Music Structural Segmentation Across Genres with Gammatone Features

TIAN, M; SANDLER, MARKB; ISMIR

http://wp.nyu.edu/ismir2016/

MUSIC STRUCTURAL SEGMENTATION ACROSS GENRES WITH GAMMATONE FEATURES

Mi Tian, Mark B. Sandler
Centre for Digital Music, Queen Mary University of London
{m.tian, mark.sandler}@qmul.ac.uk

ABSTRACT

Music structural segmentation (MSS) studies to date mainly employ audio features describing the timbral, harmonic or rhythmic aspects of the music and are evaluated using datasets consisting primarily of Western music. A new dataset of Chinese traditional Jingju music with structural annotations is introduced in this paper to complement the existing evaluation framework. We discuss some statistics of the annotations analysing the inter-annotator agreements. We present two auditory features derived from the Gammatone filters based respectively on the cepstral analysis and the spectral contrast description. The Gammatone features and two commonly used features, Mel-frequency cepstral coefficients (MFCCs) and chromagram, are evaluated on the Jingju dataset as well as two existing used ones using several state-of-the-art algorithms. The investigated Gammatone features outperform MFCCs and chromagram when evaluated on the Jingju dataset and show similar performance with the Western datasets. We identify the presented Gammatone features as effective structure descriptors, especially for music lacking notable timbral or harmonic sectional variations. Results also indicate that the design of audio features and segmentation algorithms should be adapted to specific music genres to interpret individual structural patterns.

1. INTRODUCTION

Music is primarily an event-based phenomenon comprising a series of musical elements such as melody and harmony that unfold in time. Both human listening and analysis activities can identify the musical structure of a piece that divides its contents into sections each featuring their own characteristics. *Music information retrieval* (MIR) is a research field concerning the extraction of meaningful information from the music content for real world purposes. As a popular MIR task, *Music structural segmentation* (MSS) concerns dividing music into structural parts by giving it boundaries, hence generating high-level music descriptions.

Datasets used to evaluate MIR systems consist mainly of Western popular or contemporary music. The acquisition of non-Western datasets can be highly valuable to combat the Western bias within current MIR paradigm [4, 18]. Understandings of the music structure can be genre-dependent and a segmentation algorithm designed for one corpus may have vague assumptions for another. Smith studied several segmentation algorithms and suggests that algorithms designed for the structural analysis of Western pop music are widely applicable beyond the Western context [21]. Nonetheless, the evaluation corpora used in [21] are still Western centric and collected on a basis of general structural coherence. One primary motivation of this work is to include more challenging genres to analyse the music structure beyond the Western scenario. One of these genres is Jingju, also known as Beijing Opera or Peking Opera, which is one of the most representative genres of Chinese traditional music. An analytical discovery of its song structure will greatly assist its popularisation and subsequent applications such as browsing and indexing. Although it offers intriguing research topics to challenge the existing MIR tools, little work has been done to understand its content using computational methods until very recent years with its structural analysis largely absent from the literature [1]. It should be noted that the song structure has to be differentiated from the structure of a full Jingju play, where the former relates to only the arias part of the latter [27]. In this paper, we include Jingju as a new genre in the MSS study and address the analysis of its song structure.

Various audio features have been used for the structural description of music capturing mainly its harmonic, timbral or rhythmic aspects [16]. A *chromagram* [7], also denoted *Harmonic pitch class profiles* (HPCP), along with its many variants are the most popular features for the structural analysis of Western pop music. The chromagram is a B-dimensional vector representation denoting the energy of each semitone distributed in a chromatic scale, where B is the number of semitones in an octave. The *Mel frequency cepstral coefficients* (MFCCs) feature is among the most used timbre descriptors for MSS studies. It models the shape of the spectral envelope by describing the frequency spectrum aligned on a Mel scale [10]. Rhythmicity may identify music structure beyond the timbral or harmonic variations. It is however much less employed compared to MFCCs and the chroma features [16]. The classical time-frequency (TF) representations used for timbre research are based on the short-time Fourier transform

(STFT) of the audio signal. Although MSS is considered a high-level task involving human perception, auditory cues are barely incorporated in the commonly used audio features. As the second main motivation of this work, we will explore novel timbre features modelling the frequency resolution of the human auditory system to describe the music structure.

This paper is organised as follows. Section 2 introduces the background of Jingju music and the related work on auditory features. We present the new Jingju dataset in Section 3. The investigated Gammatone features are presented in Section 4 and are evaluated in a MSS experiment introduced in Section 5. Section 6 is devoted to analyse the results. Finally we conclude this work in Section 7.

2. BACKGROUND

Compared to the spectrogram which displays the TF components of an audio signal mapped to their physical intensity levels, auditory representations attempt to emphasise its perceptually salient aspects. The *Gammatone* function has been widely used to derive the TF representation modelling human auditory responses of sound and has various applications in research areas such as auditory scene analysis, speech recognition and audio classification [19, 20, 25]. In [19], features derived from the cepstral analysis of Gammatone filterbank outputs outperform the conventional MFCC and perceptual linear prediction (PLP) features in a speech recognition experiment. In a music and audio genre classification study, features extracted from the temporal envelope of a Gammatone filterbank surpass standard features such as MFCCs [12]. In this paper, we present new features derived from Gammatone filters to describe the music timbre and investigate their applications in music structural segmentation.

Distinct from Western pop music commonly used to evaluate MSS algorithms, Jingju may hold very characteristic music form. Repetitive harmonic structures such as the chorus and verse sections typically found in Western music are hardly present in Jingju. Here we provide some essential background of this genre. The song lyrics are organised in a *couplet* structure which lays the basis of the music structural framework. A couplet contains two *melodic lines* performed by the singer with background accompaniments. Although following certain melodic, rhythmic and instrumentation regularities, each couplet unfolds in a temporal order and is hardly repeated with another. A passage of melodic lines expressing specific music ideas or motifs can be grouped into a *melodic section* which can play a rather integrate role in the overall musical form. Jingju consists mainly of three identifiable musical elements: *mode and modal systems*, *metrical patterns*, and *melodic-phrases*. When composing a Jingju play, modal systems and modes are firstly chosen to set the overall atmosphere. The metrical patterns are then accordingly arranged to portray specific content in each passage of lyrics and signal the sectional. Here a *melodic-phrase* differs from the Western understanding for a *melodic phrase* in the sense that it refers to the melodic progression and the tone for singing a single character from the lyrics [23]. It is considered the smallest meaningful unit in Jingju aesthetics. Jingju songs also have instrumental connectives to bridge the sung parts in the arias. Such connectives can serve as preludes to introduce melodic passages and as interludes to tie together successive couplets. Collectively, these musical elements are hierarchically united into composite organisations and shape the overall temporal music structure. In the next section, we will present the collection of the Jingju dataset.

3. DATASETS

3.1 Existing Corpora

A few MSS datasets are available in the literature. Two are used in this work. The first consists of 174 songs from The Beatles. The annotation was first made at Music Technology Group (MTG), Universisitat Pompeu Fabra (UPF) and corrected at Tampere University of Technology (TUT) [15]. We note this dataset *BeatlesTUT*. The second, SALAMI Internet Archive (S-IA), contains 272 pieces as a publicly available subset of the full SALAMI dataset [22]. The main design consideration of the SALAMI dataset is to cover a wide variety of music genres. S-IA also contains a large set of live recordings hence providing a diversity of audio qualities.

These datasets employ different annotation principles. BeatlesTUT is annotated on a *functional* level, i.e., the music is segmented into structural parts expressing specific musical functions. In contrast, S-IA is annotated incorporating different principles on multiple hierarchies including the *music similarity* level, the *function* level and the highest *lead instrument* level. In this paper, we use the music similarity level annotations for S-IA. The inclusion of these two datasets will provide respectively a standard example of Western pop music and a diversity of styles hence gain us meaningful reference for the structural analysis of Jingju. Additionally, they cover two different annotation principles and can serve as a comprehensive testbed for the investigated segmentation algorithms and audio features.

3.2 Composition of Jingju Dataset

The Jingju corpus presented in this paper consists of 30 excerpts from commercial CDs [2], sampled at 44.1 KHz and 16 bits per sample with a total length of 3.6 hours. The CDs were released in the recent decade with recordings of classical repertoires performed by the most renowned musicians. A full Jingju play can last up to a few hours comprising multiple arias or acts. To fulfil the computational purpose of this study, the 30 excerpts in this dataset are taken from 20 different Jingju plays, with an average length of 432 seconds. The audio samples were chosen on the criteria of repertoire coverage, structural diversity and audio quality. One prerequisite an excerpt can be selected is that various structural parts should be present characterising temporal progressions or changes of sectional units. The selected samples cover the two available modes and various metrical patterns. Half of them are performed by

female singers and half by male singers with different role types. We denote this dataset *CJ* in this paper.

In this work, annotations are made to describe the *musical similarity* setting aside the musical functions of segments, similarly to the lowest level of S-IA. There are mainly two reasons why the similarity level is chosen. First, low-level music similarity is a phenomenon that can be perceived for different genres [3]. Analysing the music structure on the similarity level would therefore grant us fair comparison across genres. However, the instrumentations and the music functions of the sectional units can be highly genre-dependent. Second, the melodic sections are never repeated with each other at a segment-level as the chorus-verse based music forms would do. This could lead to dubious decisions in defining the structural sections based on specific musical functions. Meanwhile, there exists much expressiveness in the performance, which may raise the demand of analysing the ornamentations in parallel to the functional structure units hence introducing uncertainties in locating sectional boundaries.

Three listeners (noted "A1", "A2" and "A3") participated in annotating the music. Another two engaged in verifying their annotations, one is the first author of this paper ("V1"), familiar with this music style as an amateur, the other is a Jingju musician and musicologist ("V2"). All annotators are Chinese and were provided with music scores and lyrics [26]. They were instructed to pay attention 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 expressing a unified musical idea. When multiple annotators from A1, A2 and A3 have noted a boundary, it is accepted with its final location being the average of those indicated by individual annotators. When a boundary is noted by only one of A1, A2 and A3, its acceptance rests on a conscious discussion of V1 and V2, over whether a boundary should be noted and if yes, its exact position.

The software used for annotation is Sonic Visualiser [1] which displays the annotators the waveform and the spectrogram of the track and allows them to navigate it through, as well as to add time instants and notes to mark a segment boundary. Figure 1 shows respectively the annotations by V1 and V2 and the final accepted annotation for a 60-second excerpt of the recording "Ba wang bie ji" (meaning "Farewell my concubine"), with the 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 with one sung character may last several seconds. This gives the performer lots of freedom for ornamentations in the singing such as vibratos and even intermittence. The final decision is made when agreements have been reached by V1 and V2. Additional annotation information and metadata for this dataset can be found in [24] and online [2].

Here we discuss some properties of the annotations focussing on the *inter-annotator agreement* (IAA) between

[1] http://www.sonicvisualiser.org/

[2] http://www.isophonics.net/content/jingju-structural-segmentation-dataset

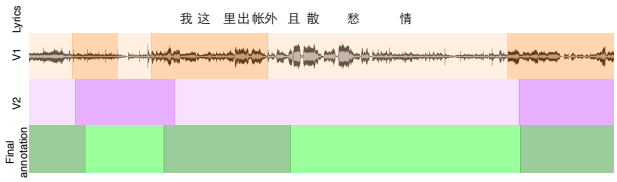

Figure 1: Boundary annotations for a 60-second excerpt of the recording "Ba wang bie ji" from Dataset *CJ*. Panes from top to bottom show respectively the lyrics of the singing (in Chinese), annotations by annotator V1 and V2 and the final annotation.

P 0.5s	R 0.5s	F 0.5s	P 3s	R 3s	F 3s	D_{ad}	D_{da}
0.891	0.675	0.693	0.911	0.689	0.743	0.27	11.88
(0.075)	(0.188)	(0.141)	(0.075)	(0.182)	(0.144)	(0.25)	(48.12)

Table 1: Average agreement between annotator V1 and V2 for recordings in dataset CJ (standard deviations into parenthesis). Statistics include: pairwise precision (P), recall (R) and F-measure (F) measured at 0.5s and 3s, and the median of distances between each annotated segment boundary to its closest detected segment boundary (D_{ad}) and that between each detected segment boundary to its closest annotated segment boundary (D_{da}).

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 2: 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).

V1 and V2. In the assessment of each of the two annotations, one is treated as the "ground truth" and the other as the "detection" and then their roles are rotated. We report the average of these two measures. With a variety of existing measures commonly used to compare multiple annotations for music structure [21], the following metrics are used: *precision* (P), *recall* (F) and *F-measure* (F) retrieved at the tolerance of 0.5s ($\pm$0.25s) and 3s ($\pm$1.5s), median of the distance between each annotated segment boundary to its closest detected segment boundary (D_{ad}) and that between each detected segment boundary to its closest annotated segment boundary (D_{da}). Statistics of the investigated metrics are given in Table 1. The IAA measured at 0.5s is relatively comparable to that measured at 3.0s in the corresponding cases. However, choosing different annotation as the ground truth each time yields lower recall than precision and lower precision than recall, hence substantial *false negative* (FN) and *false positive* (FP) respectively. This is mainly because the two annotators noted different numbers of boundaries. This observation shows that the boundary decisions do depend on the annotators' individual understanding of the music. Some statistics of the datasets used in this paper are given in Table 2. We can notice that the average segment lengths of S-IA and CJ are shorter than that of BeatlesTUT mainly due to individual annotation principles.

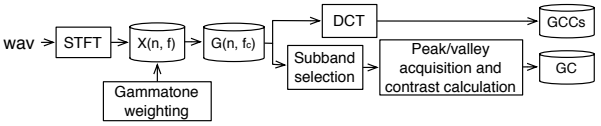

Figure 2: Gammatone feature extraction workflow.

4. GAMMATONE FEATURES

4.1 Gammatone Filters

In the Patterson *Gammatone* model, the cochlea processing is simulated by a Gammatone auditory filterbank with the bandwidth of each filter described by an *Equivalent rectangular bandwidth* (ERB) [14]. The Gammatone function is defined by the impulse response of the signal:

$$GF(t, f_c) = at^{(m-1)}e^{(-2\pi bt)}cos(2\pi f_c t + \phi), \quad (1)$$

where a is the amplitude factor, m is the order of the filter, b is the bandwidth of the filter in Hz which largely determines the duration of the impulse response, f_c in Hz is the centre frequency of the filter and t is the time in s. An efficient implementation is provided by Slaney [20]. It should be noted that $GF(t, f_c)$ keeps the original sample frequency f_s. To derive a spectrogram-like TF representation, noted *Gammatonegram* in this paper, it is necessary to sum up the energy over fixed time windows.

However, to process a signal with a bank of M Gammatone filters can still be computationally expensive. Ellis introduced an alternative method using a *fast Fourier transform* (FFT)-based approximation [5]. In this approach, a conventional fixed-bandwidth spectrogram is first calculated whose frequency bins are then aggregated into Gammatone responses with coarser resolutions via a weighting function. This approximates matches the accurate method by Slaney very closely despite neglecting phase information of each frequency bin when summing them up [5]. Another difference the approximation can introduce is a loss of temporal resolution due to the Fourier transform applied beforehand. However, this is not considered unfavourable in the structural analysis scenario as a relatively coarse temporal resolution is commonly employed aiming at a more musically meaningful scale for the analysis [16]. Many methods propose to use a window size of 0.1 - 1 second, or equal to the beat length [9,16]. In this paper, we use the FFT-based implementation following Ellis [5]. We use M = 64 channels with centre frequencies spaced between 50 Hz to f_s/2 (f_s = 44.1 KHz) on an ERB scale following the default setting of Slaney's and Ellis' toolboxes [5, 20]. The energy of the gammatone filterbank output is aggregated over a 46ms window and shifted by 23ms into the Gammatonegram $G(n, f_c)$.

4.2 Feature Extraction

Two features are extracted from the Gammatonegram capturing different properties of the signal as summarised in Figure 2. Here we describe the feature extraction process. The Discrete cosine transform (DCT) is a commonly used dimensionality reduction technique in feature extraction. It is adopted as the last step in the calculation of the MFCCs which proved highly successful in describing the sound timbre [10]. One motivation of this paper is to find alternative timbral features to describe the music structure incorporating auditory cues. To this end, we introduce a feature called *Gammatone cepstral coefficients* (GCCs) following [19] to describe the average energy distribution of each subband. Specifically, we apply a DCT to $G(n, f_c)$ to de-correlate its components.

$$GCCs(n) = \sum_{i=1}^{M} G(n)cos\left(\frac{\pi}{M}(i + \frac{1}{2})n\right) \quad (2)$$

Shao and his colleagues report that the lowest 30 orders of GCCs contain the majority information of a GF with 128 filterbanks to recover the speech signal [19]. In a sound classification work, the number of filters and GCCs coefficients are set to 48 and 13 with the later identical to MFCCs used in the study [25]. Here we use a 13-coefficient GCCs same as MFCCs to derive a fair comparison of the two. We will discuss the setting of number of coefficients in Section 6. However, a log operation is excluded as applied in common cepstral analyses since initial investigation shows degraded segmentation due to an overemphasis of the lower frequency components when using the logarithmic scale.

Similar to MFCCs, GCCs describe the average energy distribution of each subband in a compact form. Here we are also interested in the extents of flux within the spectra indicating the level of harmonicities associated with different frequency components. To this end, we present a novel feature, *Gammatone contrast* (GC). The extraction of this feature is inspired by the *spectral contrast* (SC) feature which is based on the *octave-scale* filters and is very popular in music genre classification studies [8].

The calculation of the GC feature is as follows. As the first step, the Gammatone filterbank indices $[1, ..., M]$ are grouped into C subbands with linearly equal subdivisions. Since the spectrum is originally laid out on a non-linear ERB scale, the frequency non-linearity is still reserved in the subbands. We use C=6 similar to [8] in this study yielding a subband frequency division of $[50, 363.198, 1028.195, 2440.148, 5438.074, 11803.409, 22050]$. We note $\mathbf{V}$ the Gammatonegram vector of the zth subband $[G_{z,0}(n), G_{z,1}(n), \ldots, G_{z,K-1}(n)]^T$ where $z \in [0, C-1]$, $\mathbf{V}' = \left[G'_{z,0}, G'_{z,1}, \ldots, G'_{z,K-1}\right]^T$ is $\mathbf{V}$ sorted in an ascending order such that $G'_{z,0}(n) < G'_{z,1}(n) < \ldots < G'_{z,K-1}(n)$. We calculate the difference of the strength of peaks and valleys for each subband to derive the C-dimensional GC feature:

$$GC_z(n) = log\left(G'_{z,K-1}(n) - G'_{z,0}(n)\right), \quad (3)$$

GCCs and the vector-wise concatenation of GCCs and GC will be evaluated in comparison with two commonly used features MFCCs and chromagram in the MSS scenario as introduced shortly. The reason why GC is not

evaluated individually is that it measures only the relative contrasts within subband energies hence may lack complementary spectral information. We use the LibROSA music and audio analysis library which provides feature extraction modules for MFCCs and chromagram [11]. We implement the Gammatone module into this library to obtain a uniform feature extraction environment. The extracted GCCs, GC, MFCCs and chromagram are respectively 13-, 6-, 13- and 12-dimensional features. The window and step size for feature extraction are respectively 46ms and 23ms.

5. SEGMENTATION EXPERIMENT

The investigated Gammatone features are evaluated on the presented datasets in a segmentation context using *Music Structure Analysis Framework* (MSAF) which relies on LibROSA [11] for feature extraction and includes a list of recently published segmentation algorithms [13]. Three are used in this paper covering the novelty-, homogeneity- and repetition-based segmentation principles [16]. The first one is included into MSAF by the author of this paper and the rest two are provided by MSAF.

The first method is a novelty-based one presented in a recent work [24] following Foote [6]. A Self-similarity matrix (SSM) is constructed by calculating the pairwise Euclidean distance between vectors of the feature matrix. A Gaussian-tapered "checkerboard" kernel is correlated along the main diagonal of the SSM yielding a novelty curve. Given a list of local maxima detected by the adaptive thresholding from the smoothed novelty curve, a second-degree polynomial $y = ax^2 + bx + c$ is fitted on the novelty curve centred around each local maximum with a window of five samples. In this quadratic model a and c control respectively the sharpness and amplitude of each peak. Assessing these two parameters hence allows us to assess the sharpness and the magnitude of a peak independently where it will only be selected as a segment boundary when both meet set conditions. This method is denoted *Quadratic novelty* (QN) in this paper. The second is a homogeneity-based approach which attempts to segment the music by clustering the frames into different section types [9]. First, audio frames are labelled into hidden Markov model (HMM) states derived from trained features. Then histograms of neighbouring frames are clustered into segment types where the temporal continuity on cluster assignments is obtained from the HMMs. Segment boundaries are retrieved by locating changing of segment types. We note this algorithm *Constrained clustering* (CC). The third method, SF, uses features called *structure features* which incorporate both local and global properties accounting for structural information in the recent past [17]. To construct the structure features, a multi-dimensional time series is firstly obtained by accumulating vectors of a standard audio feature ranging across a time span. A *recurrence plot* (RP) is then computed containing the pairwise resemblance $P_{i,j}$ between time series centred at different time locations i and j. Here, an *RP* differs from an SSM typically used to describe music structure in the sense that $P_{i,j}$ is calculated between feature vectors embedded with time-shifts, i.e., between multi-dimensional time series instead of static vectors. This recurrence nature enables encapsulating both homogeneity and repetition properties in the feature space. The structure features are obtained by estimating Gaussian probability density of the time lag matrix of the RP. Finally, a novelty curve is computed where segment boundaries are detected following the standard novelty approach [6]. In this way, all three basic principles – novelty, homogeneity, and repetition, are combined in the segmentation process.

While QN is newly included into MSAF along with the research presented in this paper, CC and SF are provided by the original MSAF system, with CC forked from its open source software by Levy [9] and SF reimplemented from [17] by Nieto [13].

6. RESULTS AND DISCUSSION

The segmentation boundary retrieval results are evaluated with the *precision* (P), *recall* (R) and *F-measure* (F) measured at 3s [9] using the MSAF framework [13]. Results are shown in Table 3 obtained with system configurations parameterised both globally and on each dataset individually. By doing this, we are aiming to investigate how dependent each algorithm is on parameter configurations in the context of a specific dataset. To avoid a potential overfitting, the discussions made in the remainder of this paper are based on the results obtained with the globally uniform configurations, unless noted otherwise.

Here we analyse presented Gammatone features as structural descriptors. We first compare GCCs to MFCCs, both are based on cepstral analysis of the spectra and related to the music timbre. When assessed on individual datasets, GCCs outperform MFCCs on CJ using all investigated segmentation algorithms, with statistical significance observed when using CC ($p = 0.035$) and QN ($p = 0.019$), while the two strike a tie on BeatlesTUT and S-IA.

Figure 3 shows the SSMs derived from the MFCCs and GCCs on an excerpt of the Jingju song "Ba wang bie ji" from CJ (only the first 60 seconds are shown for visualisation purposes). It can be noticed that GCCs yield more distinguished sectional variations than MFCCs. Singing-based musical works such as Jingju or Western opera may present less notable repetitive harmonic or rhythmic patterns than the popular music. However, the vocal-driven nature makes the singing voice an important discriminator of the music structure with its salient presence in the overall instrumentation. In the case of Jingju, new structural units can emerge in the same melodic passage with subtle timbral variations, as shown in Figure 3(a). The dynamics introduced by the singing voice are mainly present in the lower frequency part of the spectrum, which can be better captured by using the ERB scale than Mel. Meanwhile, emphasising the lower sound levels, the ERB warping can be robust against high-frequency transients which may interfere with the analysis. When the music presents more distinguishable timbral variations, GCCs summarise the structure equally effectively as MFCCs, as indicated by their comparable performances on BeatlesTUT and S-

	GCC						GCC+GC						MFCC						Chromagram					
	Individual config			Global config			Individual config			Global config			Individual config			Global config			Individual config			Global config		
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
BeatlesTUT	0.601	0.647	0.614	0.557	0.706	0.612	0.600	0.650	0.615	0.568	0.706	0.618	0.599	0.634	0.606	0.568	0.706	0.618	0.588	0.636	0.600	0.527	0.668	0.579
CJ	0.684	0.486	0.550	0.732	0.448	0.538	0.701	0.465	0.552	0.688	0.436	0.522	0.701	0.483	0.555	0.689	0.441	0.525	0.651	0.509	0.540	0.645	0.447	0.514
S-IA	0.550	0.525	0.524	0.500	0.562	0.515	0.555	0.536	0.535	0.514	0.586	0.533	0.572	0.559	**0.551**	0.517	0.565	0.526	0.558	0.544	0.535	0.514	0.596	0.535

(a) CC

BeatlesTUT	0.564	0.643	0.588	0.523	0.687	0.580	0.565	0.667	0.598	0.523	0.710	0.587	0.638	0.589	0.596	0.584	0.635	0.580	0.468	0.691	0.544	0.435	0..726	0.530
CJ	0.619	0.715	0.639	0.685	0.475	0.543	0.599	0.761	**0.654**	0.673	0.521	**0.574**	0.588	0.715	0.625	0.706	0.439	0.521	0.520	0.798	0.616	0.557	0.593	0.562
S-IA	0.463	0.599	0.500	0.430	0.639	0.492	0.471	0.623	0.516	0.438	0.666	0.508	0.526	0.572	0.525	0.478	0.610	0.513	0.413	0.663	0.486	0.394	0.704	0.480

(b) QN

BeatlesTUT	0.625	0.739	0.667	0.594	0.755	0.654	0.630	0.743	0.673	0.603	0.761	0.663	0.644	0.743	0.678	0.621	0.772	0.671	0.644	0.751	**0.683**	0.612	0.777	**0.679**
CJ	0.559	0.799	0.631	0.688	0.461	0.534	0.536	0.807	0.628	0.664	0.471	0.540	0.554	0.792	0.627	0.677	0.482	0.530	0.542	0.759	0.617	0.668	0.439	0.514
S-IA	0.497	0.577	0.515	0.442	0.649	**0.545**	0.502	0.586	0.520	0.433	0.635	0.493	0.504	0.588	0.523	0.443	0.635	0.497	0.494	0.588	0.524	0.438	0.630	0.500

(c) SF

Table 3: Segmentation results using investigated features on the *BeatlesTUT*, *CJ* and *S-IA* datasets with algorithm *CC*, *QN* and *SF*. Highest F-measure obtained for each dataset is shown in bold.

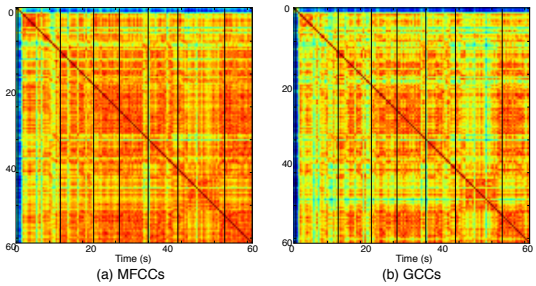

Figure 3: SSMs computed using MFCCs and GCCs on the first 60 seconds excerpt of "Ba wang bie ji" from CJ. Vertical lines indicate segment boundaries.

IA shown in Table 3. This hence suggests the GCCs as a competitive alternative to the commonly used features for music structural description.

Combining GC with GCCs by matrices concatenation has introduced improvements over using GCCs alone for most cases on the investigated datasets with each algorithm tested. However, a statistical significance is not always present as suggested by Student's t-test with related samples when comparing $GCCs + GC$ to $GCCs$. The main effect of using the additional GC feature is a more pronounced within-SSM variance. This has led to the retrieval of more boundaries as indicated by a higher recall in the general case yet an occasional degrading precision.

Investigated features and algorithms perform differently on Western and Jingju music. Chromagram feature and MFCCs work reliably on BeatlesTUT and S-IA. For Jingju, timbre features capture its structural characteristics better than the chroma feature, with auditory inspired Gammatone features outperforming MFCCs. When using the same features and algorithms, similar segmentation results in terms of F-measures tend to emerge from CJ and S-IA, both use the annotation at the music similarity level. However, it can be noticed that algorithms are more dependent on parameter configurations when evaluated on CJ than on S-IA and BeatlesTUT, reflected by the substantial degradation of the F-measures observed when changing the parameter configuration tuned for the individual dataset to the global setting. This suggests the need of designing new segmentation methods to bridge the gaps between genres. It also implies that contextual knowledge, such as the genre and the level of music structure to analyse, can assist a segmentation system to obtain better performance.

It is also noted that SF appears less effective than QN on the CJ dataset when using the chromagram feature, as shown in Table 3. This is in contrast with the many observations for Western pop music, where repetition-based segmentation algorithms are identified as useful interpreters of structural characteristics reflected by chroma features. We find that for Jingju, the chromagram feature forms mainly *block* structures as do the MFCCs instead of *stripes* in the sub-diagonals of the SSMs. This somehow contradicts with many established observations for Western pop music. As introduced in Section 2, the repetitive chord structure is lacking in Jingju in the sense of chorus and verse. The chroma feature in the Jingju scenario functions mainly to capture its low-level homogeneity in the vicinity. Therefore, the same audio feature may exhibit different structural characteristics for specific music genres and the selection of segmentation algorithms should be adapted accordingly to interpret such patterns.

7. CONCLUSION

This paper investigated novel features derived from Gammatone filters for music structural segmentation beyond the commonly studied music corpora. A new dataset with Chinese traditional Jingju music is presented to complement the existing evaluation corpora. In the music structural segmentation experiment, GCCs surpass MFCCs notably on the Jingju dataset comprising vocal-driven music. The fact that the Gammatone features also obtain comparable segmentation results to MFCCs and chromagram on the Beatles and S-IA datasets indicate them to be competitive alternatives to existing audio features for music structural analysis. Different patterns emerge for different music genres from existing algorithms and audio features, shedding new perspectives on music structural segmentation research.

8. REFERENCES

[1] R. Caro Repetto, A. Srinivasamurthy, S. Gulati, and X. Serra. Jingju music: concepts and computational tools for its analysis. Technical report, Tutorial session,

International Society for Music Information Retrieval Conference (ISMIR), 2014.

[2] China Music Group (CMG). Peking opera box set, limited edition. Audio CD, http://chinamusicgroup.com/get_music.php, 2010.

[3] D. Deutsch, editor. *The psychology of music*. Academic Press, 3rd edition, 2012.

[4] J. S. Downie. Toward the scientific evaluation of music information retrieval systems. In *Proceedings of the 4th International Society for Music Information Retrieval Conference (ISMIR)*, 2003.

[5] D. P. W. Ellis. Gammatone-like spectrograms. http://www.ee.columbia.edu/ dpwe/resources/matlab/, 2009.

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

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

[8] D. Jiang, L. Lu, H. Zhang, J. Tao, and L. Cai. Music type classification by spectral contrast feature. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, 2002.

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

[10] B. Logan. Mel frequency cepstral coefficients for music modeling. In *Proceedings of the 1st International Society for Music Information Retrieval Conference (ISMIR)*, 2000.

[11] B. McFee, M. McVicar, C. Raffel, D. Liang, O. Nieto, E. Battenberg, J. Moore, D. Ellis, R. Yamamoto, R. Bittner, D. Repetto, P. Viktorin, J. F. Santos, and A. Holovaty. Librosa: Python library for audio and music analysis. In *Proceedings of the 14th Python in Science Conference*, 2015.

[12] M. F. McKinney and J. Breebaart. Features for audio and music classification. In *Proceedings of the 4th International Society for Music Information Retrieval Conference (ISMIR)*, 2003.

[13] O. Nieto and J. P. Bello. Msaf: Music structure analytis framework. In *Late breaking session, 16th International Society for Music Information Retrieval Conference (ISMIR)*, 2015.

[14] R. D. Patterson, I. Nimmo-Smith, J. Holdsworth, and P. Rice. An efficient auditory filterbank based on the gammatone function. Applied psychology unit (apu), report 2341, Cambrige, UK, 1987.

[15] J. Paulus and A. Klapuri. Labelling the structural parts of a music piece with markov models. In *Proceedings of Computers in Music Modeling and Retrieval Conference (CMMR)*, 2008.

[16] J. Paulus, M. Müller, and A. Klapuri. State of the art report: audio-based music structure analysis. In *Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR)*, 2010.

[17] J. Serrà, M. Müller, P. Grosche, and J. L. Arcos. Unsupervised detection of music boundaries by time series structure features. In *Proceedings of the 26th AAAI Conference on Artificial Intelligence*, 2012.

[18] X. Serra. Creating research corpora creating research corpora for the computational study of music: the case of the compmusic project. In *Proceedings of International Audio Engineering Society (AES) Conference*, 2014.

[19] Y. Shao, Z. Jin, D. Wang, and S. Srinivasan. An auditory-based feature for robust speech recognition. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2009.

[20] M. Slaney. An efficient implementation of the patterson-holdsworth auditory filter bank. Technical report, Apple Technical Report, 1993.

[21] J. B. L. Smith. A comparison and evaluation of approaches to the automatic formal analysis of musical audio. Master's thesis, McGill University, 2010.

[22] J. B. L. Smith, J. A. Burgoyne, I. Fujinaga, D. De Roure, and J. S. Downie. Design and creation of a large-scale database of structural annotations. In *Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR)*, 2011.

[23] J. P. J Stock. A reassessment of the relationship between text, speech tone, melody, and aria structure in beijing opera. *Journal of Musicological Research*, 18(3):183–206, 1999.

[24] M. Tian and M. B. Sandler. Towards music structural segmentation across genres: features, structural hypotheses and annotation principles. *Special Issue on Intelligent Music Systems and Applications, Intelligent Systems and Technology, ACM Transactions on (ACM TIST)*, In press.

[25] X. Valero and F. Alías. Gammatone cepstral coefficients: biologically inspired features for non-speech audio classification. *Multimedia, IEEE Transactions on*, 14(6):1684–1689, 2012.

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

[27] E. Wichmann. *Listening to theatre: the aural dimension of Beijing opera*. University of Hawaii Press, 1991.