001

analysis of Western popular music are widely applicable [Smith 2010]. Nonetheless, the corpora used in [Smith 2010] were collected on the basis of general structural coherence that is Western-centric. One primary motivation of this article is to set out principles for the analysis of the structure of non-Western music.

In the recent decade, a few non-Western traditional music corpora have been included in MIR research [Serra 2011]. Jingju, also known as Beijing Opera or Peking Opera, is one of the most representative Chinese traditional music genres. It combines singing, dance and theatre art and can offer intriguing research topics to challenge the existing MIR paradigm [Repetto et al. 2014]. Despite its rich heritage and the sheer size of its audience, little analytic work has been done to understand its music content from an MIR perspective until very recently with few works addressing its music structure.

Because Jingju is initially improvised at its birth, an analytical discovery of its structure will assist its standardisation and popularisation as well as applications in areas such as music production and education. In modern times, the art form of Jingju has undergone changes with newly introduced popular and regional characteristics. This paper only investigates the classical repertoire of traditional songs. It should also be noted that song structure has to be differentiated from the structure of the full Jingju play, where the former relates to only the arias part of the latter. In this work, we will be addressing the analysis of the song structure.

Existing MSS tools, including audio features and segmentation algorithms, are to a certain extent encoded with assumptions or heuristics observed for specific types of music they were originally designed for or evaluated with. The overarching goal of this article is to evaluate such tools and methodologies for different music genres, and we situate our study in the scenario of Jingju. With the research presented here, we are aiming towards an audio-based music structural segmentation system, which is capable of selecting the optimal audio features and segmentation algorithms allied to contextual information such as the genre and annotation principles, hence making intelligent structural discoveries. The remainder of this article can be summarised as below. In Section 2 we will outline the background of the studied music genre and review some related work. Section 3 surveys the existing datasets used in this paper and presents a new Jingju database. We present the feature extraction work with harmonic-percussive source separation processing in Section 4. The segmentation methods and experiment conditions will be introduced in Section 5 and Section 6 presents and discusses the results. Finally, Section 7 summarises this work and outlines directions for future work.

2. BACKGROUND: JINGJU MUSIC AND MUSIC SEGMENTATION RESEARCH

2.1. Jingju Music

The music form of Jingju can have distinctively different characteristics from Western music. Unlike the well-known Western pop, harmony or chordal structure is hardly present in Jingju songs at the segment-level, i.e., the music is what musicians call through-composed. The music texture is then *heterophonic*, where variations introduced by different instruments exist in the unitary basic melody.

The Jingju music system is comprised of three major elements: *melodic-phrases* (”qiang”)[1], *metrical patterns* (”banshi”), and *modes (”diaoshi”) and modal systems (”shengqiang xitong”)*. They are hierarchically related and collectively shape the music structure [Stock 1999; Wichmann 1991].

[1]A *melodic-phrase* in this scenario differs from the Western understanding for a *melodic phrase*, in the sense that it means ”the melodic progression for singing a single written character from the lyrics”.

When composing a Jingju play, modal systems and modes are firstly chosen to set the overall atmosphere. The metrical patterns and melodic-phrases are then arranged to elaborate the specific content of each passage of lyrics. The song lyrics are organised in a *couplet* structure, each consists of two lines of lyrics, which lays the basis of the music structural framework.

The melodic lines corresponding to a couplet are considered the smallest meaningful musical units. Although following certain melodic, rhythmic and instrumentation regularities, each pair of melodic couplets unfolds in a temporal order and never repeats. A passage of melodic phrases expressing specific music ideas or motifs can be grouped into a *melodic section* ("qiangjie") which can play an integrating role in the overall musical form. The *metrical pattern* is the most expressive characteristic element of Jingju. The transitions of alternating metrical patterns in a Jingju song may indicate boundaries between sectional units [Repetto et al. 2014; Srinivasamurthy et al. 2014]. There are fixed types of metrical patterns, each associated with certain melodic tendencies and dramatic contexts. Metrical patterns can be classified into the *metred* and *free* categories based on whether their beat styles have accented beats and specific metric regulations or are free of them. Besides the sung sections, Jingju songs also have instrumental sections and percussion sections bridging the sung parts in the arias. The percussion is mainly cymbals and gongs which rarely overlap the singing and the melodic instrumental sections. Functionally, the instrumental and percussion parts serve to introduce melodic passages and to connect successive melodic lines, hence are integral to Jingju structure.

Jingju corpora have been presented in two recent works. The first addresses mood estimation in singing [Black et al. 2014] and the second for melody analysis [Repetto and Serra 2014]. However, they mainly feature singing properties and are less relevant to the present research. In the next section, we will introduce a new corpus designed for the purpose of structural analysis.

2.2. Related Work in Music Structural Segmentation

Techniques for MSS fall into three categories: *novelty*-, *homogeneity*-, and *repetition*-based. Novelty-based methods rely on the hypothesis that segment boundaries are characterised by prominent changes in audio features. One classical example of these methods is introduced in [Foote 2000] by correlating a Gaussian-tapered kernel with the main diagonal of the self-similarity matrix (SSM) computed from the audio features, resulting in a function commonly denoted the *novelty curve* indicating segment boundaries as peaks in the novelty function. Homogeneity based approaches, also referred to as the *state* approaches, assume homogeneities in local statistical properties of features in individual structural sections. One common practice is to represent the sections as *states* in a *hidden Markov model (HMM)* [Levy and Sandler 2008]. Alternatively, *repetition*-based approaches attempt to find repetitive patterns as indicators of sectional units. Such repetitions commonly form stripe structures in the sub-diagonals of the SSMs from audio features or patterns in the state sequences from statistical representations. Besides using these methods individually, some work attempts to combine them to derive descriptors integrating multiple structural principles and musical properties [Serrà et al. 2012].

Various audio features have been used to analyse music structure capturing mainly its harmonic, timbral and rhythmic content, which are identified as the most important structural descriptors [Paulus et al. 2010].

The *chromagram* [Fujishima 1999], also called *Harmonic pitch class profiles* (HPCP), along with its variants is the most frequently used feature for the structural analysis of Western pop music. The chromagram is a B-dimensional vector representation denoting the relative intensity of each semitone in a chromatic scale, where B is the

number of *bins per octave* (BPO). While the 12-BPO setting is intuitively adopted for most studies for Western pop music, several works have proposed different BPO settings for specific tasks or music genres [Harte and Sandler 2005]. In the numbered notation Jingju uses, when C is the keynote, 1, 2, 3, . . . , 7 correspond to C, D, E, . . . , B. Importantly, [Liu et al. 2009] demonstrates that when mapped into a 12-dimensional chroma scale, the energy distribution of a Chinese traditional music piece is much less dispersed than that of Western classical music, with around 90% of the energy distributed in frequency components corresponding to five notes (C, D, E, G, A). [Chen 2013] analyses the pitch histogram for a Jingju collection and confirms the use of pentatonicism with small energy distribution also presented for the 4th and the 7th degree notes, which use a different tuning scale than the equal temperament. These two have very expressive roles in modulating between keys in Jingju performance hence are indispensable for the analysis of its pitched content [Wichmann 1991]. Hence, in this work, we investigate the 7-BPO chromagram feature for MSS of Jingju music.

The *Mel frequency cepstral coefficients* (MFCCs) feature models the shape of the spectral envelope by describing the frequency spectrum transposed to a perceptual scale in a compact form [Logan 2000]. MFCCs are among the most popular timbre features in MSS research [Aucouturier et al. 2005].

Rhythmic information may also identify music structure. It is however much less employed compared to the timbre and harmony alternatives [Paulus et al. 2010]. In this article, we will revisit these three types of audio features for the structural analysis of several distinct music genres.

The success of an audio-based MSS system largely depends on the signal processing techniques used. One commonly employed technique is to use *beat-synchronised* audio features [Levy and Sandler 2008]. This is especially effective for some Western pop music with predicable beat patterns that can be considered as the basic unit of a potential structural decomposition. However, as discussed in Section 2.1, accented beats may be lacking and the tempo can be highly flexible in Jingju, resulting in limited efficacy of common beat tracking algorithms.

MFCCs have been reported as presenting problems in expressing both harmonic and percussive contents when they present at the same time in a music genre classification study [Rump et al. 2010]. Furthermore, the heavy use of cymbal and gong instruments in Jingju can mask the rest of the instrumentation components whose timbral characteristics might hold more fine detail [Tian et al. 2014]. Harmonic-percussive source separation (HPSS) is a well-studied task concerning separating the input audio signal into its harmonic and percussive components. Gkiokas et al. point out that HPSS has the tendency to improve the accuracy of music tempo estimation [Gkiokas et al. 2012]. However, the study of its effects for music structural analysis is lacking from the literature. In this article, we will investigate the HPSS technique as a pre-processing step for feature extraction in an MSS scenario.

3. MUSIC CORPORA

3.1. Existing Collections for Music Structural Segmentation

Two of the publicly available databases collected for MSS research are used in this work. The first consists of 174 songs from The Beatles. It was first manually annotated at Music Technology Group (MTG), Universisitat Pompeu Fabra (UPF) and corrected at Tampere University of Technology (TUT) [Paulus and Klapuri 2008b]. We denote this dataset *BeatlesTUT*. The SALAMI Internet Archive dataset (S-IA) is a publicly available subset of the full database collected in the SALAMI project[2] com-

[2]https://ddmal.music.mcgill.ca/research/salami

prising 272 pieces [Smith et al. 2011]. The main consideration of the SALAMI dataset was to cover a wide variety of musical genres, mainly including Western classical music, popular music, jazz and world music, and was particularly intended to provide a textbook example of Western pop music. This dataset has a diversity of audio qualities by including a large set of live recordings. It is also used for the structural segmentation evaluation in Music Information Retrieval Evaluation eXchange (MIREX), an international community-based evaluation campaign for various MIR tasks held annually[3]. Note that S-IA has an overlap with BeatlesTUT of 35 songs, although the actual recording conditions may differ.

These datasets are based on different annotation principles. BeatlesTUT is annotated with section labels mainly including: "*intro*", "*verse*", "*chorus*", "*bridge*", "*refrain*" and "*outro*" with their variations, as well as a few others such as "*break*" and "*silence*". Such annotation is made on a *functional* level, i.e., the music is segmented into structural parts expressing specific musical functions. A potential problem is that the use of function labels can conflate the notion of musical similarity with musical function and can cause uncertainties in annotation decisions [Peeters and Deruty 2009; Smith 2010]. In contrast, S-IA is annotated on multiple scales incorporating the approach proposed in [Peeters and Deruty 2009]. In the lowest *music similarity* level, the segments are identified to address similarities in "music ideas". The *function* level annotation is similar to that of BeatlesTUT but with more limited section types. Finally the highest *lead instrument* level defines structural sections by searching for dominating instrumentation they consist, such as "*vocal*" or "*guitar*". In this paper, we use the music similarity level annotations for S-IA. It is because these two datasets use different annotation principles that they serve as a comprehensive testbed for the segmentation algorithms and will offer a meaningful reference for the analysis of Jingju music.

3.2. Jingju Structural Segmentation Database

The Jingju corpus used in this article is composed of 30 excerpts from commercial CDs [(CMG) 2010], sampled at 44.1 KHz and 16 bits per sample with a total length of 3.6 hours. The CDs were released in the past decade and are recordings of classical repertoires by the most renowned performers.

A full Jingju play can last several hours, comprising multiple acts. For the purpose of this study, the excerpts consist of melodic passages taken from arias, with an average length of 432 seconds. They were selected on the criteria of repertoire coverage, structural diversity and audio quality. One prerequisite for an excerpt is that various structural parts should be present characterising temporal progressions or changes of sectional units. The selected samples in the collection cover the two main modes (*xipi* and *erhuang*) and various metrical patterns. Half of them are performed by female singers and half by male singers, covering different role types.

3.3. The Annotation Process

In this work, annotations are arranged to describe the *musical similarity* within a piece setting aside the musical functions of segments, just as in the lowest level of S-IA. This is for two reasons. First, functional or lead instrumentation annotations can be highly genre-dependent, meaning that segmentation results of one dataset are not necessarily comparable to those of another, whereas low-level music similarity is a phenomenon that can be observed across different genres [Deutsch 2012]. Assessing the structure on a music similarity level provides a fair comparison between genres and datasets. Second, the melodic sections are never repeated as chorus-verse based music

[3]http://www.music-ir.org/mirex/wiki/MIREX_HOME

forms would do and there is much expressiveness in the performance. This can necessitate the analysis of the ornamentations in parallel to defining the functional structure, thus introducing uncertainties in locating sectional boundaries. It is plausible to set a flexible and sufficient range for the temporal location of a segment boundary, but this would raise the demand for new evaluation metrics tailored for this music genre, which is outside the scope of this study. Annotations created at such a fundamental level also allows for conveying semantic or musicological meanings given further grouping.

Three listeners (”A1”, ”A2” and ”A3”) participated in annotating the music. Another two engaged in verifying their annotations, one of which is the first author of this paper (noted ”V1”) and is familiar with this music style as an amateur, the other is a Jingju musician and musicologist (noted ”V2”). All annotators are Chinese and were provided with music scores and lyrics [wenyi chubanshe 1992]. The software used for annotation is Sonic Visualiser which displays the waveform and the corresponding spectrogram of the music[4]. This dataset is denoted *CJ* in the remainder of this paper. Associated metadata is available online[5].

In this process, A1, A2 and A3 firstly worked independently, each producing annotations on their own. They were instructed to assign each syllable they hear in the audio to the (Chinese) character in the lyrics. They were asked to listen to prominent changes in music phenomena such as rhythm, melody, harmony or timbre, and mark the boundaries in places where the similarities break. Within a section, high similarity should present with a single musical idea or subject expressed. V1 and V2 then independently went through the 3 annotations to verify their disagreements and possible inconsistencies. Each would record a boundary annotated by only one of A1, A2 and A3 – hence the other two disagree with him – and then decide whether it should be marked and if yes, its exact position.

Fig. 1 shows respectively the annotations by V1 and V2 and the final accepted annotation for an 60-second excerpt of the recording ”Ba wang bie ji” (meaning ”Farewell my concubine”), with the corresponding lyrics shown on the top. The phrase shown constitutes half a couplet. We can notice that this phrase is sung at a relatively slow tempo and that a single sung character may last several seconds. This gives the performer lots of freedom in the singing, where each syllable can be sung with ornamentations such as vibrato and even intermittence.

Rather than adopting the common approach for grouping two sets of annotations by averaging event positions, the final annotation decision is a result of conscious discussions by V1 and V2 based on their individual work. The reason for this is that V1 and V2 each has noted different number of boundaries and there is not necessarily a match for a boundary from one set in another. We however realised that the discussion can produce different boundaries, i.e., the final accepted boundary location may differ from the locations indicated individually by both V1 and V2, as shown in Fig. 1. One main reason for the uncertainties in deciding the exact temporal position of an underlying boundary is that, the emergence of new sections may be accompanied by gradual changes of acoustical properties, for example, the sustaining decaying of cymbal instruments and the fade-out effect of singing. Such temporal disparities of an accepted boundary from those indicated by V1 and V2 individually however barely lead to dubious evaluation results given a sufficient acceptance window for the retrieved boundaries. In this work, detected segment boundaries are accepted to be correct if within 3s from an annotated one in the ground truth following [Levy and Sandler 2008].

[4]http://www.sonicvisualiser.org/
[5]http://isophonics.net/content/jingju-structural-segmentation-dataset

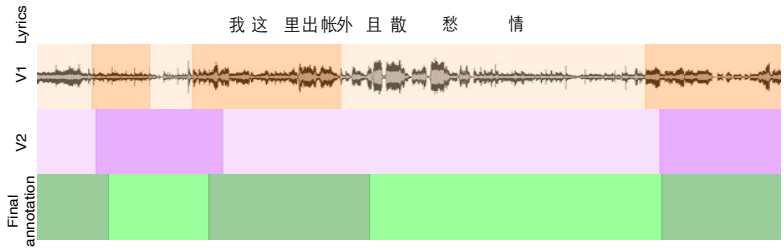


Fig. 1: Boundary annotations for a 60-second excerpt of the recording "Ba wang bie ji" from Dataset *CJ*. Panes from top to bottom pane show respectively the lyrics of the singing (in Chinese), annotations by annotator V1 and V2 and the final annotation.

$F_{0.5}$	F_3	M_{ad}	M_{da}	S_{ad}	S_{da}
0.693	0.743	11.88	0.27	74.31	0.97

Table I: Average agreement between annotator V1 and V2 for recordings in dataset CJ. $F_{0.5}$ and F_3: boundary retrieval F-measure obtained at a resolution of 0.5s and 3.0s; M_{ad} and M_{da}: median distance between each annotated segment boundary to its closest detected segment boundary (in second); S_{ad} and S_{da}: standard deviation of distance between each annotated segment boundary to its closest detected segment boundary (in second).

3.4. Statistics of the Annotations

From the variety of existing measures commonly used to compare multiple annotations [Smith 2010], we now discuss the *inter-annotator agreement* between V1 and V2. This means we analyse the accuracy first of V2 against V1, with the former playing the role of "detection" and the latter the role of "ground truth". Then their roles are reversed. Finally, averages are taken.

The analysed statistics include: F-measure retrieved at the tolerance of 0.5s ($F_{0.5}$) and 3s (F_3), median of the distance between each annotated segment boundary to its closest detected segment boundary (M_{ad}) and that between each detected segment boundary to its closest annotated segment boundary (M_{da}), standard deviation of the distance between each annotated segment boundary to its closest detected segment boundary (S_{ad}) and between each detected segment boundary to its closest annotated segment boundary (S_{da}).

As shown in Table I, the agreement between the annotators measured at 0.5s ($F_{0.5}$ = 0.693) is reasonably close to that measured at 3.0s (F_3 = 0.743). This shows that once V1 and V2 both indicate the acceptance of a boundary, they report relatively close temporal locations of it. However, there exists a large discrepancy when comparing the median or the standard deviation of the distances from one set of annotation to the another. This is mainly because the two annotators noted different numbers of segment boundaries, as shown in Fig. 1. This suggests that the structural annotations do depend on the annotators' individual understanding of the music just as is observed for Western music [Smith et al. 2014].

Statistics of datasets used in this paper describing the number and average length of the excerpts are given in Table II. We notice that the average segment length of S-IA and CJ are on average much shorter than those of BeatlesTUT.

Dataset	No. tracks	Len. track	No. segments	Len. segment
BeatlesTUT	174	159.30 (50.08)	10.21 (2.32)	17.73 (5.45)
S-IA	258	333.09 (130.78)	56.26 (32.07)	7.69 (5.28)
CJ	30	421.38 (219.02)	44.37 (19.18)	9.56 (4.57)

Table II: Statistics of datasets (standard deviations into parenthesis): number of samples in the dataset, average length of each sample (in second), average number of segments per sample, average length of each segment (in second).

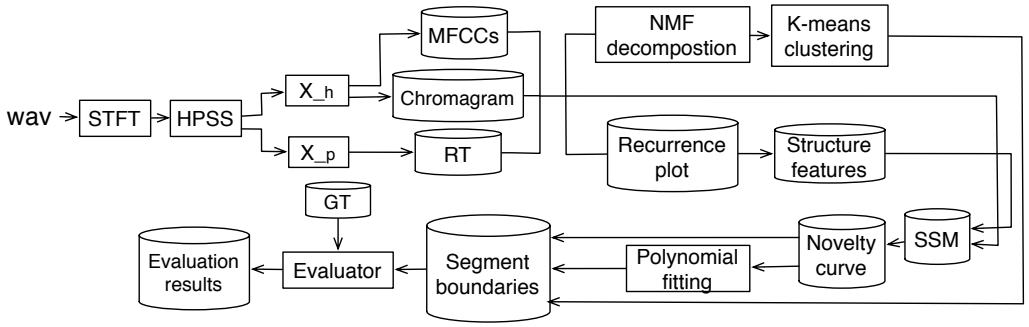


Fig. 2: Workflow of the feature extraction and music segmentation system.

4. FEATURE EXTRACTION

In this work, three features related to three musical aspects – harmony, timbre and rhythm, are evaluated. The workflow of our feature extraction and segmentation system is shown in Fig. 2. We will introduce the harmonic-percussive source separation operation as pre-processing for feature extraction.

4.1. Harmonic-percussive Source Separation

Given the complex spectrogram X of the input audio signal $x(t)$, we separate it into its harmonic component X_h and percussive component X_p. We denote as Y the magnitude spectrogram where $Y = |X|$. In this work, the separation is realised by applying a median filter to Y once in the horizontal direction and once in the vertical direction to derive respectively a harmony and percussion enhanced spectrogram following [Driedger et al. 2014] as a modification to [FitzGerald 2010].

However, this process might be influenced by vibrato in the singing voice and bowed string instruments in Jingju, which yield frequency variations over a small time period [Yang et al. 2015]. A possible solution is to consider widened frequency bins in the neighbourhood to generate a vertical mask. For the percussive component suppression in the harmony enhanced spectrogram, we also propose to widen the masking trajectory across neighbouring time instants in case the transient energies are varying. A *maximum filter* has the capacity to broaden the spectral trajectory by setting the value at certain position in the trajectory to the maximum value in the vicinity, hence can suppress the spurious positives in energy transients. As a modification to [Driedger et al. 2014], we introduce a one-dimensional maximum filter processing before generating the mask. The maximum filter is applied to the magnitude spectrogram vertically for the separation of the harmonic component and horizontally for the separation of the percussive component, taking the opposite directions of the median filter applied subsequently. In this way, the harmonic/percussive slices can be strengthened before the separation, leading to a highlighting effect of the corresponding sources combating

the possible interference of vibratos or energy sways. The whole process is described as follows:

$$Y_h{}^m(t,k) = \mathsf{max}(Y(t,k-m_h:k+m_h)) \tag{1a}$$

$$Y_p{}^m(t,k) = \mathsf{max}(Y(t-m_p:t+m_p,k)), \tag{1b}$$

$$\tilde{Y_h}{}^m(t,k) = \mathsf{median}(Y_h{}^m(t-l_h:t+l_h,k)) \tag{2a}$$

$$\tilde{Y_p}{}^m(t,k) = \mathsf{median}(Y_p{}^m(t,k-l_p:k+l_p)), \tag{2b}$$

Next, the derived masks are applied to the original X to compute correspondingly the separated source X_h and X_p,

$$X_h(t,k) = X(t,k)\cdot\left(\tilde{Y_h}{}^m(t,k)/(\tilde{Y_p}{}^m(t,k)+\epsilon) > \beta\right) \tag{3a}$$

$$X_p(t,k) = X(t,k)\cdot\left(\tilde{Y_p}{}^m(t,k)/(\tilde{Y_h}{}^m(t,k)+\epsilon) > \beta\right), \tag{3b}$$

The sample rate, window and step size used in this study are respectively 44100 Hz, 0.046s and 0.023s. For l_h, l_p, m_h, $m_p \in \mathbb{N}$, $2l_h+1$, $2l_p+1$, $2m_h+1$ and $2m_p+1$ are respectively the sizes of the median and maximum filters equal to 0.23s, 350 Hz, 0.07s and 70 Hz. ϵ is a small constant to avoid zero division and β is the separation factor to control the ratio of specific component to separate experimentally set to 0.5. We will discuss the involved parameters in Section 6.

4.2. Audio Features

Features extracted from the spectrogram after the harmonic-percussive source separation processing (HPSS) with the maximum filter applied are denoted $hMFCCs_m$ and $hChromagram_m$ in the rest of this article as opposed to $MFCCs$ and $chromagram$ that are extracted from the raw spectrogram X. We also compute the features with HPSS but without maximum filtering, and label these $hMFCCs$ and $hChromagram$.

In this paper the 13-dimension MFCCs are extracted with a 0.046s window and a 50% overlap where the number of Mel filters is 40.

As discussed in Section 2, Jingju uses a different chroma scale to Western pop music. In this paper, we use the $BPO = 7$ setting to extract the chromagram feature for Jingu, i.e., the CJ dataset, while the conventional 12-bin chromatogram is used for the two Western datasets, S-IA and BeatlesTUT. We will discuss the effect of the number of pitch classes per octave in Section 6. The window and the step size used for feature extraction are set to 0.372s and 0.023s respectively.

For most music genres, various instruments can be prominent at different metrical levels and play diverse roles to produce the overall rhythmic structure in a piece [Parncutt 1994]. Although rhythmic features have been investigated in previous work [Jensen 2005; Paulus and Klapuri 2008a], the novelty of the featureset investigated in this paper lies in the incorporation of tempo perception cues introduced shortly. An autocorrelation tempogram is first calculated following [Davies and Plumbley 2004]. The time window used is 6s with a step size of 0.2s. Instead of targeting a rigorous tempo or beat tracking, features are extracted inspired by acoustical perception experiments [Moore et al. 1997]. To characterise the specific tempo strength, we first group the tempogram BPM bins into quasi-logarithmic spaced bands. Two features, *Tempo intensity* (TI) and *Tempo intensity ratio* (TIR) are extracted respectively

by compressing the bandwise intensity values inspired by the calculation of the perceptual feature *specific loudness* [Rodet 2001] and by measuring the intensity ratio of each band to describe the perceived relative salience of individual rhythmic components. Although additional features are presented, the concatenation of TI and TIR feature vectors achieves the best segmentation in the evaluation [Tian et al. 2015]. In this paper, we replicate this process and use the concatenation of TI and TIR as the tempogram-derived rhythmic feature used in this paper and note the feature extracted with no HPSS applied RT. We use the percussive spectrogram X_p after HPSS to extract the feature which we note pRT_m and pRT respectively for cases with the maximum filtering included and excluded.

All features including their HPSS variants were further resampled to obtain a uniform frame rate of 0.2s. Features were then subjected to a 6-dimensional Principal component analysis (PCA) before used for segmentation because preliminary results have shown that the PCA has introduced marginal improvements in the segmentation. The use of a relatively large window would assist forming the structure description on a musically meaningful scale [Paulus et al. 2010].

5. MUSIC STRUCTURAL SEGMENTATION

The first segmentation algorithm, denoted *Quadratic novelty* (QN), retrieves segment boundaries using a polynomial fitting mechanism based on [Foote 2000]. We first compute the Self-Similarity Matrix (SSM) using the pairwise Euclidean distance of the feature matrix. A novelty curve is generated from the SSM following [Foote 2000]. A series of post-processing and peak picking procedures are applied to the derived novelty curve to select boundaries following [Tian et al. 2015]. First, normalisation and DC removal are applied to the raw novelty curve. The normalised novelty curve is then passed through a low-pass filter for noise removal. Subsequently, adaptive thresholds are generated from the smoothed novelty curve using a median filter. Finally, boundaries are retrieved using a polynomial fitting based method.

We fit a second-degree polynomial on the smoothed novelty curve centred around each local maximum obtained from the adaptive thresholding using a window of 5 samples. This estimates the coefficients of the second-degree quadratic function $y = ax^2 + bx + c$, where coefficients a and c correspond respectively to the sharpness and the amplitude of each peak. The coefficient b is not assessed as the acceptance of a peak depends on only the shape and amplitude of the parabola in our system. A peak will be accepted as a segment boundary when both the following conditions are satisfied: $a > th_a$ and $c > th_c$, where th_a and th_c are computed from a single sensitivity parameter $sens$ and two experimentally defined values using $th_a = (100 - sens)/1000$ and $th_c = (100 - sens)/1500$ ($sens \in [0, 100]$). This method is inspired by the QM Vamp Onset Detector[6]. Hence the higher the sensitivity, the looser the condition is and the more boundaries will be retrieved from the novelty curve. In this work, we use an experimentally defined setting of $sens = 30$.

The second segmentation algorithm relies on Non-negative matrix factorisation (NMF) with a convex constraint [Nieto and Jehan 2013], denoted *CNMF* in this paper. In NMF-based segmentation, the input matrix V represents a feature matrix or its SSM, where $V \in R^{N \times M}$ for the first and $V \in R^{N \times N}$ for the latter, and N is the number of frames and M is the number of features. With an NMF decomposition, V can be approximated as the product of two non-negative matrices W and H, where the $N \times R$ matrix W contains the basis vectors, the $R \times M$ matrix H supplies in its columns the coefficients to approximate each column of V as the linear combination of

[6]QM-DSP audio analysis C++ library: https://code.soundsoftware.ac.uk/hg/qm-dsp

the columns of W, and R is the rank of decomposition. Finally, segment boundaries can be detected by clustering the frames in the decomposition matrices as row-vector features [Kaiser and Sikora 2010; Grohganz et al. 2013]. Based on this approach, CNMF uses the feature matrix as the input V and introduces a convex constrain to W such that it becomes the convex combinations of the input feature matrix V, expressed as $W = VC$ where $C \in R^M \times r$. In this way, each observation frame of W can be interpreted as weighted cluster centroids representing potential sections of the music piece. To detect segment boundaries from the decomposition matrices, a k-means clustering with the cluster number set to 2 is carried out where the 2 classes represent respectively if there is a boundary or not. Finally, boundaries detected from each rank are grouped and merged into the average locations when they locate in a given temporal window where the final boundary decisions are made. One parameter involved in this process is the decomposition rank r. Kaiser and Sikora reported a maximum of separability with $r = 9$ [Kaiser and Sikora 2010] while in [Nieto and Jehan 2013] it is set to 2. We found from preliminary research that the setting of this parameter does not have a profound effect in the segmentation results when using values in a moderate range between 3 and 7. In this work r is set to 3.

The third segmentation method from [Serrà et al. 2012] uses a feature called *structure features* incorporating global properties which account for structural information in the recent past. It is denoted ***SF*** in this paper. To construct the structure features, a multi-dimensional time series is firstly obtained by accumulating vectors of the standard audio feature ranging across a span centred at different time locations. A *recurrence plot* P is then computed from the pairwise resemblance between time series. An element $P_{i,j}$ of the recurrence plot is set to 1 when two time series centred at time i and j are sufficiently close and to 0 otherwise. The homogeneous and periodic nature of the typology of a recurrence plot enables addressing the local stationarity and the global repetition from the time series [Eckmann et al. 1987]. Subsequently, the structure features are obtained by estimating the temporally spanned Gaussian probability density of the time lag matrix of P. Finally, a novelty curve is computed denoting the distances between consecutive samples of structure features where segment boundaries are detected using a standard thresholding mechanism following [Foote 2000]. In this way, all three segmentation mechanisms (novelty, homogeneity and repetition) are combined.

Fig. 3, 4 and 5 show respectively the segmentation processes using the three algorithms on an excerpt of Jingju song "Hong niang" (meaning "The red maid") from dataset CJ with feature $hChromagram_m$. From Fig. 3 we can see that the novelty scores associated with the annotated segment boundaries in the raw novelty curve can be relatively subtle. Compared to standard boundary retrieval algorithms which decides the acceptance of potential peak by comparing its value to a threshold, polynomial fitting is able to eliminate "flatter" peaks which in our scenario indicates higher similarity within feature vectors in the vicinity. Fig. 4 shows the k-means clustering results from each decomposition matrix with each rank shown in each row from the segmentation process of CNMF. The white vertical lines in the upper pane show the detected boundaries by merging boundary decisions derived from each decomposition matrix. We can notice that the final grouping has removed many boundaries indicated by only few decomposition matrices especially when the sections are of short durations. Fig. 5 shows the recurrence plot and the derived novelty curve using ***SF***. Instead of exhibiting stripe structures as Western pop music normally does [Serrà et al. 2012], only block structures are presented. This demonstrates the non-repeating nature of the music from a global perspective as discussed in Section 2. Consequently, boundaries are mainly derived from the local homogeneities in audio features.

Fig. 3: Segmentation process on Jingju music excerpt "Hong niang" using $hChromagram_m$ by algorithm *QN*. The black vertical lines, green triangles and red crosses represent respectively the annotations, detected boundaries and those would also have been retrieved without the polynomial fitting (using adaptive thresholding).

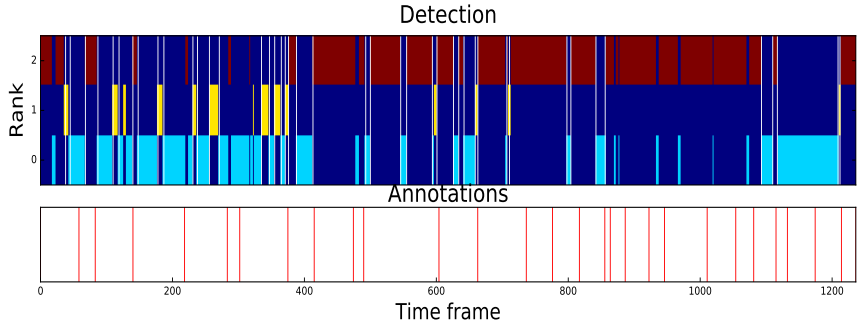


Fig. 4: Segmentation process on Jingju music excerpt "Hong niang" using $hChromagram_m$ by algorithm *CNMF*. The upper pane shows the k-means clustering results from each decomposition matrix where the duo colours in each row indicate section divisions. The white and red vertical lines in the upper and bottom pane shows the retrieved boundaries and the annotations.

6. RESULTS AND DISCUSSION

6.1. Evaluation Metrics

There has been a wealth of research investigating the evaluation frameworks for MSS tasks [Smith 2010; Schedl et al. 2014]. In this paper, a detected boundary is accepted as a *true positive* (TP) if located within a 3s-window from a boundary in the ground truth [Levy and Sandler 2008]. The quality of segmentation is assessed with the standard segment boundary recovery *precision* (P), *recall* (R) and *F-measure* (F)[7].

[7]Music structural segmentation evaluation metrics: http://www.music-ir.org/mirex/wiki/2009:Structural_Segmentation#Evaluation_Measures

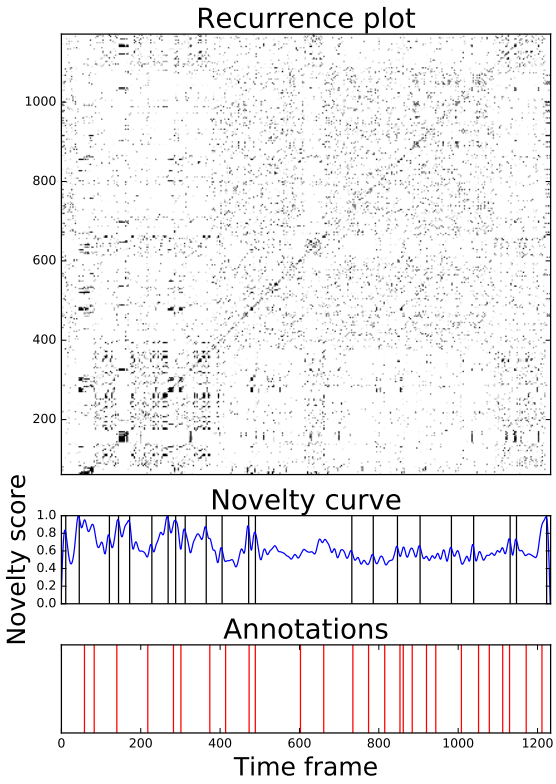


Fig. 5: Segmentation process on Jingju music excerpt "Hong niang" using $hChromagram_m$ by algorithm *SF*. The upper, middle and bottom pane show the recurrence plot, the novelty curve and retrieved boundaries, and the annotations respectively.

6.2. Harmonic-percussive Source Separation for Music Structural Segmentation

Table III illustrates the average segmentation results for audio samples in dataset S-IA and CJ using investigated features (MFCCs have been rescaled for non-negativity for CNMF). We report evaluation for these two datasets here because their annotations are made based on comparable principles. Results reported for CNMF and SF in Table III differ from the MIREX results of [Nieto and Jehan 2013] and [Serrà et al. 2012][8] mainly due to different system configurations, and that S-IA and the MIREX testset are both subsets of the full SALAMI dataset hence not equivalent to each other.

The effect of the harmonic-percussive source separation (HPSS) is illustrated in Table IIIa comparing results obtained using features without HPSS, with HPSS and without/with maximum filtering. The significance level of the differences between a feature with and without HPSS in terms of segmentation F-measures obtained for

[8]http://nema.lis.illinois.edu/nema_out/mirex2012/results/struct/sal/summary.html, http://nema.lis.illinois.edu/nema_out/mirex2014/results/struct/sal/summary.html

	P	R	F	P	R	F	P	R	F
	$Chromagram$			$MFCCs$			RT		
S-IA	0.414	0.728	0.476	0.521	0.693	0.505	0.438	0.753	0.489
CJ	0.390	0.594	0.394	0.384	0.612	0.445	0.415	0.557	0.426
	$hChromagram$			$hMFCCs$			pRT		
S-IA	0.427	0.741	0.501	0.546	0.654	0.507	0.449	0.701	0.476
CJ	0.421	0.615	0.443†	0.427	0.621	0.462	0.437	0.528	0.424
	$hChromagram_m$			$hMFCCs_m$			pRT_m		
S-IA	0.436	0.767	0.516†	0.558	0.678	0.513*	0.467	0.636	0.501
CJ	0.441	0.659	0.455†	0.458	0.612	**0.487***	0.451	0.519	0.446*

(a) *QN*

	$Chromagram$			$MFCCs$			RT		
S-IA	0.448	0.411	0.418	0.425	0.473	0.431	0.457	0.387	0.407
CJ	0.444	0.353	0.367	0.451	0.372	0.396	0.403	0.314	0.332
	$hChromagram$			$hMFCCs$			pRT		
S-IA	0.479	0.445	0.462*	0.454	0.485	0.460*	0.487	0.395	0.416
CJ	0.465	0.388	0.392	0.470	0.406	0.408	0.463	0.352	0.366
	$hChromagram_m$			$hMFCCs_m$			$pRTm$		
S-IA	0.495	0.486	0.491†	0.489	0.496	0.486†	0.460	0.411	0.427*
CJ	0.478	0.401	0.404*	0.489	0.393	0.421*	0.468	0.376	0.396†

(b) *CNMF*

	$Chromagram$			$MFCCs$			RT		
S-IA	0.445	0.539	0.463	0.423	0.550	0.478	0.436	0.567	0.484
CJ	0.454	0.307	0.322	0.470	0.321	0.354	0.445	0.332	0.349
	$hChromagram$			$hMFCCs$			pRT		
S-IA	0.459	0.545	0.497*	0.471	0.566	0.486	0.454	0.501	0.468
CJ	0.458	0.335	0.372*	0.476	0.342	0.383*	0.473	0.368	0.379*
	$hChromagram_m$			$hMFCCs_m$			pRT_m		
S-IA	0.487	0.575	**0.521**†	0.486	0.567	0.505*	0.456	0.509	0.480
CJ	0.471	0.343	0.397†	0.502	0.384	0.418‡	0.466	0.388	0.387*

(c) *SF*

Table III: Segmentation results using selected features on *S-IA* and *CJ* dataset with method *QN*, *SF* and *CNMF*. *P*, *R*, *F*: Segment boundary recovery precision, recall and F-measure measured at *3s*. Highest F-measure for each dataset is shown in bold. *, † and ‡ denote the presence of significant improvement in segmentation F-measure for features extracted with HPSS over the standard versions (without HPSS) on each dataset at the level of 0.05, 0.01 and 0.001 using the Wilcoxon signed-rank test.

each audio sample in a dataset is measured using Wilcoxon signed-rank test. The most notable improvements are observed for *chromagram*, with $p < 0.01$ for most cases with a maximum filtering, second to which are MFCCs. Although HPSS has improved the segmentation when using MFCCs and chromagram features in general, its actual effects for each lies in improving respectively the *precision* and the *recall*. This is mainly because the low-level timbre similarities encoded in MFCCs may incur limited segmentation precision in the first place, especially for novelty-based methods; while chromagram tends to depict the long-term repetition structures and can overlook the low-level novelty-associated boundaries, so that there might have been room for improvements in recall, especially for repetition-based methods.

However, the opposite is observed for the two datasets using the *RT* feature. For *CJ*, HPSS yields improved segmentation. We found the onset detection algorithm and subsequently the tempo tracking may work less effectively in the presence of singing voice the constant tempo variation. Removal of the harmonic components in the spectrogram therefore is beneficial to the tempogram computation and the following fea-

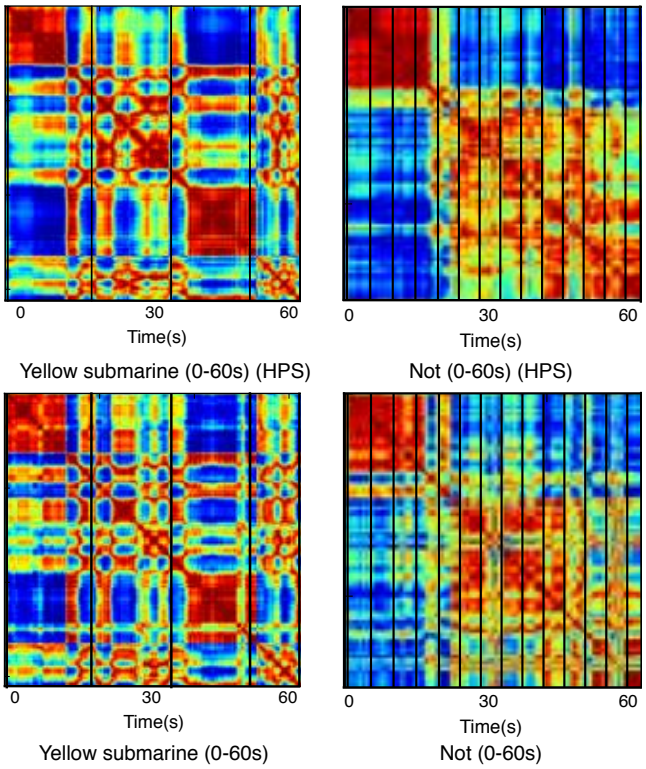
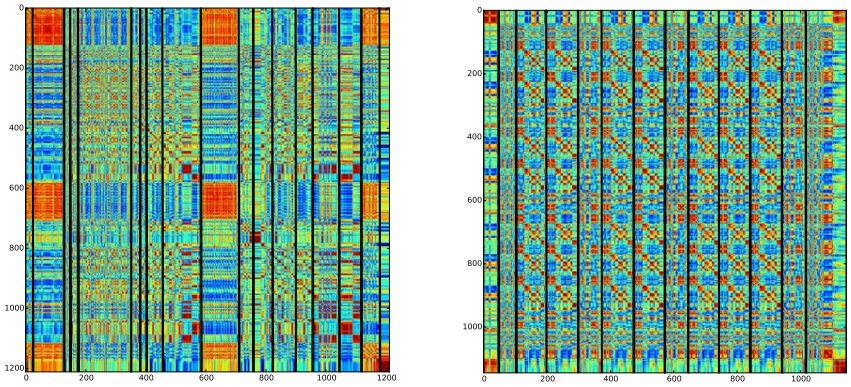


Fig. 6: SSMs computed with pRT_m (top) and RT features (bottom) for 60-second excerpt of music "*Yellow Submarine*" By *Beatles* (left) from *BeatlesTUT* and "*Not*" by *Box O'Laffs* from *S-IA* (right). Black vertical lines indicate segment boundaries.

ture extraction. We also observed that the occasional absence of accented beats in the music has degraded the accuracy of the tempogram calculation. Analysis of the results for *S-IA* indicates, however, that using RT features with HPSS may lead to an *under-segmentation*. The block structures are made cleaner in the SSM, resulting in less false positives and a higher precision ($p < 0.05$). Nevertheless, this is achieved at the cost of a substantial degradation in true positives, generating degraded recall.

To validate this observation, we repeated this experiment on dataset *BeatlesTUT*, whose annotations are made on a functional level (see Section 3.1). The segmentation using all three investigated features are significantly improved after HPSS including RT ($p < 0.001$). Fig. 6 shows SSMs of the RT features with and without HPSS (maximum filter applied) for two pieces from BeatlesTUT and S-IA (only the first 60 seconds of the tracks are shown for visualisation purposes). Smaller blocks are aggregated into bigger ones as a result of HPSS, yielding less local dynamics hence less false positives. However, this may also cause SSMs to fail to represent structural details corresponding to low-level annotations. One limitation of the RT feature is a degraded temporal resolution as a result of using long window (6s) with considerable overlap during the tempogram calculation. Applying the maximum filter in the HPSS operation can accentuate this resolution deficiency and lead to more missed boundaries. Therefore



(a) Tian nu san hua (b) Dig a pony

Fig. 7: SSMs computed using $hChromagram_m$ for music "*Tian nv san hua*" from *CJ* (left) and "*Dig a pony*" By *Beatles* (right).

when applying HPSS, its influence on feature resolution can be an additional factor to consider for an effective segmentation.

The improvement due to HPSS in segmentation is altered by its parameters. The most influential parameter in our case is the separation factor β. [Driedger et al. 2014] report that when $\beta = 1$, the residual is roughly equally distributed in both X_h and X_p while when $\beta = 3$ only clearly horizontal and vertical structures are preserved in the spectrogram. We tested values ranging from 0.3 to 3 (in steps of 0.1 when β ranges from 0.3 to 1 and of 0.5 from 1 to 3). Applying a maximum filter also has the tendency of leaving residual components in the resulting signal. A β ranging from 0.4 to 0.5 is optimal for all investigated features in our system (Table III results are obtained with $\beta = 0.5$). When β exceeds 1.5, extracted features yield worse segmentation results than when not using HPSS. In the case of music structural analysis, it is not desirable to have the opposite source and the residuals tightly removed when using only X_h or X_p for feature extraction, given each may contain complementary structural information.

6.3. Effect of Bins per Octave in Chroma for Jingju

Although chroma features were originally designed for chord recognition for Western music [Fujishima 1999], they measure the relative intensity of each pitch class of an equal-tempered scale in a tuning independent way. This justifies their use for Jingju music which uses equal temperament except for its 4th and 7th degrees which have more musical expressiveness functions. However, because there are no repetitive harmonic patterns, the chroma feature does not form stripes in the sub-diagonals in the SSM. This somehow contradicts observations for Western pop music, as in Fig. 7. Therefore, the same audio feature may capture different structural characteristics for different genres and the design of segmentation algorithms should be adapted accordingly to interpret such patterns. In this paper, we use chromagrams with 7 bins per octave (BPO) setting for Jingju. For both the original chromagram and its variants with HPSS, we compare the segmentation results to those adopted using the 12-BPO setting commonly used for Western music. 7-BPO achieves better segmentation for all cases. The difference is however significant only for $hChromagram$ and $hChromagram_m$ with $p = 0.036$ and 0.014 respectively. Another interpretation of this is that the chroma feature may be more effective, if it is specially adapted or enhanced for Jingju. A direction

	CNMF			SF			QN		
	P	R	F	P	R	F	P	R	F
BeatlesTUT	0.489	0.636	0.540	0.681	0.737	**0.699**‡	0.465	0.646	0.527

Table IV: Segmentation results using beat-synchronised chroma features on *BeatlesTUT* dataset using the three investigated algorithms *CNMF*, *SF* and *QN*.

for future work also lies in analysing the behaviour of the 4th and the 7th pitch classes in the Jingju chroma scale.

6.4. Features, Segmentation Methods and Genres

Here we discuss the three types of features regardless of the effect of HPSS. Chromagram and MFCCs work reliably well for Western music, confirming the conclusion of previous studies [Paulus and Klapuri 2008a]. MFCCs have consistent performance for both CJ and S-IA and significantly surpass chromagram and RT on CJ as shown in Table III. This can be mainly because these two datasets are annotated at the lowest music similarity level as perceived by human listeners, which is well captured by the timbral description of the music content [Logan 2000]. Jingju music tends to have specific leading instruments presenting distinct timbre characteristics at different structural sections which makes timbral features very appealing for structural description of Jingju.

It is noted that *SF* and *CNMF* do not work as effectively as *QN* on the Jingju dataset, as shown in Table III. To validate this, we evaluated these algorithms on the *BeatlesTUT* dataset which is often used to test the repetition-based segmentation algorithms. We used the *Music Structure Analysis Framework* by Nieto[9] which contains the segmentation algorithms of interest[10]. We plugged our *QN* algorithm into the same framework to obtain a direct comparison and used *beat-synchronised chromagram* as the feature descriptor following the default setting. Results are shown in Table IV. *SF* outperforms the other two while *QN* performs the worst ($p < 0.001$). Although the boundary extraction method of this algorithm is designed to be generic [Serrà et al. 2012], its advantage is more pronounced on music with discernible repetitions. When the music is less repetitive, SF may give a low recall rate regardless of the feature used. *CNMF* exhibits more balanced performance for different music genres and feature types. It does not rely on the repetition hypothesis of the music structure and can detect the patterns encoded in the feature matrices. This property of many homogeneity-based algorithms makes them susceptible to noise and tend to yield limited precision rates. This is a shared weakness noted also for novelty based methods which may overlook the global music pattern. The QN algorithm attempts to mitigate this by assessing the sharpness of a peak in the novelty curve hence comparing the extent of novelty of the current frame to the recent past and near future. However, we still observe an unbalanced precision and recall rate in Table IIIa and Table IV. A low boundary retrieval sensitivity of 30 over the range of $[0, 100]$ (see Section 5) gives the optimal segmentation in our system. When using higher sensitivities, the increase in the recall rate does not compensate the downgrade in precision, resulting in degrading F-measures overall.

[9]https://github.com/urinieto/msaf/tree/devel

[10]It has to be noted that the processing techniques used may be different from our system where results from Table III are obtained.

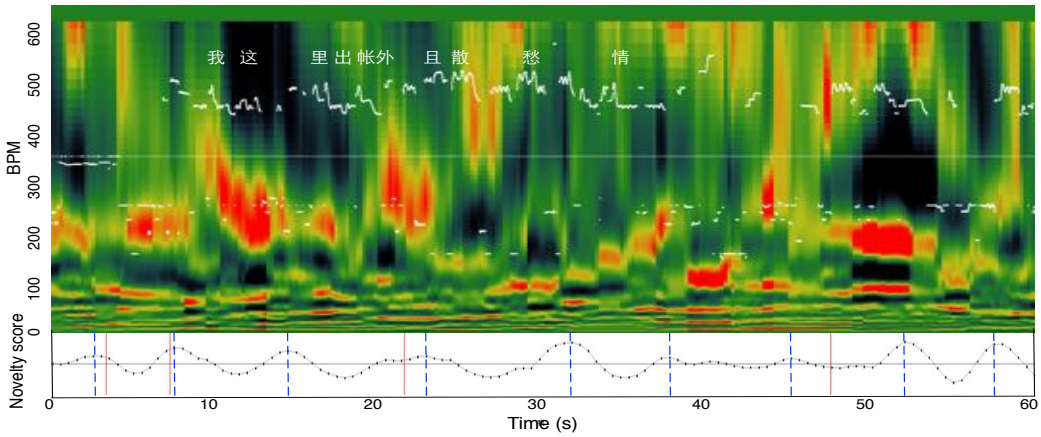


Fig. 8: Tempogram, melody contour and segmentation for a 60-second excerpt of the recording "Ba wang bie ji". The top pane shows the extracted melody (in white curves) and tempogram for the excerpt. Bottom pane shows the smoothed novelty curve using the QN algorithm using $hMFCCs_m$ feature (dotted curve in black), segment annotations (solid lines in red) and retrieved boundaries (dotted lines in navy).

7. CONCLUSION AND FUTURE WORK

In this article, we applied conventionally used features and algorithms for audio structural segmentation to an important non-Western music style, Jingju. As part of this, a new database of Jingju music is presented to complement existing evaluation corpora. The harmonic-percussive source separation technique introduced in the feature extraction process has brought significant improvements in segmentation for both Western and Chinese music categories. Furthermore, we have demonstrated that novelty or homogeneity based segmentation algorithms using timbral features may surpass repetition based ones for Jingju due to its lack of global repetition structures. However, this method produces only limited precision. A possible solution is to use musically meaningful audio features or statistical models to discriminate the non-section related noises in the low-level feature representations and target characterising the melodic and metrical patterns as the underlying structural units (see Section 2.1). Fig. 8 shows the tempogram [Tian et al. 2015] and the predominant melody [Salamon and Gómez 2012] for a 60s excerpt of Jingju music (see Fig. 1). Both the melodic contour and the predominant pulses have the tendency to remain stable or to show steadily evolving patterns within a structural segment, whose sudden break can indicate emergence of new sections. It is also noticeable that a peak in the novelty curve is more likely to indicate a structural boundary when accompanied by prominent rhythmic or melodic changes. In future work, we propose to use the novelty information from timbre features to derive intermediate structural descriptions and meanwhile, rely on rhythmic and melodic modelling for verified segmentations.

The outcomes of this study give strong indications to direct the creation of an intelligent system that automates the selection of audio features and segmentation algorithms, given contextual knowledge of the audio signal such as genre and the level of music structure to analyse. A semi-supervised system capable of encoding human knowledge into the audio signal analysis process in an interactive fashion is also considered as a future direction.

ACKNOWLEDGMENTS

The authors would like to thank the community for making their research reproducible and the anonymous reviewers for their valuable comments and suggestions, and the colleagues from CompMusic project, Universitat Pompeu Fabra for the inspiring discussions.

REFERENCES

J-J Aucouturier, François Pachet, and Mark Sandler. 2005. "The way it Sounds": timbre models for analysis and retrieval of music signals. *Multimedia, IEEE Transactions on* 7, 6 (2005), 1028–1035.

Dawn A. A. Black, Ma Li, and Mi Tian. 2014. Automatic identification of emotional cues in Chinese opera singing. In *Proceedings of the 13th International Conference on Music Perception and Cognition and the 5th Conference for the Asian-Pacific Society for Cognitive Sciences of Music (ICMPC 13-APSC0M 5)*.

Kainan Chen. 2013. *Characterization of Pitch Intonation of Beijing Opera*. Master's thesis. Universitat Pompeu Fabra.

China Music Group (CMG). 2010. Peking Opera Box set, Limited Edition. Audio CD. (2010).

Matthew E.P. Davies and Mark D. Plumbley. 2004. Causal tempo tracking of audio. In *Proceedings of the 5th International Conference on Music Information Retrieval (ISMIR)*.

Lloyd A. Dawe, John R. Platf, and Ronald J. Racine. 1993. Harmonic accents in inference of metrical structure and perception of rhythm patterns. *Perception & psychophysics* 54, 6 (1993), 794–807.

Diana Deutsch. 2012. *The psychology of music*. Academic Press.

Jonathan Driedger, Meinard Müller, and Sascha Disch. 2014. Extending harmonic-percussive separation of audio signals. In *Proceedings of the 15th International Conference on Music Information Retrieval (ISMIR)*.

Jean-Pierre Eckmann, S. Oliffson Kamphorst, and David Ruelle. 1987. Recurrence plots of dynamical systems. *Europhysics Letters* 4, 9 (1987), 973–977.

Derry FitzGerald. 2010. Harmonic/percussive separation using median filtering. In *Proceedings of the 13th International Conference on Digital Audio Effects (DAFx-10)*.

Jonathan Foote. 2000. Automatic audio segmentation using a measure of audio novelty. In *Proceedings of IEEE International Conference on Multimedia and Expo (ICME)*.

Takuya Fujishima. 1999. Realtime chord recognition of musical sound: A system using common lisp music. In *Proceedings of the International Computer Music Conference (ICMC)*.

Aggelos Gkiokas, Vassilios Katsouros, George Carayannis, and Themos Stajylakis. 2012. Music tempo estimation and beat tracking by applying source separation and metrical relations. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Harald Grohganz, Michael Clausen, Nanzhu Jiang, and Meinard Müller. 2013. Converting path structures into block structures using eigenvalue decompositions of self-similarity matrices. In *Proceedings of the 14th International Conferece on Music Information Retrieval (ISMIR)*.

Christopher A. Harte and Mark B. Sandler. 2005. Automatic chord identification using a quantised chromagram. In *Proceedings of the 118th Convention of the Audio Engineering Society*.

Kristoffer Jensen. 2005. A causal rhythm grouping. In *Computer Music Modeling and Retrieval*. Springer.

Florian Kaiser and Thomas Sikora. 2010. Music structure discovery in popular music using non-negative matrix factorization. In *Proceedings of the 11th International Conference on Music Information Retrieval (ISMIR)*.

Mark Levy and Mark B. Sandler. 2008. Structural segmentation of musical audio by constrained clustering. *Audio, Speech, and Language Processing, IEEE Transactions on* 16, 2 (2008), 318–326.

Yuxiang Liu, Qiaoliang Xiang, Ye Wang, and Lianhong Cai. 2009. Cultural style based music classification of audio signals. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Beth Logan. 2000. Mel frequency cepstral coefficients for music modeling. In *Proceedings of the 1st International Conference on Music Information Retrieval (ISMIR)*.

Brian C.J. Moore, Brian R. Glasberg, and Thomas Baer. 1997. A model for the prediction of thresholds, loudness, and partial loudness. *Journal of the Audio Engineering Society* 45, 4 (1997), 224–240.

Oriol Nieto and Tristan Jehan. 2013. Convex non-negative matrix factorization for automatic music structure identification. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Richard Parncutt. 1994. A perceptual model of pulse salience and metrical accent in musical rhythms. *Music Perception* 11 (1994), 409–464.

Jouni Paulus and Anssi Klapuri. 2008a. Acoustic features for music piece structure analysis. In *Proceedings of the 11th International Conference on Digital Audio Effects (Dafx)*.

Jouni Paulus and Anssi Klapuri. 2008b. Labelling the structural parts of a music piece with Markov models. In *Proceedings of Computers in Music Modeling and Retrieval Conference (CMMR)*.

Jouni Paulus, Meinard Müller, and Anssi Klapuri. 2010. State of the art report: audio-based music structure analysis. In *11th International Conference on Music Information Retrieval (ISMIR)*.

Geoffroy Peeters and Emmanuel Deruty. 2009. Is music structure annotation multi-dimensional? A proposal for robust local music annotational music annotation. In *Proceedings of the 10th International Conference on Music Information Retrieval (ISMIR)*.

Rafael Caro Repetto and Xavier Serra. 2014. Creating a Corpus of Jingju (Beijing Opera) Music and Possibilities for Melodic Analysis. In *15th International Conference on Music Information Retrieval (ISMIR 2014)*.

Rafael Caro Repetto, Ajay Srinivasamurthy, Sankalp Gulati, and Xavier Serra. 2014. *Jingju music: concepts and computational tools for its analysis*. Technical Report. Tutorial session, International Conference on Music Information Retrieval (ISMIR).

Xavier Rodet. 2001. *Project Ecrins: calcul des descripteur de bas-niveaux*. Technical Report. Ircam.

Halfdan Rump, Shigeki Miyabe, Emiru Tsunoo, and Nobutaka Ono. 2010. Autoregressive MFCC models for genre classification improved by harmonic-percussion separation. In *Proceedings of the 11th International Conference on Music Information Retrieval (ISMIR)*.

Justin Salamon and Emilia Gómez. 2012. Melody extraction from polyphonic music signals using pitch contour characteristics. *Audio, Speech, and Language Processing, IEEE Transactions on* 20, 6 (2012), 1759–1770.

Craig Stuart Sapp. 2005. Tempo change JND experiment, Mazurka Project. (2005). http://www.mazurka.org.uk/experiments/tempojnd/

Markus Schedl, Emilia Gómez, and Julián Urbano. 2014. Music information retrieval: recent developments and applications. *Foundations and Trends in Information Retrieval* 8, 2-3 (2014), 127–261.

Joan Serrà, Meinard Müller, Peter Grosche, and Josep Lluis Arcos. 2012. Unsupervised detection of music boundaries by time series structure features. In *Proceedings of the 26th AAAI Conference on Artificial Intelligence*.

Xavier Serra. 2011. A multicultural approach in music information research. In *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR)*.

Jordan B. L. Smith. 2010. *A comparison and evaluation of approaches to the automatic formal analysis of musical audio*. Master's thesis. McGill University.

Jordan B. L. Smith, J. Ashley Burgoyne, Ichiro Fujinaga, David De Roure, and J. Stephen Downie. 2011. Design and creation of a large-scale database of structural annotations. In *Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR)*.

Jordan B. L Smith, Isaac Schankler, and Elaine Chew. 2014. Listening as a creative act: Meaningful differences in structural annotations of improvised performances. *Music Theory Online* 20, 3 (2014).

Ajay Srinivasamurthy, Rafael Caro Repetto, Harshavardhan Sundar, and Xavier Serra. 2014. Transcription and recognition of syllable based percussion patterns: the case of Beijing opera. In *Proceedings of the 15th International Conference on Music Information Retrieval (ISMIR)*.

Jonathan P. J. Stock. 1999. A reassessment of the relationship between text, speech tone, melody, and aria structure in Beijing Opera. *Journal of Musicological Research* 18, 3 (1999), 183–206.

Mi Tian, György Fazekas, Dawn A. A. Black, and Mark Sandler. 2015. On the use of tempogram to describe audio content and its application to music structural segmentation. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Mi Tian, Ajay Srinivasamurthy, Mark Sandler, and Xavier Serra. 2014. A study of instrument-wise onset detection in Beijing opera percussion ensembles. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Shanghai wenyi chubanshe. 1992. *Collection of jingju scores ("Jingju qupu jicheng")*.

Elizabeth Wichmann. 1991. *Listening to theatre: the aural dimension of Beijing opera*. University of Hawaii Press.

Luwei Yang, Mi Tian, and Elaine Chew. 2015. Vibrato characteristics and frequency histogram envelopes in Beijing opera singing. In *Proceedings of the 5th International Workshop on Folk Music Analysis (FMA)*.

Received October 2015; revised March 2009; accepted June 2009